Yandex Cloud

Атлас безопасности AI-приложений

AI SAFE 2.0

AI SAFE 2.0 — фреймворк для системного рассмотрения угроз для приложений, использующих AI. Он выполнен в форме атласа: набора карт, которые связывают между собой воздействия, недостатки-первопричины, защитные меры и компоненты архитектуры AI-приложений и позволяют перемещаться по этим связям.

Команда проекта

Yandex Cloud × solidlab

Содержание

Введение 9

Схема компонентов 11

Архитектура AI-системы 13

Классы недостатков 18

Предметный указатель 21

По компонентам архитектуры 21
По классификационным признакам 27
По области применения 32
По сопоставлению со стандартом 32
По сервисам Yandex Cloud 33

Компоненты 35

AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения 35
AISAFE2-COMP-002 · Модель и интерфейс инференса 37
AISAFE2-COMP-003 · Управление контекстом общения с моделью 40
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core) 42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы 44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний 48
AISAFE2-COMP-007 · Долговременная память 50
AISAFE2-COMP-008 · Данные для подсистемы обучения 52
AISAFE2-COMP-009 · Конвейер обучения и дообучения 54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели 56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения 58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты 61
AISAFE2-COMP-013 · Телеметрия и обнаружение 64
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения 65
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска 67
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса 69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика 72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши 75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов 77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером 79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов 82
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки 84
AISAFE2-COMP-023 · Подсистема формирования базы знаний 86
AISAFE2-COMP-024 · Артефакты работы AI-системы 88
AISAFE2-COMP-025 · Источники данных для инструментов 90

Воздействия 92

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой 92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы 93
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы 95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных 97
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы 98
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей 100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core 101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов) 104
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов 106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов 107
AISAFE2-IMP-011 · Компрометация баз знаний 109
AISAFE2-IMP-012 · Кража модели 111
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели 113
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных 114
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий) 116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий) 118
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям 120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов 121
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов 123
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы 125
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов 126
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала 127
AISAFE2-IMP-023 · Перегрузка контура человеческой проверки 129
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур 130
AISAFE2-IMP-025 · Прямые финансовые потери 133
AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов 134
AISAFE2-IMP-027 · Несоответствие требованиям регуляторов 134
AISAFE2-IMP-028 · Нарушение этических норм 135
AISAFE2-IMP-029 · Репутационный ущерб 136
AISAFE2-IMP-030 · Утеря конкурентного преимущества 136
AISAFE2-IMP-031 · Утрата датасета в результате атаки 137
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами 139
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки 141

Недостатки 143

AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу 143
AISAFE2-WEAK-002 · Отсутствие привязки утверждений AI-ответа к источникам 145
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия 146
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции 148
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе 151
AISAFE2-WEAK-006 · Отсутствие раскрытия факта взаимодействия с AI 152
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными 154
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью 158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций 160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст 162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение 164
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI 167
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных 171
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи 173
AISAFE2-WEAK-015 · Переобучение модели 175
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам 176
AISAFE2-WEAK-017 · Отсутствие проверяемых критериев выбора модели 179
AISAFE2-WEAK-018 · Избыточная наблюдаемость и воспроизводимость поведения модели через интерфейс 180
AISAFE2-WEAK-019 · Избыточное раскрытие внутренних оценок модели 182
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации 184
AISAFE2-WEAK-023 · Незашифрованное хранение 186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи 188
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами 190
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов 194
AISAFE2-WEAK-027 · Отсутствие экономического ограничения массовых запросов 196
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений 199
AISAFE2-WEAK-030 · Отсутствие проверяемой маркировки модели и её выходов 201
AISAFE2-WEAK-031 · Недостаточная защита модели при поставке в недоверенную среду 203
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре 204
AISAFE2-WEAK-033 · Раскрытие сведений об отдельных обучающих записях через модель 206
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных 207
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации) 209
AISAFE2-WEAK-037 · Непропорциональный вклад отдельных обучающих записей в результат обучения 211
AISAFE2-WEAK-039 · Выдача инструментом чувствительных или избыточных данных 212
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных 214
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам 216
AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария 217
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции 220
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы 222
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps 223
AISAFE2-WEAK-046 · Недостаточное управление соответствием требованиям 225

AISAFE2-WEAK-047 · Недостаточное управление уязвимостями 227
AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление 229
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A 232
AISAFE2-WEAK-050 · Остаточные данные в общей инфраструктуре исполнения моделей 234
AISAFE2-WEAK-051 · Выдача результата без завершения обязательных стадий вычислений 235
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки 237
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации 240
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели 242
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель 245
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов 246
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий 248
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора 251
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели 253
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе 254
AISAFE2-WEAK-061 · Избыточно открытая внутренняя телеметрия 255
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов 257
AISAFE2-WEAK-063 · Недостаточные пользовательские средства управления расходами и прозрачность тарификации 260
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей 262
AISAFE2-WEAK-065 · Избыточная маршрутизация задач на человеческую проверку 265
AISAFE2-WEAK-066 · Недостаточная мощность и резервирование человеческой проверки 266
AISAFE2-WEAK-067 · Недостаточная инструментальная поддержка оператора HITL 267
AISAFE2-WEAK-068 · Недостаточный контроль качества решений и утомления операторов HITL 268
AISAFE2-WEAK-069 · Неопределённый режим переполнения и недоступности человеческой проверки 269
AISAFE2-WEAK-070 · Отсутствие управляемой политики допустимого использования AI-сервисов 270
AISAFE2-WEAK-071 · Отсутствие явной модели доверенных зон и потоков данных 271
AISAFE2-WEAK-072 · Отсутствие обязательного технического контроля межзонных потоков данных 273
AISAFE2-WEAK-073 · Отсутствие независимого проверочного материала для AI-результата 274
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели 276
AISAFE2-WEAK-075 · Отсутствие совокупных пределов масштабирования и расходов инфраструктуры 277
AISAFE2-WEAK-076 · Отсутствие ресурсной изоляции арендаторов в разделяемой среде 280
AISAFE2-WEAK-077 · Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку 281
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке 282
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения 284
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода 286
AISAFE2-WEAK-081 · Неаутентифицированный внешний вход в AI-приложение 289
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей 290
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей 292

AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов 294
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов 296
AISAFE2-WEAK-086 · Утрата атрибутов доступа исходного материала при индексировании базы знаний 298
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями 299
AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы 301
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента 303
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода 306
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели 308
AISAFE2-WEAK-092 · Недостаточно ограниченные делегированные полномочия пользователя у агента 310
AISAFE2-WEAK-093 · Неверная атрибуция тарифицируемого потребления между арендаторами 311
AISAFE2-WEAK-094 · Небезопасное хранение и распространение делегированных учётных данных пользователя 312
AISAFE2-WEAK-095 · Недостаточно ограниченные делегированные платёжные полномочия 314

Меры защиты **316**

AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам 316
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода 320
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста 324
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений 328
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода 331
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости 333
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG 339
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности 342
AISAFE2-MIT-009 · Версионирование и откат базы знаний 346
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур 349
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ 356
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели 359
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском 364
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели 368
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов 371
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов 375
AISAFE2-MIT-017 · Изолированное исполнение инструментов 379
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями 383
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения 389
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние 392
AISAFE2-MIT-021 · Управление доверием и сроком жизни долговременной памяти 396
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки 400
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений 405
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов 410
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы 415

AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации 419
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений 423
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений 426
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом 429
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов 434
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером 438
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы 441
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти 444
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр 448
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы 451
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации 456
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений 459
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом 462
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария 466
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков 469
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения 473
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель 478
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения 481
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс 485
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями 488
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости 492
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным 494
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений 497
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса 500
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных 503
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата 507
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации 509
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска 514
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных 519
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps 525
AISAFE2-MIT-056 · Исполнение версионируемой политики допустимости содержимого 529
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов 531
AISAFE2-MIT-058 · Аутентификация на внешнем входе AI-приложения 535
AISAFE2-MIT-059 · Усиленная аутентификация привилегированных человеческих учётных записей 538
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями 541
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию 544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении 547
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами 550
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы 554
AISAFE2-MIT-065 · Защищённое сохранение материалов расследования 558
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация 561

AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений 564
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов 568
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями 571
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы 573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском 577

Реестр связей **581**

Введение

AI SAFE 2.0 — фреймворк для системного рассмотрения угроз для приложений, использующих AI. Он выполнен в форме атласа: набора карт, которые связывают между собой воздействия, недостатки-первопричины, защитные меры и компоненты архитектуры AI-приложений и позволяют перемещаться по этим связям.

Мотивация и цель

В первой версии AI SAFE мы предложили модель угроз, построенную на декомпозиции типовой архитектуры AI-приложений: выделили пять уровней и рассмотрели угрозы на каждом из них. Такой взгляд — со стороны архитектуры и решаемых задач — оказался полезным, но по своей природе односторонним.

Цель второй версии — дать возможность рассматривать область защиты AI-приложений с нескольких точек зрения:

- со стороны рисков и последствий реализации угроз;
- со стороны первопричин — недостатков в процессах, архитектуре, реализации и эксплуатации, позволяющих угрозам реализоваться;
- со стороны защитных мер;
- со стороны архитектуры приложений — как и в первой версии.

Поэтому AI SAFE 2.0 выпускается в форме атласа. Карта в нём — это срез, который пользователь формирует сам: он выбирает интересующие его сущности и перемещается по связям между воздействиями, недостатками, защитными мерами и компонентами архитектуры.

При подготовке документа мы опирались на опубликованные ранее фреймворки, каталоги и рекомендации: материалы OWASP GenAI Security Project, Google SAIF, Databricks AI Security Framework, MITRE ATLAS и другие.

Аудитория

- Архитекторам, бизнес-аналитикам и разработчикам, создающим AI-решения или встраивающим их в собственные продукты и сервисы, — для реализации свойств защищённости на уровне архитектуры и бизнес-сценариев.
- CISO и руководителям служб безопасности — для планирования защитных мер и оценки их покрытия относительно возможных рисков и последствий.
- Инженерам безопасности, DevSecOps-специалистам, участникам процессов MLOps и MLSecOps — для понимания и устранения первопричин в своей зоне ответственности, ведущих к реализации специфических угроз в отношении AI-приложений, а также поддерживающих их инфраструктур и процессов.
- Специалистам по оценке рисков, юристам и руководителям — чтобы узнать о практических последствиях для бизнес-задач, возникающих при создании и эксплуатации AI-приложений.
- Аналитикам, преподавателям и студентам — как системный способ изучения области угроз и защиты AI-приложений.

Ограничения

1. AI SAFE 2.0 сфокусирован на угрозах, недостатках и защитных мерах, специфичных для AI-приложений и сервисов. Атлас также учитывает общие компоненты инфраструктуры и процессы разработки и эксплуатации, влияющие на безопасность AI-приложений, но не претендует на исчерпывающее описание их защиты. Для более подробного рассмотрения общих вопросов безопасности приложений рекомендуем обращаться к моделям зрелости OWASP SAMM и DSOMM, требованиям OWASP ASVS, каталогам недостатков CWE и шаблонов атак CAPEC, а также к материалам по безопасности Yandex Cloud.
2. AI SAFE 2.0 не сводится к таксономии. Классификации помогают ориентироваться в материалах, а основа атласа — связи между воздействиями, недостатками, защитными мерами и компонентами архитектуры.
3. Помимо прямых связей между сущностями существуют многошаговые причинные пути. Например, реализация одной угрозы (несанкционированный доступ) может открыть путь к реализации другой (кража модели). Исходный недостаток при этом может участвовать в причинной цепочке, ведущей к последующим воздействиям. Для каждого перехода необходимо проверить условия его реализации. Мы не превращаем все многошаговые пути в прямые связи между их началом и концом. Отдельные причинные цепочки представлены явно, с условиями применимости. При анализе конкретных сценариев стоит учитывать и другие возможные цепочки.

Как пользоваться атласом

Основной раздел перечисляет четыре типа сущностей атласа — воздействия, недостатки, защитные меры и компоненты архитектуры — и связи между ними. Начинать работу с атласом можно с любой точки зрения, близкой вашей роли: от возможных воздействий и их последствий, от недостатков, от защитных мер или от архитектуры приложения. Выбрав интересующие сущности определённых категорий, вы формируете собственную карту и переходите по связям к другим сущностям атласа.

Команда проекта

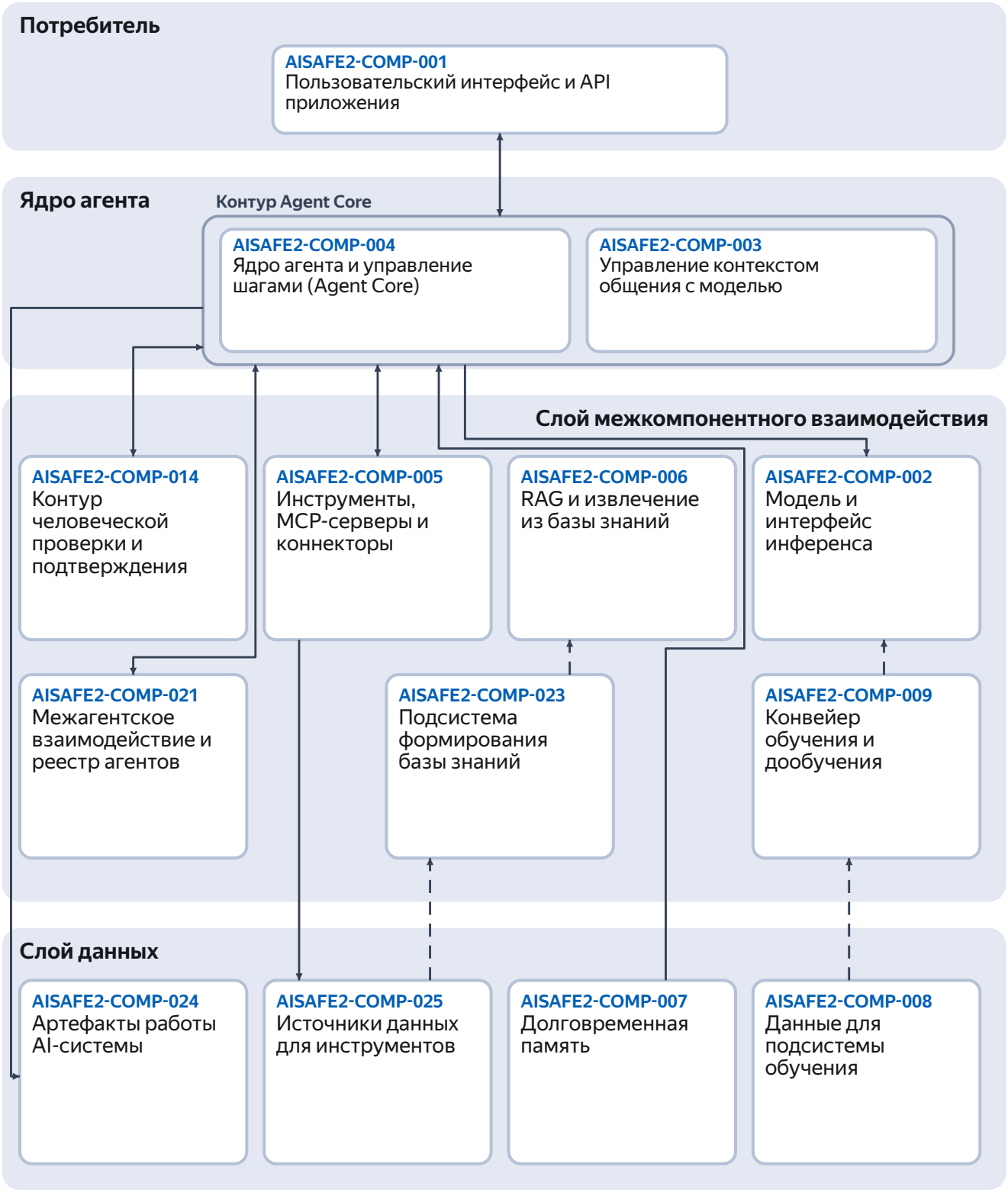
Команда Безопасности ИИ Яндекс Облака и команда AI security SolidLab.

Yandex Cloud × solidlab

Схема компонентов

Основные слои

4 слоя · 14 компонентов



Сквозные и периферийные контуры

Сквозные контуры

6 контуров · 11 компонентов

Прикладное исполнение и операционные данные

- **AISAFE2-COMP-011** Вычислительная и сетевая среда выполнения
- **AISAFE2-COMP-017** Серверная часть приложения и бизнес-логика
- **AISAFE2-COMP-018** Операционные данные, очереди и кэши

Разработка, выпуск и цепочка поставки

- **AISAFE2-COMP-022** Программные артефакты и цепочка поставки
- **AISAFE2-COMP-015** Контур разработки, тестирования и выпуска
- **AISAFE2-COMP-010** Хранилище и реестр артефактов модели

Идентификация и управление доступом

- **AISAFE2-COMP-012** Идентификация, авторизация, ключи и секреты

Исполнение политик и защита запросов

- **AISAFE2-COMP-019** Контур исполнения политик и защиты AI-запросов

Внешний AI-провайдер

- **AISAFE2-COMP-016** Адаптер внешнего AI-сервиса
- **AISAFE2-COMP-020** Контур управления внешним AI-провайдером

Телеметрия и обнаружение

- **AISAFE2-COMP-013** Телеметрия и обнаружение

Архитектура AI-системы

Для анализа конкретного AI-решения сопоставьте его архитектуру с представленными компонентами. Компоненты описаны обобщённо, чтобы схему можно было применять к разным AI-системам.

Они сгруппированы в четыре слоя: потребитель, ядро агента, межкомпонентное взаимодействие и данные. Компоненты могут находиться внутри или вне основного контура и принадлежать разным зонам ответственности.

Потребитель

Человек, бизнес-процесс, подсистема автоматической обработки созданного кода либо другая программа или система, которые запускают работу AI-системы и получают результат.

- AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения

  Потребитель

  Человек задаёт вопрос и получает результат; бизнес-процесс, кодовая подсистема или другая программа могут потреблять результат автоматически.

Ядро агента

Бизнес-логика и композиция AI-решения: координация задачи, управление шагами, контекстом и обращениями к моделям и доступным возможностям.

- AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)

  Ядро агента

  Контур бизнес-логики и композиции AI-решения, который ведёт задачу, выбирает шаги и координирует модели и доступные возможности.

  Управление контекстом общения с моделью

  В составе: Ядро агента

  Возможность Agent Core, которая формирует контекст обращения к моделям: с учётом происхождения, доверия и доступа объединяет инструкции, пользовательский ввод, результаты работы инструментов, данные из кратковременных и долговременных хранилищ и другие сведения.

  - AISAFE2-COMP-003 · Управление контекстом общения с моделью

Слой межкомпонентного взаимодействия

Граница запросов и ответов с моделями, инструментами, RAG, другими агентами и человеческой проверкой, в том числе когда компонент поставляет другая команда или вендор.

Человек в контуре AI-системы

Обязательная человеческая проверка и подтверждение значимых решений или действий.

- AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения

Инструменты

Исполняемые возможности AI-системы и внешние прикладные системы, доступные через авторизованные вызовы с проверкой результата.

- AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы

Формирование базы знаний

Допуск источников, преобразование и защита данных, их индексирование и публикация сформированной базы знаний.

- AISAFE2-COMP-023 · Подсистема формирования базы знаний

RAG и извлечение данных

Авторизованный поиск, ранжирование и возврат фрагментов из подготовленной базы знаний для обогащения текущего запроса.

- AISAFE2-COMP-006 · RAG и извлечение из базы знаний

Межагентское взаимодействие

Обнаружение независимых агентов и обращение к агентам или субагентам через A2A и другие протоколы: аутентификация и проверка доверия, защищённые конверты сообщений, корреляция, защита от повтора и контролируемое делегирование задач.

- AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов

Модели

Раздельные поверхности исполнения обученной модели и подготовки новой версии модели.

Инференс

В составе: Модели

Исполнение обращения к выбранной версии модели и возврат вероятностного результата.

- AISAFE2-COMP-002 · Модель и интерфейс инференса

Обучение

В составе: Модели

Подготовка новой версии модели из разрешённых наборов данных на основе настроек обучения.

- AISAFE2-COMP-009 · Конвейер обучения и дообучения

Слой данных

Источники, артефакты, состояние ядра агента и наборы данных, поддерживающие работу системы.

Артефакты работы AI-системы

Недоверенные переиспользуемые результаты, включая созданные код, изображения, документы и сохраняемый текст.

- AISAFE2-COMP-024 · Артефакты работы AI-системы

Источники данных для инструментов

Внешние и внутренние исходные данные, используемые инструментами, включая RAG: API, файлы, хранилища и общедоступные источники, состав, целостность, происхождение, правила доступа и конфиденциальность которых требуют проверки. Данные обучения и оценки относятся к отдельному контуру.

- AISAFE2-COMP-025 · Источники данных для инструментов

Хранилище ядра агента

Системные инструкции, состояние сессии, предпочтения и другие сохраняемые сведения, используемые Agent Core, включая общее состояние нескольких агентов.

- AISAFE2-COMP-007 · Долговременная память

Данные для подсистемы обучения

Защищённые специализированные наборы примеров, разметки и эталонных результатов, передаваемые в контур обучения, дообучения или оценки моделей.

- AISAFE2-COMP-008 · Данные для подсистемы обучения

Не AI-специфичные аспекты разработки и эксплуатации

Прикладное исполнение и операционные данные

Серверная логика, вычислительная среда, сеть, очереди, кэши и рабочие данные пересекают несколько слоёв. Проверьте границы исполнения, изоляции и хранения.

- AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения
- AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика
- AISAFE2-COMP-018 · Операционные данные, очереди и кэши

Идентификация и управление доступом

Идентичности, авторизация, ключи и секреты защищают все исполняемые и внешние пути. Проверьте минимальные полномочия, выдачу и отзыв доступа.

- AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты

Телеметрия и обнаружение

События и сигналы собираются на нескольких слоях. Проверьте полноту наблюдения, обнаружение отклонений и достаточность данных для расследования.

- AISAFE2-COMP-013 · Телеметрия и обнаружение

Исполнение политик и защита запросов

Детерминированные проверки действуют до, во время и после обращения к модели или инструменту. Проверьте точки принудительного применения и отказобезопасность.

- AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов

Внешний AI-провайдер

Адаптер и административный контур поставщика находятся за локальной границей, но участвуют в инференсе. Проверьте маршрут, настройки данных, модели, регионы и квоты.

- AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса
- AISAFE2-COMP-020 · Контур управления внешним AI-провайдером

Разработка, выпуск и цепочка поставки

Разработка, тестирование, выпуск, реестры моделей и программные артефакты относятся к жизненному циклу. Проверьте происхождение, целостность и контроль допуска.

- AISAFE2-COMP-010 · Хранилище и реестр артефактов модели
- AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска
- AISAFE2-COMP-022 · Программные артефакты и цепочка поставки

Защитные ограничения зависят от реализации

Защитные ограничения не образуют один универсальный компонент. В конкретной системе они могут быть реализованы в модели, детерминированном контуре политик, инструменте или внешнем API либо через обязательную человеческую проверку.

- Контроль на стороне модели: Инференс
- Детерминированное исполнение политики: Исполнение политик и защита запросов
- Инструмент или внешний API: Инструменты
- Обязательная человеческая проверка: Человек в контуре AI-системы

Высокоуровневые бизнес-последствия

Это последствия второго порядка: они показывают, зачем бизнесу анализировать технические воздействия, но не подменяют их. Они намеренно не сопоставляются напрямую с компонентами или мерами защиты; защиту выбирают через технические воздействия и недостатки.

- AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
- AISAFE2-IMP-028 · Нарушение этических норм
- AISAFE2-IMP-027 · Несоответствие требованиям регуляторов
- AISAFE2-IMP-025 · Прямые финансовые потери
- AISAFE2-IMP-029 · Репутационный ущерб
- AISAFE2-IMP-030 · Утеря конкурентного преимущества

Классы недостатков

AISAFE2-WC-002 · Поведение при обучении и дообучении
AISAFE2-WEAK-059, AISAFE2-WEAK-037, AISAFE2-WEAK-015, AISAFE2-WEAK-033

AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений
AISAFE2-WEAK-009, AISAFE2-WEAK-074, AISAFE2-WEAK-011, AISAFE2-WEAK-010, AISAFE2-WEAK-054, AISAFE2-WEAK-008

AISAFE2-WC-004 · Сигналы вывода и извлечения модели
AISAFE2-WEAK-018, AISAFE2-WEAK-019

AISAFE2-WC-005 · Выбор и допуск модели
AISAFE2-WEAK-042, AISAFE2-WEAK-017

AISAFE2-WC-007 · Типизация доверия в контексте
AISAFE2-WEAK-007

AISAFE2-WC-008 · Системные инструкции и детерминированные контроли
AISAFE2-WEAK-055, AISAFE2-WEAK-043, AISAFE2-WEAK-074

AISAFE2-WC-010 · Контроль доступа и авторизация
AISAFE2-WEAK-089, AISAFE2-WEAK-041, AISAFE2-WEAK-022, AISAFE2-WEAK-016, AISAFE2-WEAK-086

AISAFE2-WC-011 · Передача идентичности и A2A-доверие
AISAFE2-WEAK-058, AISAFE2-WEAK-095, AISAFE2-WEAK-092, AISAFE2-WEAK-049

AISAFE2-WC-013 · Проверка вывода и контекстная изоляция
AISAFE2-WEAK-003, AISAFE2-WEAK-080, AISAFE2-WEAK-004

AISAFE2-WC-014 · Контракты вызова инструментов
AISAFE2-WEAK-056

AISAFE2-WC-016 · Управление и происхождение данных
AISAFE2-WEAK-035, AISAFE2-WEAK-087, AISAFE2-WEAK-013, AISAFE2-WEAK-086

AISAFE2-WC-017 · Контроль допуска источников
AISAFE2-WEAK-014

AISAFE2-WC-018 · Минимизация данных и приватность
AISAFE2-WEAK-039, AISAFE2-WEAK-087, AISAFE2-WEAK-040, AISAFE2-WEAK-036, AISAFE2-WEAK-033

AISAFE2-WC-019 · Изоляция данных
AISAFE2-WEAK-059, AISAFE2-WEAK-050

AISAFE2-WC-020 · Лимиты, ресурсы и контроль затрат
AISAFE2-WEAK-075

AISAFE2-WC-021 · Ограничение частоты, ресурсов и стоимости
AISAFE2-WEAK-052, AISAFE2-WEAK-026, AISAFE2-WEAK-062, AISAFE2-WEAK-027

AISAFE2-WC-022 · Пользовательский контроль тарифицируемых операций
AISAFE2-WEAK-093, AISAFE2-WEAK-063

AISAFE2-WC-024 · Доверие человека и прозрачность интерфейса
AISAFE2-WEAK-001, AISAFE2-WEAK-005, AISAFE2-WEAK-073, AISAFE2-WEAK-002, AISAFE2-WEAK-006

AISAFE2-WC-025 · Подтверждение и пропускная способность HITL
AISAFE2-WEAK-065, AISAFE2-WEAK-067, AISAFE2-WEAK-066, AISAFE2-WEAK-068, AISAFE2-WEAK-069, AISAFE2-WEAK-057

AISAFE2-WC-027 · Криптографическая защита хранения и каналов
AISAFE2-WEAK-023, AISAFE2-WEAK-024

AISAFE2-WC-028 · Управление ключами и секретами
AISAFE2-WEAK-094, AISAFE2-WEAK-025, AISAFE2-WEAK-041

AISAFE2-WC-030 · Безопасная обработка ошибок и деградация
AISAFE2-WEAK-051, AISAFE2-WEAK-064, AISAFE2-WEAK-060

AISAFE2-WC-031 · Готовность реагирования и восстановления
AISAFE2-WEAK-048, AISAFE2-WEAK-085

AISAFE2-WC-032 · Цепочка поставки, недоверенная среда и физическая защита
AISAFE2-WEAK-091, AISAFE2-WEAK-031, AISAFE2-WEAK-032, AISAFE2-WEAK-012

AISAFE2-WC-033 · Наблюдаемость, обнаружение и доказательства
AISAFE2-WEAK-061, AISAFE2-WEAK-028, AISAFE2-WEAK-030

AISAFE2-WC-035 · Безопасный выпуск и управление уязвимостями
AISAFE2-WEAK-044, AISAFE2-WEAK-047, AISAFE2-WEAK-084

AISAFE2-WC-036 · Соответствие и допустимое использование
AISAFE2-WEAK-046, AISAFE2-WEAK-088, AISAFE2-WEAK-070

AISAFE2-WC-037 · Закрытие прав и возможностей жизненного цикла
AISAFE2-WEAK-053, AISAFE2-WEAK-045

AISAFE2-WC-038 · Доверенные зоны и потоки данных
AISAFE2-WEAK-083

AISAFE2-WC-039 · Модель зон и межзонное принуждение
AISAFE2-WEAK-072, AISAFE2-WEAK-071

AISAFE2-WC-040 · Производительность, ёмкость и управление перегрузкой
AISAFE2-WEAK-078, AISAFE2-WEAK-077, AISAFE2-WEAK-076

AISAFE2-WC-041 · Зависимость от поставщика и переносимость
AISAFE2-WEAK-079

AISAFE2-WC-042 · Аутентификация субъектов
AISAFE2-WEAK-081, AISAFE2-WEAK-082

AISAFE2-WC-043 · Изоляция среды исполнения
AISAFE2-WEAK-090

Предметный указатель

Указатели помогают найти карточки по компоненту, классификационному признаку, области применения, сопоставлению или сервису. Для компонентов перечислены только карточки с прямой связью; многошаговые пути и условия связей приведены в самих карточках и реестре связей. Число рядом с ID — страница карточки.

По компонентам архитектуры

AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения

Воздействия (12): AISAFE2-IMP-001 92 · AISAFE2-IMP-002 93 · AISAFE2-IMP-003 95 · AISAFE2-IMP-004 97 · AISAFE2-IMP-006 100 · AISAFE2-IMP-009 106 · AISAFE2-IMP-010 107 · AISAFE2-IMP-014 114 · AISAFE2-IMP-020 125 · AISAFE2-IMP-022 127 · AISAFE2-IMP-024 130 · AISAFE2-IMP-032 139

Недостатки (14): AISAFE2-WEAK-001 143 · AISAFE2-WEAK-002 145 · AISAFE2-WEAK-003 146 · AISAFE2-WEAK-004 148 · AISAFE2-WEAK-005 151 · AISAFE2-WEAK-006 152 · AISAFE2-WEAK-026 194 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-057 248 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-063 260 · AISAFE2-WEAK-073 274 · AISAFE2-WEAK-078 282 · AISAFE2-WEAK-081 289

Меры защиты (11): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-005 331 · AISAFE2-MIT-038 462 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-056 529 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-063 550 · AISAFE2-MIT-071 577

AISAFE2-COMP-002 · Модель и интерфейс инференса

Воздействия (7): AISAFE2-IMP-006 100 · AISAFE2-IMP-012 111 · AISAFE2-IMP-013 113 · AISAFE2-IMP-014 114 · AISAFE2-IMP-017 120 · AISAFE2-IMP-022 127 · AISAFE2-IMP-032 139

Недостатки (22): AISAFE2-WEAK-008 158 · AISAFE2-WEAK-009 160 · AISAFE2-WEAK-010 162 · AISAFE2-WEAK-011 164 · AISAFE2-WEAK-015 175 · AISAFE2-WEAK-018 180 · AISAFE2-WEAK-019 182 · AISAFE2-WEAK-026 194 · AISAFE2-WEAK-028 199 · AISAFE2-WEAK-030 201 · AISAFE2-WEAK-033 206 · AISAFE2-WEAK-042 217 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-054 242 · AISAFE2-WEAK-059 253 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-074 276 · AISAFE2-WEAK-076 280 · AISAFE2-WEAK-080 286 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-091 308

Меры защиты (17): AISAFE2-MIT-002 320 · AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-014 368 · AISAFE2-MIT-024 410 · AISAFE2-MIT-034 448 · AISAFE2-MIT-036 456 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-042 478 · AISAFE2-MIT-044 485 · AISAFE2-MIT-046 492 · AISAFE2-MIT-047 494 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573

AISAFE2-COMP-003 · Управление контекстом общения с моделью

Воздействия (5): AISAFE2-IMP-004 97 · AISAFE2-IMP-010 107 · AISAFE2-IMP-014 114 · AISAFE2-IMP-024 130 · AISAFE2-IMP-032 139

Недостатки (7): AISAFE2-WEAK-007 154 · AISAFE2-WEAK-010 162 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-043 220 · AISAFE2-WEAK-074 276

Меры защиты (7): AISAFE2-MIT-002 320 · AISAFE2-MIT-003 324 · AISAFE2-MIT-004 328 · AISAFE2-MIT-010 349 · AISAFE2-MIT-015 371 · AISAFE2-MIT-054 519 · AISAFE2-MIT-066 561

AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)

Воздействия (7): AISAFE2-IMP-007 101 · AISAFE2-IMP-016 118 · AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-019 123 · AISAFE2-IMP-021 126 · AISAFE2-IMP-032 139

Недостатки (7): AISAFE2-WEAK-052 237 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-055 245 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-089 303 · AISAFE2-WEAK-092 310

Меры защиты (15): AISAFE2-MIT-003 324 · AISAFE2-MIT-006 333 · AISAFE2-MIT-016 375 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-029 429 · AISAFE2-MIT-048 497 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-066 561 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы

Воздействия (11): AISAFE2-IMP-004 97 · AISAFE2-IMP-007 101 · AISAFE2-IMP-008 104 · AISAFE2-IMP-009 106 · AISAFE2-IMP-010 107 · AISAFE2-IMP-016 118 · AISAFE2-IMP-018 121 · AISAFE2-IMP-021 126 · AISAFE2-IMP-022 127 · AISAFE2-IMP-024 130 · AISAFE2-IMP-032 139

Недостатки (18): AISAFE2-WEAK-003 146 · AISAFE2-WEAK-004 148 · AISAFE2-WEAK-007 154 · AISAFE2-WEAK-012 167 · AISAFE2-WEAK-024 188 · AISAFE2-WEAK-039 212 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-056 246 · AISAFE2-WEAK-058 251 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-064 262 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-089 303 · AISAFE2-WEAK-090 306 · AISAFE2-WEAK-092 310 · AISAFE2-WEAK-095 314

Меры защиты (27): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-013 364 · AISAFE2-MIT-015 371 · AISAFE2-MIT-016 375 · AISAFE2-MIT-017 379 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-031 438 · AISAFE2-MIT-032 441 · AISAFE2-MIT-035 451 · AISAFE2-MIT-046 492 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-057 531 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

AISAFE2-COMP-006 · RAG и извлечение из базы знаний

Воздействия (6): AISAFE2-IMP-008 104 · AISAFE2-IMP-009 106 · AISAFE2-IMP-011 109 · AISAFE2-IMP-014 114 · AISAFE2-IMP-022 127 · AISAFE2-IMP-032 139

Недостатки (4): AISAFE2-WEAK-002 145 · AISAFE2-WEAK-007 154 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-087 299

Меры защиты (8): AISAFE2-MIT-005 331 · AISAFE2-MIT-008 342 · AISAFE2-MIT-036 456 · AISAFE2-MIT-046 492 · AISAFE2-MIT-054 519 · AISAFE2-MIT-063 550 · AISAFE2-MIT-069 571 · AISAFE2-MIT-070 573

AISAFE2-COMP-007 · Долговременная память

Воздействия (5): AISAFE2-IMP-008 104 · AISAFE2-IMP-009 106 · AISAFE2-IMP-015 116 · AISAFE2-IMP-022 127 · AISAFE2-IMP-032 139

Недостатки (6): AISAFE2-WEAK-007 154 · AISAFE2-WEAK-014 173 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-053 240

Меры защиты (9): AISAFE2-MIT-010 349 · AISAFE2-MIT-021 396 · AISAFE2-MIT-024 410 · AISAFE2-MIT-033 444 · AISAFE2-MIT-046 492 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547

AISAFE2-COMP-008 · Данные для подсистемы обучения

Воздействия (3): AISAFE2-IMP-013 113 · AISAFE2-IMP-015 116 · AISAFE2-IMP-031 137

Недостатки (14): AISAFE2-WEAK-012 167 · AISAFE2-WEAK-013 171 · AISAFE2-WEAK-014 173 · AISAFE2-WEAK-015 175 · AISAFE2-WEAK-016 176 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-033 206 · AISAFE2-WEAK-035 207 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-037 211 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-059 253 · AISAFE2-WEAK-085 296

Меры защиты (15): AISAFE2-MIT-010 349 · AISAFE2-MIT-024 410 · AISAFE2-MIT-027 423 · AISAFE2-MIT-035 451 · AISAFE2-MIT-036 456 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-042 478 · AISAFE2-MIT-050 503 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-068 568

AISAFE2-COMP-009 · Конвейер обучения и дообучения

Воздействия (1): AISAFE2-IMP-013 113

Недостатки (11): AISAFE2-WEAK-011 164 · AISAFE2-WEAK-012 167 · AISAFE2-WEAK-015 175 · AISAFE2-WEAK-016 176 · AISAFE2-WEAK-033 206 · AISAFE2-WEAK-037 211 · AISAFE2-WEAK-045 223 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-059 253 · AISAFE2-WEAK-091 308

Меры защиты (13): AISAFE2-MIT-010 349 · AISAFE2-MIT-013 364 · AISAFE2-MIT-018 383 · AISAFE2-MIT-026 419 · AISAFE2-MIT-035 451 · AISAFE2-MIT-036 456 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-042 478 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-060 541 · AISAFE2-MIT-064 554

AISAFE2-COMP-010 · Хранилище и реестр артефактов модели

Воздействия (5): AISAFE2-IMP-012 111 · AISAFE2-IMP-013 113 · AISAFE2-IMP-014 114 · AISAFE2-IMP-015 116 · AISAFE2-IMP-033 141

Недостатки (15): AISAFE2-WEAK-012 167 · AISAFE2-WEAK-013 171 · AISAFE2-WEAK-016 176 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-024 188 · AISAFE2-WEAK-030 201 · AISAFE2-WEAK-031 203 · AISAFE2-WEAK-032 204 · AISAFE2-WEAK-045 223 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-085 296 · AISAFE2-WEAK-091 308

Меры защиты (13): AISAFE2-MIT-012 359 · AISAFE2-MIT-013 364 · AISAFE2-MIT-014 368 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-043 481 · AISAFE2-MIT-052 509 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568

AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения

Воздействия (5): AISAFE2-IMP-012 111 · AISAFE2-IMP-016 118 · AISAFE2-IMP-019 123 · AISAFE2-IMP-022 127 · AISAFE2-IMP-032 139

Недостатки (19): AISAFE2-WEAK-004 148 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-024 188 · AISAFE2-WEAK-031 203 · AISAFE2-WEAK-032 204 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-050 234 · AISAFE2-WEAK-052 237 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-075 277 · AISAFE2-WEAK-076 280 · AISAFE2-WEAK-077 281 · AISAFE2-WEAK-078 282 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-090 306

Меры защиты (20): AISAFE2-MIT-006 333 · AISAFE2-MIT-012 359 · AISAFE2-MIT-017 379 · AISAFE2-MIT-018 383 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419 · AISAFE2-MIT-031 438 · AISAFE2-MIT-041 473 · AISAFE2-MIT-043 481 · AISAFE2-MIT-046 492 · AISAFE2-MIT-049 500 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-057 531 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573

AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты

Воздействия (5): AISAFE2-IMP-007 101 · AISAFE2-IMP-011 109 · AISAFE2-IMP-012 111 · AISAFE2-IMP-015 116 · AISAFE2-IMP-016 118

Недостатки (11): AISAFE2-WEAK-016 176 · AISAFE2-WEAK-022 184 · AISAFE2-WEAK-025 190 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-045 223 · AISAFE2-WEAK-049 232 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-058 251 · AISAFE2-WEAK-082 290 · AISAFE2-WEAK-092 310 · AISAFE2-WEAK-094 312

Меры защиты (21): AISAFE2-MIT-008 342 · AISAFE2-MIT-010 349 · AISAFE2-MIT-011 356 · AISAFE2-MIT-012 359 · AISAFE2-MIT-015 371 · AISAFE2-MIT-018 383 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-031 438 · AISAFE2-MIT-041 473 · AISAFE2-MIT-043 481 · AISAFE2-MIT-048 497 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-055 525 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-064 554 · AISAFE2-MIT-070 573

AISAFE2-COMP-013 · Телеметрия и обнаружение

Воздействия (2): AISAFE2-IMP-009 106 · AISAFE2-IMP-015 116

Недостатки (7): AISAFE2-WEAK-023 186 · AISAFE2-WEAK-028 199 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-061 255 · AISAFE2-WEAK-084 294

Меры защиты (16): AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-014 368 · AISAFE2-MIT-022 400 · AISAFE2-MIT-023 405 · AISAFE2-MIT-024 410 · AISAFE2-MIT-036 456 · AISAFE2-MIT-038 462 · AISAFE2-MIT-044 485 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-065 558 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573

AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения

Воздействия (3): AISAFE2-IMP-001 92 · AISAFE2-IMP-002 93 · AISAFE2-IMP-023 129

Недостатки (6): AISAFE2-WEAK-001 143 · AISAFE2-WEAK-057 248 · AISAFE2-WEAK-065 265 · AISAFE2-WEAK-067 267 · AISAFE2-WEAK-068 268 · AISAFE2-WEAK-069 269

Меры защиты (3): AISAFE2-MIT-019 389 · AISAFE2-MIT-037 459 · AISAFE2-MIT-045 488

AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска

Воздействия (1): AISAFE2-IMP-033 141

Недостатки (13): AISAFE2-WEAK-012 167 · AISAFE2-WEAK-017 179 · AISAFE2-WEAK-035 207 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-042 217 · AISAFE2-WEAK-044 222 · AISAFE2-WEAK-045 223 · AISAFE2-WEAK-046 225 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-085 296 · AISAFE2-WEAK-088 301

Меры защиты (20): AISAFE2-MIT-013 364 · AISAFE2-MIT-015 371 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-039 466 · AISAFE2-MIT-042 478 · AISAFE2-MIT-045 488 · AISAFE2-MIT-047 494 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-055 525 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса

Воздействия (10): AISAFE2-IMP-004 97 · AISAFE2-IMP-006 100 · AISAFE2-IMP-009 106 · AISAFE2-IMP-012 111 · AISAFE2-IMP-013 113 · AISAFE2-IMP-016 118 · AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-022 127 · AISAFE2-IMP-024 130

Недостатки (11): AISAFE2-WEAK-024 188 · AISAFE2-WEAK-026 194 · AISAFE2-WEAK-028 199 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-052 237 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-064 262 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-087 299

Меры защиты (19): AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-014 368 · AISAFE2-MIT-018 383 · AISAFE2-MIT-024 410 · AISAFE2-MIT-030 434 · AISAFE2-MIT-032 441 · AISAFE2-MIT-036 456 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-042 478 · AISAFE2-MIT-044 485 · AISAFE2-MIT-046 492 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-057 531 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-069 571

AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика

Воздействия (16): AISAFE2-IMP-001 92 · AISAFE2-IMP-002 93 · AISAFE2-IMP-004 97 · AISAFE2-IMP-006 100 · AISAFE2-IMP-007 101 · AISAFE2-IMP-008 104 · AISAFE2-IMP-009 106 · AISAFE2-IMP-014 114 · AISAFE2-IMP-016 118 · AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-019 123 · AISAFE2-IMP-020 125 · AISAFE2-IMP-021 126 · AISAFE2-IMP-022 127 · AISAFE2-IMP-032 139

Недостатки (15): AISAFE2-WEAK-003 146 · AISAFE2-WEAK-004 148 · AISAFE2-WEAK-051 235 · AISAFE2-WEAK-052 237 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-055 245 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-063 260 · AISAFE2-WEAK-064 262 · AISAFE2-WEAK-073 274 · AISAFE2-WEAK-078 282 · AISAFE2-WEAK-081 289 · AISAFE2-WEAK-089 303 · AISAFE2-WEAK-093 311

Меры защиты (20): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-016 375 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419 · AISAFE2-MIT-029 429 · AISAFE2-MIT-038 462 · AISAFE2-MIT-045 488 · AISAFE2-MIT-046 492 · AISAFE2-MIT-051 507 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-058 535 · AISAFE2-MIT-066 561 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

AISAFE2-COMP-018 · Операционные данные, очереди и кэши

Воздействия (8): AISAFE2-IMP-008 104 · AISAFE2-IMP-009 106 · AISAFE2-IMP-010 107 · AISAFE2-IMP-014 114 · AISAFE2-IMP-015 116 · AISAFE2-IMP-019 123 · AISAFE2-IMP-022 127 · AISAFE2-IMP-032 139

Недостатки (8): AISAFE2-WEAK-023 186 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-052 237 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-076 280 · AISAFE2-WEAK-078 282

Меры защиты (12): AISAFE2-MIT-006 333 · AISAFE2-MIT-008 342 · AISAFE2-MIT-010 349 · AISAFE2-MIT-015 371 · AISAFE2-MIT-024 410 · AISAFE2-MIT-046 492 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-065 558

AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов

Воздействия (4): AISAFE2-IMP-007 101 · AISAFE2-IMP-016 118 · AISAFE2-IMP-022 127 · AISAFE2-IMP-024 130

Недостатки (10): AISAFE2-WEAK-003 146 · AISAFE2-WEAK-026 194 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-051 235 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-072 273 · AISAFE2-WEAK-080 286 · AISAFE2-WEAK-089 303

Меры защиты (20): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-004 328 · AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-016 375 · AISAFE2-MIT-019 389 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419 · AISAFE2-MIT-029 429 · AISAFE2-MIT-032 441 · AISAFE2-MIT-038 462 · AISAFE2-MIT-044 485 · AISAFE2-MIT-051 507 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-056 529 · AISAFE2-MIT-066 561 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

AISAFE2-COMP-020 · Контур управления внешним AI-провайдером

Воздействия (7): AISAFE2-IMP-012 111 · AISAFE2-IMP-015 116 · AISAFE2-IMP-016 118 · AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-020 125 · AISAFE2-IMP-022 127

Недостатки (7): AISAFE2-WEAK-045 223 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-075 277 · AISAFE2-WEAK-077 281 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-082 290

Меры защиты (13): AISAFE2-MIT-006 333 · AISAFE2-MIT-014 368 · AISAFE2-MIT-018 383 · AISAFE2-MIT-024 410 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-042 478 · AISAFE2-MIT-044 485 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-054 519 · AISAFE2-MIT-059 538 · AISAFE2-MIT-064 554

AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов

Воздействия (7): AISAFE2-IMP-007 101 · AISAFE2-IMP-008 104 · AISAFE2-IMP-016 118 · AISAFE2-IMP-021 126 · AISAFE2-IMP-022 127 · AISAFE2-IMP-024 130 · AISAFE2-IMP-032 139

Недостатки (3): AISAFE2-WEAK-024 188 · AISAFE2-WEAK-049 232 · AISAFE2-WEAK-053 240

Меры защиты (14): AISAFE2-MIT-001 316 · AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-018 383 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419 · AISAFE2-MIT-046 492 · AISAFE2-MIT-048 497 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-063 550 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

AISAFE2-COMP-022 · Программные артефакты и цепочка поставки

Воздействия (2): AISAFE2-IMP-014 114 · AISAFE2-IMP-015 116

Недостатки (3): AISAFE2-WEAK-012 167 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-053 240

Меры защиты (6): AISAFE2-MIT-013 364 · AISAFE2-MIT-015 371 · AISAFE2-MIT-026 419 · AISAFE2-MIT-035 451 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514

AISAFE2-COMP-023 · Подсистема формирования базы знаний

Воздействия (3): AISAFE2-IMP-011 109 · AISAFE2-IMP-015 116 · AISAFE2-IMP-033 141

Недостатки (10): AISAFE2-WEAK-013 171 · AISAFE2-WEAK-014 173 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-085 296 · AISAFE2-WEAK-086 298 · AISAFE2-WEAK-087 299

Меры защиты (17): AISAFE2-MIT-007 339 · AISAFE2-MIT-009 346 · AISAFE2-MIT-010 349 · AISAFE2-MIT-011 356 · AISAFE2-MIT-018 383 · AISAFE2-MIT-024 410 · AISAFE2-MIT-033 444 · AISAFE2-MIT-036 456 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-MIT-069 571

AISAFE2-COMP-024 · Артефакты работы AI-системы

Воздействия (1): AISAFE2-IMP-032 139

Недостатки (1): AISAFE2-WEAK-003 146

Меры защиты (3): AISAFE2-MIT-002 320 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547

AISAFE2-COMP-025 · Источники данных для инструментов

Воздействия (1): AISAFE2-IMP-015 116

Недостатки (2): AISAFE2-WEAK-013 171 · AISAFE2-WEAK-035 207

Меры защиты (2): AISAFE2-MIT-010 349 · AISAFE2-MIT-069 571

По классификационным признакам

Каждый признак относится только к указанному типу карточек, как соответствующий фильтр на сайте.

Воздействия · Категория воздействия

Высокоуровневые бизнес-последствия (6): AISAFE2-IMP-025 133 · AISAFE2-IMP-026 134 · AISAFE2-IMP-027 134 · AISAFE2-IMP-028 135 · AISAFE2-IMP-029 136 · AISAFE2-IMP-030 136

I.1 Негативные последствия для потребителя за счет использования результатов AI-системы (5): AISAFE2-IMP-001 92 · AISAFE2-IMP-002 93 · AISAFE2-IMP-003 95 · AISAFE2-IMP-004 97 · AISAFE2-IMP-005 98

I.2 Нарушение предусмотренных сценариев работы AI-системы (3): AISAFE2-IMP-006 100 · AISAFE2-IMP-007 101 · AISAFE2-IMP-008 104

I.3 Раскрытие чувствительных данных (6): AISAFE2-IMP-009 106 · AISAFE2-IMP-010 107 · AISAFE2-IMP-011 109 · AISAFE2-IMP-012 111 · AISAFE2-IMP-013 113 · AISAFE2-IMP-014 114

I.4 Компрометация компонента системы, получения контроля над ним (5): AISAFE2-IMP-015 116 · AISAFE2-IMP-016 118 · AISAFE2-IMP-031 137 · AISAFE2-IMP-032 139 · AISAFE2-IMP-033 141

I.5 Повышенное потребление (ресурсов, токенов) (4): AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-019 123 · AISAFE2-IMP-020 125

I.6 Отказ в обслуживании (3): AISAFE2-IMP-021 126 · AISAFE2-IMP-022 127 · AISAFE2-IMP-023 129

I.7 Утечка чувствительных данных в недоверенный контур (2): AISAFE2-IMP-024 130 · AISAFE2-IMP-032 139

Воздействия · Область последствий

Информационные активы (16): AISAFE2-IMP-004 97 · AISAFE2-IMP-006 100 · AISAFE2-IMP-007 101 · AISAFE2-IMP-008 104 · AISAFE2-IMP-009 106 · AISAFE2-IMP-010 107 · AISAFE2-IMP-011 109 · AISAFE2-IMP-012 111 · AISAFE2-IMP-013 113 · AISAFE2-IMP-014 114 · AISAFE2-IMP-015 116 · AISAFE2-IMP-016 118 · AISAFE2-IMP-024 130 · AISAFE2-IMP-031 137 · AISAFE2-IMP-032 139 · AISAFE2-IMP-033 141

Конкурентная позиция (1): AISAFE2-IMP-030 136

Организация (2): AISAFE2-IMP-002 93 · AISAFE2-IMP-026 134

Правовые и нормативные обязательства (1): AISAFE2-IMP-027 134

Работа сервиса (12): AISAFE2-IMP-007 101 · AISAFE2-IMP-008 104 · AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-019 123 · AISAFE2-IMP-020 125 · AISAFE2-IMP-021 126 · AISAFE2-IMP-022 127 · AISAFE2-IMP-023 129 · AISAFE2-IMP-026 134 · AISAFE2-IMP-031 137 · AISAFE2-IMP-033 141

Репутация (1): AISAFE2-IMP-029 136

Финансы (9): AISAFE2-IMP-001 92 · AISAFE2-IMP-002 93 · AISAFE2-IMP-003 95 · AISAFE2-IMP-017 120 · AISAFE2-IMP-018 121 · AISAFE2-IMP-019 123 · AISAFE2-IMP-020 125 · AISAFE2-IMP-025 133 · AISAFE2-IMP-030 136

Человек (6): AISAFE2-IMP-001 92 · AISAFE2-IMP-002 93 · AISAFE2-IMP-003 95 · AISAFE2-IMP-004 97 · AISAFE2-IMP-005 98 · AISAFE2-IMP-023 129

Этические последствия (2): AISAFE2-IMP-005 98 · AISAFE2-IMP-028 135

Недостатки · Источник недостатка

Проектирование (42): AISAFE2-WEAK-001 143 · AISAFE2-WEAK-002 145 · AISAFE2-WEAK-006 152 · AISAFE2-WEAK-007 154 · AISAFE2-WEAK-008 158 · AISAFE2-WEAK-009 160 · AISAFE2-WEAK-010 162 · AISAFE2-WEAK-015 175 · AISAFE2-WEAK-018 180 · AISAFE2-WEAK-019 182 · AISAFE2-WEAK-022 184 · AISAFE2-WEAK-027 196 · AISAFE2-WEAK-031 203 · AISAFE2-WEAK-033 206 · AISAFE2-WEAK-042 217 · AISAFE2-WEAK-054 242 · AISAFE2-WEAK-055 245 · AISAFE2-WEAK-059 253 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-063 260 · AISAFE2-WEAK-065 265 · AISAFE2-WEAK-069 269 · AISAFE2-WEAK-071 271 · AISAFE2-WEAK-073 274 · AISAFE2-WEAK-074 276 · AISAFE2-WEAK-075 277 · AISAFE2-WEAK-076 280 · AISAFE2-WEAK-078 282 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-080 286 · AISAFE2-WEAK-081 289 · AISAFE2-WEAK-082 290 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-086 298 · AISAFE2-WEAK-087 299 · AISAFE2-WEAK-089 303 · AISAFE2-WEAK-090 306 · AISAFE2-WEAK-091 308 · AISAFE2-WEAK-092 310 · AISAFE2-WEAK-093 311 · AISAFE2-WEAK-094 312 · AISAFE2-WEAK-095 314

Процесс (20): AISAFE2-WEAK-011 164 · AISAFE2-WEAK-012 167 · AISAFE2-WEAK-013 171 · AISAFE2-WEAK-014 173 · AISAFE2-WEAK-017 179 · AISAFE2-WEAK-035 207 · AISAFE2-WEAK-044 222 · AISAFE2-WEAK-045 223 · AISAFE2-WEAK-046 225 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-048 229 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-066 266 · AISAFE2-WEAK-068 268 · AISAFE2-WEAK-070 270 · AISAFE2-WEAK-077 281 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-085 296 · AISAFE2-WEAK-088 301

Реализация (56): AISAFE2-WEAK-003 146 · AISAFE2-WEAK-004 148 · AISAFE2-WEAK-005 151 · AISAFE2-WEAK-009 160 · AISAFE2-WEAK-012 167 · AISAFE2-WEAK-016 176 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-024 188 · AISAFE2-WEAK-025 190 · AISAFE2-WEAK-026 194 · AISAFE2-WEAK-028 199 · AISAFE2-WEAK-030 201 · AISAFE2-WEAK-032 204 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-037 211 · AISAFE2-WEAK-039 212 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-043 220 · AISAFE2-WEAK-049 232 · AISAFE2-WEAK-050 234 · AISAFE2-WEAK-051 235 · AISAFE2-WEAK-052 237 · AISAFE2-WEAK-055 245 · AISAFE2-WEAK-056 246 · AISAFE2-WEAK-057 248 · AISAFE2-WEAK-058 251 · AISAFE2-WEAK-059 253 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-061 255 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-063 260 · AISAFE2-WEAK-064 262 · AISAFE2-WEAK-065 265 · AISAFE2-WEAK-067 267 · AISAFE2-WEAK-068 268 · AISAFE2-WEAK-069 269 · AISAFE2-WEAK-072 273 · AISAFE2-WEAK-073 274 · AISAFE2-WEAK-075 277 · AISAFE2-WEAK-076 280 · AISAFE2-WEAK-077 281 · AISAFE2-WEAK-078 282 · AISAFE2-WEAK-081 289 · AISAFE2-WEAK-082 290 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-085 296 · AISAFE2-WEAK-086 298 · AISAFE2-WEAK-089 303 · AISAFE2-WEAK-090 306 · AISAFE2-WEAK-091 308 · AISAFE2-WEAK-092 310 · AISAFE2-WEAK-093 311 · AISAFE2-WEAK-094 312 · AISAFE2-WEAK-095 314

Меры защиты · Механизм защиты

Авторизация (22): AISAFE2-MIT-008 342 · AISAFE2-MIT-011 356 · AISAFE2-MIT-016 375 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-021 396 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-032 441 · AISAFE2-MIT-043 481 · AISAFE2-MIT-054 519 · AISAFE2-MIT-055 525 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-066 561

Валидация (14): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-004 328 · AISAFE2-MIT-007 339 · AISAFE2-MIT-010 349 · AISAFE2-MIT-016 375 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-042 478 · AISAFE2-MIT-047 494 · AISAFE2-MIT-056 529 · AISAFE2-MIT-058 535 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Восстановление (6): AISAFE2-MIT-009 346 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-046 492 · AISAFE2-MIT-050 503 · AISAFE2-MIT-068 568

Жизненный цикл (16): AISAFE2-MIT-009 346 · AISAFE2-MIT-018 383 · AISAFE2-MIT-021 396 · AISAFE2-MIT-026 419 · AISAFE2-MIT-034 448 · AISAFE2-MIT-036 456 · AISAFE2-MIT-046 492 · AISAFE2-MIT-049 500 · AISAFE2-MIT-052 509 · AISAFE2-MIT-060 541 · AISAFE2-MIT-062 547 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Изоляция (10): AISAFE2-MIT-003 324 · AISAFE2-MIT-011 356 · AISAFE2-MIT-017 379 · AISAFE2-MIT-031 438 · AISAFE2-MIT-041 473 · AISAFE2-MIT-043 481 · AISAFE2-MIT-049 500 · AISAFE2-MIT-057 531 · AISAFE2-MIT-061 544 · AISAFE2-MIT-068 568

Контроль ресурсов (3): AISAFE2-MIT-006 333 · AISAFE2-MIT-038 462 · AISAFE2-MIT-044 485

Минимизация данных (4): AISAFE2-MIT-010 349 · AISAFE2-MIT-032 441 · AISAFE2-MIT-042 478 · AISAFE2-MIT-064 554

Наблюдаемость (6): AISAFE2-MIT-005 331 · AISAFE2-MIT-022 400 · AISAFE2-MIT-028 426 · AISAFE2-MIT-035 451 · AISAFE2-MIT-053 514 · AISAFE2-MIT-065 558

Обнаружение (6): AISAFE2-MIT-004 328 · AISAFE2-MIT-023 405 · AISAFE2-MIT-036 456 · AISAFE2-MIT-038 462 · AISAFE2-MIT-044 485 · AISAFE2-MIT-056 529

Реагирование (3): AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-026 419

Управление и надзор (5): AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-053 514 · AISAFE2-MIT-055 525 · AISAFE2-MIT-069 571

Управление конфигурацией (13): AISAFE2-MIT-003 324 · AISAFE2-MIT-014 368 · AISAFE2-MIT-015 371 · AISAFE2-MIT-054 519 · AISAFE2-MIT-056 529 · AISAFE2-MIT-057 531 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-066 561 · AISAFE2-MIT-067 564 · AISAFE2-MIT-069 571

Целостность (15): AISAFE2-MIT-007 339 · AISAFE2-MIT-009 346 · AISAFE2-MIT-012 359 · AISAFE2-MIT-013 364 · AISAFE2-MIT-033 444 · AISAFE2-MIT-035 451 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-048 497 · AISAFE2-MIT-050 503 · AISAFE2-MIT-060 541 · AISAFE2-MIT-063 550 · AISAFE2-MIT-065 558 · AISAFE2-MIT-066 561 · AISAFE2-MIT-068 568

Человеческий контроль (4): AISAFE2-MIT-005 331 · AISAFE2-MIT-019 389 · AISAFE2-MIT-037 459 · AISAFE2-MIT-045 488

Шлюз допуска (14): AISAFE2-MIT-013 364 · AISAFE2-MIT-014 368 · AISAFE2-MIT-025 415 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-039 466 · AISAFE2-MIT-051 507 · AISAFE2-MIT-053 514 · AISAFE2-MIT-060 541 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Меры защиты · Форма меры защиты

Архитектурная (15): AISAFE2-MIT-003 324 · AISAFE2-MIT-011 356 · AISAFE2-MIT-017 379 · AISAFE2-MIT-031 438 · AISAFE2-MIT-041 473 · AISAFE2-MIT-049 500 · AISAFE2-MIT-053 514 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-066 561 · AISAFE2-MIT-069 571

Выполняемая человеком (5): AISAFE2-MIT-019 389 · AISAFE2-MIT-037 459 · AISAFE2-MIT-045 488 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Договорная (2): AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519

Организационная (12): AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-037 459 · AISAFE2-MIT-038 462 · AISAFE2-MIT-045 488 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-054 519 · AISAFE2-MIT-055 525 · AISAFE2-MIT-065 558

Процессная (42): AISAFE2-MIT-005 331 · AISAFE2-MIT-007 339 · AISAFE2-MIT-009 346 · AISAFE2-MIT-013 364 · AISAFE2-MIT-014 368 · AISAFE2-MIT-018 383 · AISAFE2-MIT-022 400 · AISAFE2-MIT-024 410 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-036 456 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-042 478 · AISAFE2-MIT-043 481 · AISAFE2-MIT-044 485 · AISAFE2-MIT-045 488 · AISAFE2-MIT-046 492 · AISAFE2-MIT-050 503 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-054 519 ·

AISAFE2-MIT-055 525 · AISAFE2-MIT-056 529 · AISAFE2-MIT-059 538 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-MIT-069 571 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Техническая (67): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-003 324 · AISAFE2-MIT-004 328 · AISAFE2-MIT-005 331 · AISAFE2-MIT-006 333 · AISAFE2-MIT-007 339 · AISAFE2-MIT-008 342 · AISAFE2-MIT-009 346 · AISAFE2-MIT-010 349 · AISAFE2-MIT-011 356 · AISAFE2-MIT-012 359 · AISAFE2-MIT-013 364 · AISAFE2-MIT-014 368 · AISAFE2-MIT-015 371 · AISAFE2-MIT-016 375 · AISAFE2-MIT-017 379 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-021 396 · AISAFE2-MIT-022 400 · AISAFE2-MIT-023 405 · AISAFE2-MIT-024 410 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-031 438 · AISAFE2-MIT-032 441 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-036 456 · AISAFE2-MIT-037 459 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-042 478 · AISAFE2-MIT-043 481 · AISAFE2-MIT-044 485 · AISAFE2-MIT-045 488 · AISAFE2-MIT-046 492 · AISAFE2-MIT-047 494 · AISAFE2-MIT-048 497 · AISAFE2-MIT-049 500 · AISAFE2-MIT-050 503 · AISAFE2-MIT-051 507 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-056 529 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-066 561 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-MIT-069 571 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Меры защиты · Ответственность за меру

Заказчик (70): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-003 324 · AISAFE2-MIT-004 328 · AISAFE2-MIT-005 331 · AISAFE2-MIT-006 333 · AISAFE2-MIT-007 339 · AISAFE2-MIT-008 342 · AISAFE2-MIT-009 346 · AISAFE2-MIT-010 349 · AISAFE2-MIT-011 356 · AISAFE2-MIT-012 359 · AISAFE2-MIT-013 364 · AISAFE2-MIT-014 368 · AISAFE2-MIT-015 371 · AISAFE2-MIT-016 375 · AISAFE2-MIT-017 379 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-022 400 · AISAFE2-MIT-023 405 · AISAFE2-MIT-024 410 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-031 438 · AISAFE2-MIT-032 441 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-036 456 · AISAFE2-MIT-037 459 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-042 478 · AISAFE2-MIT-043 481 · AISAFE2-MIT-044 485 · AISAFE2-MIT-045 488 · AISAFE2-MIT-046 492 · AISAFE2-MIT-047 494 · AISAFE2-MIT-048 497 · AISAFE2-MIT-049 500 · AISAFE2-MIT-050 503 · AISAFE2-MIT-051 507 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-054 519 · AISAFE2-MIT-055 525 · AISAFE2-MIT-056 529 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-066 561 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-MIT-069 571 · AISAFE2-MIT-070 573 · AISAFE2-MIT-071 577

Поставщик (13): AISAFE2-MIT-049 500 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-066 561 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-MIT-069 571

Совместная ответственность (1): AISAFE2-MIT-021 396

Третья сторона (6): AISAFE2-MIT-014 368 · AISAFE2-MIT-024 410 · AISAFE2-MIT-027 423 · AISAFE2-MIT-042 478 · AISAFE2-MIT-046 492 · AISAFE2-MIT-052 509

Меры защиты · Режим отказа

Безопасная деградация (3): AISAFE2-MIT-004 328 · AISAFE2-MIT-036 456 · AISAFE2-MIT-046 492

Блокировка процессом (10): AISAFE2-MIT-007 339 · AISAFE2-MIT-023 405 · AISAFE2-MIT-024 410 · AISAFE2-MIT-041 473 · AISAFE2-MIT-045 488 · AISAFE2-MIT-053 514 · AISAFE2-MIT-055 525 · AISAFE2-MIT-065 558 · AISAFE2-MIT-067 564 · AISAFE2-MIT-070 573

Отказ с блокировкой (58): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-003 324 · AISAFE2-MIT-005 331 · AISAFE2-MIT-006 333 · AISAFE2-MIT-008 342 · AISAFE2-MIT-009 346 · AISAFE2-MIT-010 349 · AISAFE2-MIT-011 356 · AISAFE2-MIT-012 359 · AISAFE2-MIT-013 364 · AISAFE2-MIT-014 368 · AISAFE2-MIT-015 371 · AISAFE2-MIT-016 375 · AISAFE2-MIT-017 379 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-021 396 · AISAFE2-MIT-022 400 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-031 438 · AISAFE2-MIT-032 441 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-037 459 · AISAFE2-MIT-038 462 · AISAFE2-MIT-039 466 · AISAFE2-MIT-040 469 · AISAFE2-MIT-042 478 · AISAFE2-MIT-043 481 · AISAFE2-MIT-044 485 · AISAFE2-MIT-047 494 · AISAFE2-MIT-048 497 · AISAFE2-MIT-049 500 · AISAFE2-MIT-050 503 · AISAFE2-MIT-051 507 · AISAFE2-MIT-052 509 · AISAFE2-MIT-054 519 · AISAFE2-MIT-056 529 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-066 561 · AISAFE2-MIT-068 568 · AISAFE2-MIT-069 571 · AISAFE2-MIT-071 577

По области применения

Общая безопасность (74): AISAFE2-MIT-001 316 · AISAFE2-MIT-006 333 · AISAFE2-MIT-010 349 · AISAFE2-MIT-013 364 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-022 400 · AISAFE2-MIT-024 410 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-027 423 · AISAFE2-MIT-028 426 · AISAFE2-MIT-035 451 · AISAFE2-MIT-046 492 · AISAFE2-MIT-049 500 · AISAFE2-MIT-050 503 · AISAFE2-MIT-051 507 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-054 519 · AISAFE2-MIT-055 525 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-059 538 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-067 564 · AISAFE2-MIT-068 568 · AISAFE2-WEAK-004 148 · AISAFE2-WEAK-012 167 · AISAFE2-WEAK-013 171 · AISAFE2-WEAK-014 173 · AISAFE2-WEAK-016 176 · AISAFE2-WEAK-022 184 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-024 188 · AISAFE2-WEAK-025 190 · AISAFE2-WEAK-026 194 · AISAFE2-WEAK-028 199 · AISAFE2-WEAK-032 204 · AISAFE2-WEAK-035 207 · AISAFE2-WEAK-036 209 · AISAFE2-WEAK-040 214 · AISAFE2-WEAK-041 216 · AISAFE2-WEAK-044 222 · AISAFE2-WEAK-045 223 · AISAFE2-WEAK-046 225 · AISAFE2-WEAK-047 227 · AISAFE2-WEAK-048 229 · AISAFE2-WEAK-050 234 · AISAFE2-WEAK-051 235 · AISAFE2-WEAK-053 240 · AISAFE2-WEAK-057 248 · AISAFE2-WEAK-060 254 · AISAFE2-WEAK-061 255 · AISAFE2-WEAK-062 257 · AISAFE2-WEAK-063 260 · AISAFE2-WEAK-064 262 · AISAFE2-WEAK-071 271 · AISAFE2-WEAK-072 273 · AISAFE2-WEAK-075 277 · AISAFE2-WEAK-076 280 · AISAFE2-WEAK-077 281 · AISAFE2-WEAK-078 282 · AISAFE2-WEAK-079 284 · AISAFE2-WEAK-081 289 · AISAFE2-WEAK-082 290 · AISAFE2-WEAK-083 292 · AISAFE2-WEAK-084 294 · AISAFE2-WEAK-085 296

По сопоставлению со стандартом

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud (53): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-003 324 · AISAFE2-MIT-004 328 · AISAFE2-MIT-006 333 · AISAFE2-MIT-007 339 · AISAFE2-MIT-008 342 · AISAFE2-MIT-010 349 · AISAFE2-MIT-011 356 · AISAFE2-MIT-012 359 · AISAFE2-MIT-014 368 · AISAFE2-MIT-015 371 · AISAFE2-MIT-016 375 · AISAFE2-MIT-017 379 · AISAFE2-MIT-018 383 · AISAFE2-MIT-019 389 · AISAFE2-MIT-020 392 · AISAFE2-MIT-021 396 · AISAFE2-MIT-022 400 · AISAFE2-MIT-023 405 · AISAFE2-MIT-024 410 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-032 441 · AISAFE2-MIT-033 444 · AISAFE2-MIT-034 448 · AISAFE2-MIT-035 451 · AISAFE2-MIT-039 466 · AISAFE2-MIT-040 469 · AISAFE2-MIT-041 473 · AISAFE2-MIT-043 481 · AISAFE2-MIT-048 497 · AISAFE2-MIT-049 500 · AISAFE2-MIT-050 503 · AISAFE2-MIT-052 509 · AISAFE2-MIT-053 514 · AISAFE2-MIT-054 519 · AISAFE2-MIT-055 525 · AISAFE2-MIT-056 529 · AISAFE2-MIT-057 531 · AISAFE2-MIT-058 535 · AISAFE2-MIT-060 541 · AISAFE2-MIT-061 544 · AISAFE2-MIT-062 547 · AISAFE2-MIT-063 550 · AISAFE2-MIT-064 554 · AISAFE2-MIT-065 558 · AISAFE2-MIT-066 561 · AISAFE2-WEAK-023 186 · AISAFE2-WEAK-024 188 · AISAFE2-WEAK-025 190

По сервисам Yandex Cloud

Перечислены меры защиты с опубликованным сопоставлением сервиса. Наличие сервиса не подтверждает настройку или эффективность меры.

СолидЛаб — анализ защищённости систем с AI (1): AISAFE2-MIT-070 573

SourceCraft Security (4): AISAFE2-MIT-013 364 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-071 577

Yandex API Gateway (4): AISAFE2-MIT-001 316 · AISAFE2-MIT-002 320 · AISAFE2-MIT-016 375 · AISAFE2-MIT-029 429

Yandex Audit Trails (3): AISAFE2-MIT-015 371 · AISAFE2-MIT-022 400 · AISAFE2-MIT-033 444

Yandex Cloud AI Studio (7): AISAFE2-MIT-002 320 · AISAFE2-MIT-003 324 · AISAFE2-MIT-005 331 · AISAFE2-MIT-006 333 · AISAFE2-MIT-014 368 · AISAFE2-MIT-016 375 · AISAFE2-MIT-052 509

Yandex Cloud Backup (1): AISAFE2-MIT-024 410

Yandex Cloud Billing (1): AISAFE2-MIT-006 333

Yandex Cloud Detection and Response (2): AISAFE2-MIT-023 405 · AISAFE2-MIT-024 410

Yandex Cloud Logging (1): AISAFE2-MIT-022 400

Yandex Container Registry (4): AISAFE2-MIT-012 359 · AISAFE2-MIT-013 364 · AISAFE2-MIT-026 419 · AISAFE2-MIT-035 451

Yandex DataSphere (2): AISAFE2-MIT-034 448 · AISAFE2-MIT-041 473

Yandex Identity and Access Management (7): AISAFE2-MIT-008 342 · AISAFE2-MIT-018 383 · AISAFE2-MIT-020 392 · AISAFE2-MIT-024 410 · AISAFE2-MIT-029 429 · AISAFE2-MIT-030 434 · AISAFE2-MIT-041 473

Yandex Key Management Service (3): AISAFE2-MIT-011 356 · AISAFE2-MIT-012 359 · AISAFE2-MIT-043 481

Yandex Lockbox (2): AISAFE2-MIT-010 349 · AISAFE2-MIT-015 371

Yandex Managed Service for GitLab (3): AISAFE2-MIT-025 415 · AISAFE2-MIT-027 423 · AISAFE2-MIT-071 577

Yandex Managed Service for Kubernetes (1): AISAFE2-MIT-017 379

Yandex Managed Service for OpenSearch (3): AISAFE2-MIT-008 342 · AISAFE2-MIT-009 346 · AISAFE2-MIT-011 356

Yandex Managed Service for YDB (1): AISAFE2-MIT-021 396

Yandex MetaData Hub (3): AISAFE2-MIT-027 423 · AISAFE2-MIT-035 451 · AISAFE2-MIT-040 469

Yandex Monitoring (3): AISAFE2-MIT-006 333 · AISAFE2-MIT-023 405 · AISAFE2-MIT-036 456

Yandex Object Storage (5): AISAFE2-MIT-009 346 · AISAFE2-MIT-021 396 · AISAFE2-MIT-028 426 · AISAFE2-MIT-040 469 · AISAFE2-MIT-043 481

Yandex Search API (1): AISAFE2-MIT-005 331

Yandex Security Deck (7): AISAFE2-MIT-007 339 · AISAFE2-MIT-010 349 · AISAFE2-MIT-018 383 · AISAFE2-MIT-023 405 · AISAFE2-MIT-025 415 · AISAFE2-MIT-026 419 · AISAFE2-MIT-042 478

Yandex Serverless Containers (2): AISAFE2-MIT-006 333 · AISAFE2-MIT-017 379

Yandex SIEM (1): AISAFE2-MIT-023 405

Yandex Smart Web Security (3): AISAFE2-MIT-006 333 · AISAFE2-MIT-038 462 · AISAFE2-MIT-044 485

Yandex SmartCaptcha (2): AISAFE2-MIT-038 462 · AISAFE2-MIT-044 485

Yandex Virtual Private Cloud (3): AISAFE2-MIT-017 379 · AISAFE2-MIT-024 410 · AISAFE2-MIT-031 438

Yandex Workflows (1): AISAFE2-MIT-046 492

Компоненты

Компонент · AISAFE2-COMP-001

Пользовательский интерфейс и API приложения

Краткое описание

Принимает запросы пользователей и внешних клиентов, передаёт их в приложение и отображает полученные результаты.

Архитектурная область

Образует входную и выходную границу приложения: принимает пользовательские данные, приводит запрос к прикладному контракту, управляет представлением текущего взаимодействия и безопасно выводит результат. Компонент не определяет бизнес-решение и не формирует обращения к модели самостоятельно. В данном случае потребитель понимается в широком смысле и не ограничивается пользователем, отправившим запрос в AI-систему. В зависимости от назначения AI-системы потребителем может быть, в частности, программный компонент, сторонний агент, бизнес-процесс или инструмент разработки кода.

На схеме

Потребитель

Что входит

- Веб-интерфейсы, мобильные интерфейсы и консоли
- Публичные API приложения и обработчики входящих запросов
- Формы загрузки данных и представление результата
- Клиентское состояние текущего взаимодействия

Что не входит

- Серверная бизнес-логика и прикладные рабочие процессы
- Интерфейс инференса и поведение модели
- Общий контур исполнения политик для AI-запросов

Положение на архитектурной схеме

Компонент задаёт техническую UI/API-границу слоя потребителя, через которую поступает запрос и возвращается результат.

- Потребитель: Прямое соответствие

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	1	143

Карточка	Связи	Стр.
AISAFE2-WEAK-002 · Отсутствие привязки утверждений AI-ответа к источникам	3	145
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	5	146
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	10	148
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе	14	151
AISAFE2-WEAK-006 · Отсутствие раскрытия факта взаимодействия с AI	15	152
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	66	194
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	136	240
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	160	248
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	167	254
AISAFE2-WEAK-063 · Недостаточные пользовательские средства управления расходами и прозрачность тарификации	179	260
AISAFE2-WEAK-073 · Отсутствие независимого проверочного материала для AI-результата	189	274
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке	200	282
AISAFE2-WEAK-081 · Неаутентифицированный внешний вход в AI-приложение	210	289
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	718	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	721	93
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	724	95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	725	97
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	730	100
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	746	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	754	107
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	772	114
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	813	125
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	820	127
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	833	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	840	139
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	855	316
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	860	320
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	871	331

Карточка	Связи	Стр.
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1017	462
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1069	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1092	519
AISAFE2-MIT-056 · Исполнение версионируемой политики допустимости содержимого	1113	529
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	1115	531
AISAFE2-MIT-058 · Аутентификация на внешнем входе AI-приложения	1119	535
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1142	550
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1189	577

Компонент · AISAFE2-COMP-002

Модель и интерфейс инференса

Краткое описание

Объединяет используемые AI-системой модели и интерфейс инференса, возвращающий результат, достоверность и безопасность применения которого требуют проверки.

Архитектурная область

Охватывает одну или несколько собственных или предоставленных поставщиком моделей: LLM, мультимодальные, Mixture-of-Experts и иные ML-модели, включая специальные модели противодействия атакам. Независимо от назначения следует учитывать риск ошибочного или недостоверного вывода. Граница включает контракт инференса, но не обучение или дообучение, хранение пакета модели, вычислительную платформу и технический маршрут к внешнему провайдеру.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы
- Собственные и предоставленные поставщиком модели
- Модельные защитные ограничения и классификаторы
- API и протокол обращения к модели
- Параметры генерации и ограничения вывода
- Подготовка и выдача результата инференса

Что не входит

- Хранилище и реестр артефактов модели
- Вычислительная и сетевая среда выполнения
- Адаптер внешнего AI-сервиса и его маршрут

Положение на архитектурной схеме

Компонент соответствует моделям AI-системы, интерфейсу инференса и получению вероятностного результата.

- Инференс: Прямое соответствие

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	20	158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций	21	160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	22	162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	24	164
AISAFE2-WEAK-015 · Переобучение модели	39	175
AISAFE2-WEAK-018 · Избыточная наблюдаемость и воспроизводимость поведения модели через интерфейс	47, 48	180
AISAFE2-WEAK-019 · Избыточное раскрытие внутренних оценок модели	49, 50, 51	182
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	67	194
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	70	199
AISAFE2-WEAK-030 · Отсутствие проверяемой маркировки модели и её выходов	74	201
AISAFE2-WEAK-033 · Раскрытие сведений об отдельных обучающих записях через модель	80, 81	206
AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария	111	217
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	137	240
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели	156	242
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели	164	253
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	168	254
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели	191	276
AISAFE2-WEAK-076 · Отсутствие ресурсной изоляции арендаторов в разделяемой среде	195	280
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	208	286
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	214	292
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	218	294
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	241	308
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	731	100

Карточка	Связи	Стр.
AISAFE2-IMP-012 · Кража модели	761	111
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	767	113
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	773	114
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	799	120
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	821	127
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	841	139
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	861	320
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	873	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	889	349
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	914	368
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	954	410
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1000	448
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1009	456
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1018	462
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1024	466
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1034	478
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1043	485
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1051	492
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1060	494
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1070	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1093	519
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1162	564
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1178	573

Компонент · AISAFE2-COMP-003

Управление контекстом общения с моделью

Краткое описание

Безопасно формирует краткоживущий набор инструкций и данных, доступный модели при обработке одного запроса или шага Agent Core.

Архитектурная область

Отдельная возможность Agent Core для безопасного управления контекстом при работе с моделями. При формировании запроса она объединяет системные инструкции, пользовательские данные, результаты инструментов и данные кратковременных и долговременных хранилищ. Каждая единица информации включается с учётом происхождения, доверия, правил доступа, порядка и возможного влияния атакующего. Контекст не является самостоятельным хранилищем, состоянием задачи агента или источником данных.

На схеме

Ядро агента

Что входит

- Системные, разработческие и пользовательские инструкции текущего запроса
- Выбранная часть истории текущего диалога
- Извлечённые фрагменты и память с метками происхождения, доверия и доступа
- Описания и результаты работы инструментов
- Правила разделения и порядка частей контекста

Что не входит

- Состояние задачи и переходы между шагами агента
- RAG-корпус, поисковый индекс и конвейер извлечения
- Долговременная память и операционные хранилища

Положение на архитектурной схеме

Компонент соответствует выделенной внутри Agent Core возможности доверительной сборки краткоживущего контекста общения с моделью.

- Управление контекстом общения с моделью: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM01:2026 · Площадка реализации или атаки

Сборка контекста соединяет инструкции и внешние данные в точке, где вредоносный фрагмент может повлиять на поведение модели.

Граница связи: Наличие компонента не означает утрату границы доверия или успешное внедрение; решает фактическая типизация и обработка фрагментов.

Субъект	Аналитик угроз AI-приложения
Условие применения	Недоверенное содержимое попадает в контекст модели и может пересечь границу между данными и инструкциями.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	16	154
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	23	162
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	87	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	98	214
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	106	216
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции	113	220
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели	192	276
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	726	97
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	755	107
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	774	114
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	834	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	842	139
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	862	320
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	867	324
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	869	328

Карточка	Связи	Стр.
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	890	349
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	919	371
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1094	519
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	1158	561

Компонент · AISAFE2-COMP-004

Ядро агента и управление шагами (Agent Core)

Краткое описание

Определяет бизнес-логику и композицию AI-решения, управляет состоянием задачи, шагами и обращениями к моделям и разрешённым возможностям.

Архитектурная область

Реализует цикл и прикладную композицию агентного решения: оценивает состояние, выбирает действие, передаёт данные между шагами и определяет условия повтора, остановки или отката. Обычно эта логика находится в границе ответственности команды-разработчика AI-решения.

На схеме

Ядро агента

Что входит

- Бизнес-логика и композиция агентного AI-решения
- Планирование и последовательность шагов
- Логическое состояние и контрольные точки задачи
- Выбор действия, повтор, остановка и обработка отказа
- Координация вызовов модели и инструментов

Что не входит

- Исполнение конкретного инструмента или коннектора
- Межагентские сообщения и реестр агентов
- Серверная логика неагентных приложений

Положение на архитектурной схеме

Компонент задаёт внешний блок Agent Core с бизнес-логикой, композицией решения, состоянием задачи и порядком шагов.

- Ядро агента: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI01** · Площадка реализации или атаки
>
> Agent Core — место, где недоверенный материал может изменить цель или последовательность шагов агента.
>
> **Граница связи:** Наличие ядра не подтверждает разделение инструкций и данных или запрет неразрешённых действий.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент получает недоверенный ввод или сообщение другого агента, способное изменить его разрешённую цель.
Условие связи с карточкой	Контролируемый атакующим материал влияет на формирование цели или выбор следующего шага агента.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	131	237
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	138	240
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель	157	245
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	174	257
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	219	294
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	235	303
AISAFE2-WEAK-092 · Недостаточно ограниченные делегированные полномочия пользователя у агента	244	310
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	734	101
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	790	118
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	800	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	804	121
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	809	123

Карточка	Связи	Стр.
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	816	126
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	843	139
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	868	324
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	874	333
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	925	375
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	931	383
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	940	389
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	945	392
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	955	410
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	983	429
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1062	497
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1071	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1095	519
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	1159	561
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1163	564
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1179	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1190	577

Компонент · AISAFE2-COMP-005

Инструменты, MCP-серверы и коннекторы

Краткое описание

Предоставляет AI-системе общие исполняемые возможности и доступ к внешним прикладным системам через управляемую границу запросов и ответов.

Архитектурная область

Охватывает реализацию и вызов общих функций, через которые приложение или агент читает данные, изменяет внешние системы либо запускает вычисления. Компонент может поставляться другой командой или вендором, но разработчик решения отвечает за выбор, безопасное подключение, раскрытие данных, авторизацию и проверку результата. RAG, формирование базы знаний и A2A рассматриваются отдельно.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- Общие инструменты и их среды выполнения
- MCP-клиенты и MCP-серверы
- Коннекторы к прикладным API и данным
- Авторизованный вызов, преобразование аргументов и проверка результата

Что не входит

- Решение ядра агента о выборе действия
- Адаптер внешнего AI-сервиса
- RAG-извлечение и формирование базы знаний
- Межагентские шлюзы, сообщения и реестр агентов

Положение на архитектурной схеме

Компонент соответствует общим инструментам, MCP-серверам и коннекторам слоя межкомпонентного взаимодействия.

- Инструменты: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

> **C10.2.4** · Площадка реализации или атаки
>
> Компонент инструментов и MCP-серверов показывает границу, на которой клиент получает доступный список инструментов.
>
> **Граница связи:** Наличие компонента не подтверждает фильтрацию `tools/list` по правам клиента.
>
> | | |
> |---|---|
> | **Субъект** | Владелец или проверяющий AI-приложения |
> | **Условие применения** | В MCP-клиенте проверяется, какие инструменты доступны для обнаружения в пределах его полномочий по AISVS 1.0. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	6	146
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	11	148
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	17	154
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	26	167

Карточка	Связи	Стр.
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	60	188
AISAFE2-WEAK-039 · Выдача инструментом чувствительных или избыточных данных	97	212
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	139	240
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	159	246
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	162	251
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	169	254
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей	181	262
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения	204	284
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	215	292
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	220	294
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	236	303
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	239	306
AISAFE2-WEAK-092 · Недостаточно ограниченные делегированные полномочия пользователя у агента	245	310
AISAFE2-WEAK-095 · Недостаточно ограниченные делегированные платёжные полномочия	249	314
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	727	97
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	735	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	740	104
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	747	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	756	107
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	791	118
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	805	121
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	817	126
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	822	127
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	835	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	844	139

Карточка	Связи	Стр.
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	856	316
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	863	320
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	875	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	891	349
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	909	364
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	920	371
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	926	375
AISAFE2-MIT-017 · Изолированное исполнение инструментов	929	379
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	932	383
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	941	389
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	946	392
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	956	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	971	419
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	984	429
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	988	434
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	991	438
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	994	441
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1003	451
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1052	492
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1072	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1096	519
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	1116	531
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1143	550
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1150	554
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1164	564
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1180	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1191	577

Компонент · AISAFE2-COMP-006

RAG и извлечение из базы знаний

Краткое описание

Выполняет запросное авторизованное извлечение и ранжирование данных из подготовленной базы знаний для обогащения контекста модели.

Архитектурная область

Подсистема извлечения информации охватывает авторизованное извлечение из внутренней базы знаний посредством RAG, но не обращения общих инструментов к произвольным сторонним источникам. Компонент рассматривается как специализированный инструмент на пути текущего запроса: принимает поисковый запрос, применяет ограничения доступа, выполняет поиск и ранжирование и возвращает результаты с информацией о происхождении. Подготовка источников, преобразование, индексирование и публикация базы знаний относятся к отдельной подсистеме формирования базы знаний.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- Интерфейс запросного обращения к подготовленной базе знаний
- Авторизованный поиск и фильтрация доступных данных
- Ранжирование и выбор релевантных фрагментов
- Возврат фрагментов с происхождением и метками доверия

Что не входит

- Допуск источников, подготовка корпуса, индексирование и публикация базы знаний
- Исходные внешние и внутренние наборы данных для инструментов
- Наборы данных для обучения и оценки модели
- Краткоживущий контекст и долговременная память

Положение на архитектурной схеме

Компонент соответствует запросному RAG-извлечению и ранжированию из уже подготовленной базы знаний.

- RAG и извлечение данных: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C5.2.2 · Площадка реализации или атаки

Компонент RAG обозначает поиск и выдачу фрагментов, где проверяется право конечного пользователя.

Граница связи: Описание компонента не подтверждает повторную авторизацию после ранжирования и при сборке контекста.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверку прав конечного пользователя на каждом этапе поиска и сборки контекста.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-002 · Отсутствие привязки утверждений AI-ответа к источникам	4	145
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	18	154
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	170	254
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	231	299
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	741	104
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	748	106
AISAFE2-IMP-011 · Компрометация баз знаний	758	109
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	775	114
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	823	127
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	845	139
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	872	331
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	885	342
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1010	456
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1053	492

Карточка	Связи	Стр.
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1097	519
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1144	550
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1174	571
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1181	573

Компонент · AISAFE2-COMP-007

Долговременная память

Краткое описание

Хранит сведения между запросами, сессиями или задачами для последующего использования компонентом Agent Core.

Архитектурная область

Представляет собой сервис и хранилище данных с правилами записи, обновления, срока жизни, удаления и повторного использования сохранённой информации. Память отличается от временного контекста, логического состояния текущей агентной задачи и общих бизнес-данных приложения.

На схеме

Слой данных

Что входит

- Параметры долговременной памяти
- Сохранённые единицы информации
- Правила записи, обновления, истечения и удаления
- Извлечение данных для последующих запросов

Что не входит

- Краткоживущий контекст одного запроса
- Логическое состояние текущей задачи агента
- Операционные базы данных и кэши приложения

Положение на архитектурной схеме

Долговременная память хранится в области состояния Agent Core и может поставлять отобранные сведения в контекст последующих запросов.

- Хранилище ядра агента: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C8.2.3 · Площадка реализации или атаки

Долговременная память обозначает место, где ответ агента или инструмента может стать данными для последующих запросов.

Граница связи: Описание компонента не устанавливает проверку источника ответа до записи.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверку источника до попадания вывода агента или инструмента в доверенную память.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	19	154
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	36	173
AISAFE2-WEAK-023 · Незашифрованное хранение	53	186
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	88	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	99	214
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	140	240
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	742	104
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	749	106
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	780	116
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	824	127
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	846	139
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	892	349
AISAFE2-MIT-021 · Управление доверием и сроком жизни долговременной памяти	951	396
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	957	410
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	997	444

Карточка	Связи	Стр.
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1054	492
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1073	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1098	519
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1127	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1134	547

Компонент · AISAFE2-COMP-008

Данные для подсистемы обучения

Краткое описание

Защищённо хранит и передаёт сырые и размеченные данные, содержащие примеры, признаки и эталонные результаты для обучения, дообучения и независимой оценки моделей.

Архитектурная область

Применяется, когда в AI-решении присутствует подсистема обучения или дообучения моделей, и представляет используемые данные как управляемые активы с определённым назначением, происхождением и жизненным циклом. Охватывает обучающие и реально используемые оценочные наборы, но не процессы обучения, инструменты оценки, источники инструментов или продуктивные данные.

На схеме

Слой данных

Что входит

- Обучающие и валидационные выборки
- Тестовые, контрольные и состязательные наборы
- Разметка, признаки и эталонные ответы
- Версии и состав наборов данных

Что не входит

- Источники данных инструментов и подготовленные базы знаний
- Операционные данные и результаты приложения
- Инфраструктура обучения, тестирования и выпуска моделей

Положение на архитектурной схеме

Компонент соответствует защищённым данным для обучения, дообучения и, когда применимо, независимой оценки моделей.

- Данные для подсистемы обучения: Прямое соответствие

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	27	167
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	32	171
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	37	173
AISAFE2-WEAK-015 · Переобучение модели	40	175
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	42	176
AISAFE2-WEAK-023 · Незашифрованное хранение	54	186
AISAFE2-WEAK-033 · Раскрытие сведений об отдельных обучающих записях через модель	82	206
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	84	207
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	89	209
AISAFE2-WEAK-037 · Непропорциональный вклад отдельных обучающих записей в результат обучения	95	211
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	100	214
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	141	240
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели	165	253
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	226	296
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	768	113
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	781	116
AISAFE2-IMP-031 · Утрата датасета в результате атаки	839	137
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	893	349
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	958	410
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений	980	423
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1004	451
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1011	456
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1028	469
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1030	473
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1035	478
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных	1066	503

Карточка	Связи	Стр.
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1074	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1099	519
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	1124	541
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1128	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1135	547
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	1170	568

Компонент · AISAFE2-COMP-009

Конвейер обучения и дообучения

Краткое описание

При наличии собственной обучаемой модели преобразует подготовленные данные в новые версии и управляет заданиями обучения или дообучения.

Архитектурная область

Применяется, если хотя бы одна модель, не получаемая только как готовая услуга поставщика, обучается или дообучается в контуре решения. Если конвейер обучения принадлежит команде AI-решения, связанные угрозы и меры входят в её ответственность. Компонент охватывает воспроизводимую подготовку входов, запуск обучения и контрольные точки, но не независимую оценку, решение о выпуске или хранение окончательной модели в реестре.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- Подготовка данных и признаков для обучения
- Задания обучения и дообучения
- Гиперпараметры и конфигурация обучения
- Промежуточные контрольные точки и результаты задания

Что не входит

- Независимая оценка и решение о выпуске
- Хранилище и реестр окончательных моделей
- Хранилище данных для обучения
- Продуктивный инференс

Положение на архитектурной схеме

Компонент соответствует вычислительному и управляющему контуру обучения или дообучения собственных моделей решения.

- Обучение: Прямое соответствие

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	25	164
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	28	167
AISAFE2-WEAK-015 · Переобучение модели	41	175
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	43	176
AISAFE2-WEAK-033 · Раскрытие сведений об отдельных обучающих записях через модель	83	206
AISAFE2-WEAK-037 · Непропорциональный вклад отдельных обучающих записей в результат обучения	96	211
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	115	223
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	121	227
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	142	240
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели	166	253
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	242	308
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	769	113
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	894	349
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	910	364
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	933	383
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	972	419
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1005	451
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1012	456
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1029	469
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1031	473
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1036	478
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1075	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1100	519
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	1125	541
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1151	554

Компонент · AISAFE2-COMP-010

Хранилище и реестр артефактов модели

Краткое описание

Хранит, проверяет, версионирует и предоставляет разрешённые артефакты модели.

Архитектурная область

Управляет целостностью и идентичностью пакета модели после обучения: связывает веса, формат, манифест и метаданные версии, а затем выдаёт конкретный артефакт разрешённым потребителям.

На схеме

Периферийная область

Что входит

- Веса, конфигурация и формат модели
- Манифесты, подписи и метаданные версии
- Хранилища и реестры моделей
- Техническое продвижение и выдача выбранной версии

Что не входит

- Конвейер обучения и дообучения
- Решение о выпуске и независимые проверки
- Программные пакеты и образы без модели

Положение на архитектурной схеме

Артефакты модели и их реестр относятся только к жизненному циклу и не смешиваются с создаваемыми AI-системой рабочими результатами.

- Разработка, выпуск и цепочка поставки: Периферийная область

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM04:2026** · Площадка реализации или атаки
>
> Реестр хранит и выдаёт точные версии весов и связанных файлов; их подмена на этом этапе затрагивает цепочку поставки модели.
>
> **Граница связи:** Реестр охватывает артефакты модели; данные, программные зависимости и поставщик имеют отдельные границы. Наличие реестра не подтверждает безопасность содержимого.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	29	167
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	33	171
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	44	176
AISAFE2-WEAK-023 · Незашифрованное хранение	55	186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	61	188
AISAFE2-WEAK-030 · Отсутствие проверяемой маркировки модели и её выходов	75	201
AISAFE2-WEAK-031 · Недостаточная защита модели при поставке в недоверенную среду	76	203
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	78	204
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	116	223
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	122	227
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	143	240
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	216	292
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	221	294
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	227	296
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	243	308

Карточка	Связи	Стр.
AISAFE2-IMP-012 · Кража модели	762	111
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	770	113
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	776	114
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	782	116
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	852	141
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	906	359
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	911	364
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	915	368
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	959	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	973	419
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1001	448
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1006	451
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1040	481
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1076	509
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1129	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1136	547
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1165	564
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	1171	568

Компонент · AISAFE2-COMP-011

Вычислительная и сетевая среда выполнения

Краткое описание

Предоставляет вычислительные ресурсы, сеть, изоляцию и временное состояние для работающих компонентов AI-системы.

Архитектурная область

Образует техническую среду, в которой запускаются модели, приложения, агенты и вспомогательные сервисы. Среда распределяет ресурсы и обеспечивает изоляцию исполнения.

На схеме

Сквозная область

Что входит

- Виртуальные машины, контейнеры и оркестраторы
- Бессерверные среды и процессы
- CPU, GPU и другие вычислительные ускорители
- Сеть, временное хранилище и границы изоляции

Что не входит

- Серверная бизнес-логика приложения
- Долговременные базы данных и очереди
- Шлюзы и правила защиты AI-запросов

Положение на архитектурной схеме

Вычислительная и сетевая среда поддерживает несколько слоёв и поэтому показана как сквозная область.

- Прикладное исполнение и операционные данные: Сквозная область

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 32(1)(b)** · Площадка реализации или атаки
>
> Среда выполнения обозначает вычислительные и сетевые границы, на которых проверяют защищённость обработки персональных данных по статье 32.
>
> **Граница связи:** Само наличие среды не подтверждает изоляцию, доступность и устойчивость; нужны настройки, испытания и оценка риска для этого потока.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Персональные данные проходят через эту среду; контролёр или обработчик управляет ею либо располагает проверяемыми гарантиями поставщика.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	12	148
AISAFE2-WEAK-023 · Незашифрованное хранение	56	186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	62	188

Карточка	Связи	Стр.
AISAFE2-WEAK-031 · Недостаточная защита модели при поставке в недоверенную среду	77	203
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	79	204
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	107	216
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	123	227
AISAFE2-WEAK-050 · Остаточные данные в общей инфраструктуре исполнения моделей	128	234
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	132	237
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	144	240
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	175	257
AISAFE2-WEAK-075 · Отсутствие совокупных пределов масштабирования и расходов инфраструктуры	193	277
AISAFE2-WEAK-076 · Отсутствие ресурсной изоляции арендаторов в разделяемой среде	196	280
AISAFE2-WEAK-077 · Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку	198	281
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке	201	282
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения	205	284
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	217	292
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	222	294
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	240	306
AISAFE2-IMP-012 · Кража модели	763	111
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	792	118
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	810	123
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	825	127
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	847	139
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	876	333
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	907	359
AISAFE2-MIT-017 · Изолированное исполнение инструментов	930	379

Карточка	Связи	Стр.
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	934	383
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	947	392
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	960	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	974	419
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	992	438
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1032	473
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1041	481
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1055	492
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	1065	500
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1077	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1101	519
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	1117	531
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1137	547
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1145	550
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1152	554
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1166	564
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1182	573

Компонент · AISAFE2-COMP-012

Идентификация, авторизация, ключи и секреты

Краткое описание

Устанавливает идентичность субъектов и сервисов, принимает решения о доступе и управляет ключами и секретами AI-контура.

Архитектурная область

Охватывает базовые механизмы доверия, через которые пользователи, сервисы и рабочие нагрузки получают проверяемую идентичность и минимальные полномочия.

На схеме

Сквозная область

Что входит

- Каталоги пользователей и сервисных идентичностей
- IAM-роли, политики доступа и решения авторизации

- Хранилища ключей, секретов и учётных данных
- Выдача, ротация и отзыв полномочий

Что не входит

- Бизнес-правила допустимости действий
- Сетевые и вычислительные границы изоляции
- Администрирование учётной записи внешнего AI-провайдера

Положение на архитектурной схеме

Идентификация, авторизация, ключи и секреты образуют сквозную границу для всех исполняемых путей.

- Идентификация и управление доступом: Сквозная область

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	45	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	52	184
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	65	190
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	108	216
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	117	223
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	126	232
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	145	240
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	163	251
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	212	290
AISAFE2-WEAK-092 · Недостаточно ограниченные делегированные полномочия пользователя у агента	246	310
AISAFE2-WEAK-094 · Небезопасное хранение и распространение делегированных учётных данных пользователя	248	312
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	736	101
AISAFE2-IMP-011 · Компрометация баз знаний	759	109
AISAFE2-IMP-012 · Кража модели	764	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	783	116

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	793	118
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	886	342
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	895	349
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	904	356
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	908	359
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	921	371
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	935	383
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	948	392
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	961	410
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	985	429
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	989	434
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	993	438
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1033	473
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1042	481
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1063	497
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1078	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1102	519
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1111	525
AISAFE2-MIT-058 · Аутентификация на внешнем входе AI-приложения	1120	535
AISAFE2-MIT-059 · Усиленная аутентификация привилегированных человеческих учётных записей	1122	538
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1153	554
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1183	573

Компонент · AISAFE2-COMP-013

Телеметрия и обнаружение

Краткое описание

Собирает и сохраняет события AI-системы, сопоставляет их с правилами обнаружения и формирует проверяемые сигналы.

Архитектурная область

Охватывает технический путь от получения журналов, трассировок и метрик до нормализации, корреляции, хранения и выпуска предупреждения.

На схеме

Сквозная область

Что входит

- Сборщики журналов, трассировок и метрик
- Конвейеры обработки и хранилища телеметрии
- Правила и механизмы обнаружения
- Предупреждения и технические данные для расследования

Что не входит

- Команды и организационные процедуры реагирования
- Аварийное восстановление и резервное копирование сервисов
- Исполнение основного пользовательского запроса

Положение на архитектурной схеме

Телеметрия собирает события с разных поверхностей и не сводится к одному узлу исходной схемы.

- Телеметрия и обнаружение: Сквозная область

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-023 · Незашифрованное хранение	57	186
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	71, 72	199
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	90	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	101	214
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	146	240
AISAFE2-WEAK-061 · Избыточно открытая внутренняя телеметрия	173	255
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	223	294

Карточка	Связи	Стр.
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	750	106
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	784	116
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	877	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	896	349
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	916	368
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	952	400
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	953	405
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	962	410
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1013	456
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1019	462
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1044	485
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1079	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1103	519
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1130	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1138	547
AISAFE2-MIT-065 · Защищённое сохранение материалов расследования	1156	558
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1167	564
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1184	573

Компонент · AISAFE2-COMP-014

Контур человеческой проверки и подтверждения

Краткое описание

Организует человеческую проверку значимых решений в сценариях, где она предусмотрена, и фиксирует её исход.

Архитектурная область

Человеческое участие может требоваться, поскольку результат модели не гарантирует правильность. Проверка человеком промежуточных результатов AI-системы может повысить её надёжность, но создаёт специфические угрозы: ошибочную маршрутизацию, ошибки при

принятии решений, нехватку пропускной способности из-за усталости проверяющих, формальное подтверждение и небезопасную деградацию AI-системы при недоступности человека.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- Правила внедрения человеческой проверки при обработке запроса
- Очереди и интерфейсы подтверждения действий
- Рабочий процесс назначения, эскалации и замещения
- Запись решения, основания и времени подтверждения

Что не входит

- Проверяющие как команда или организационная единица
- Вывод промежуточных результатов работы AI-системы или рассуждений без принятия решений пользователем
- Автоматическое исполнение политик обработки запросов без участия человека

Положение на архитектурной схеме

Компонент соответствует обязательной человеческой проверке вероятностных результатов, подтверждению, отклонению и безопасной эскалации.

- Человек в контуре AI-системы: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

> **C9.2.1** · Площадка реализации или атаки
>
> Контур человеческой проверки обозначает интерфейс и запись решения для подтверждения значимого действия.
>
> **Граница связи:** Наличие контура не доказывает, что привилегированное или необратимое действие блокируется до проверки решения.
>
Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверенное подтверждение человека до привилегированного или необратимого действия.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	2	143

Карточка	Связи	Стр.
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	161	248
AISAFE2-WEAK-065 · Избыточная маршрутизация задач на человеческую проверку	184	265
AISAFE2-WEAK-067 · Недостаточная инструментальная поддержка оператора HITL	185	267
AISAFE2-WEAK-068 · Недостаточный контроль качества решений и утомления операторов HITL	186	268
AISAFE2-WEAK-069 · Неопределённый режим переполнения и недоступности человеческой проверки	187	269
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	719	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	722	93
AISAFE2-IMP-023 · Перегрузка контура человеческой проверки	832	129
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	942	389
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	1016	459
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1048	488

Компонент · AISAFE2-COMP-015

Контур разработки, тестирования и выпуска

Краткое описание

Связывает изменения AI-системы с независимыми проверками, решением о выпуске и доказательствами прохождения контрольных точек жизненного цикла.

Архитектурная область

Охватывает управляемый путь изменения кода, моделей, данных и конфигурации от рецензирования до сборки, оценки, проверки безопасности и допуска в среду. Здесь выполняются верификация, валидация, моделирование атак и выпускные проверки; постоянное применение политик к продуктивным запросам AI-системы относится к отдельному контуру исполнения.

На схеме

Периферийная область

Что входит

- Репозитории изменений, рецензирование и CI/CD
- Функциональное, защитное и регрессионное тестирование
- Независимая оценка, верификация, валидация и моделирование атак
- Контрольные точки выпуска, решение о допуске и пакет подтверждений

Что не входит

- Принудительное применение правил к продуктивным AI-запросам
- Эксплуатационная телеметрия и обнаружение событий
- Конвейер обучения как способ получения модели

Положение на архитектурной схеме

Разработка, тестирование и выпуск относятся к жизненному циклу за пределами рабочего контура исходной схемы.

- Разработка, выпуск и цепочка поставки: Периферийная область

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	30	167
AISAFE2-WEAK-017 · Отсутствие проверяемых критериев выбора модели	46	179
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	85	207
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	109	216
AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария	112	217
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	114	222
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	118	223
AISAFE2-WEAK-046 · Недостаточное управление соответствием требованиям	120	225
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	124	227
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	147	240
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	224	294
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	228	296
AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы	234	301
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	853	141
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	912	364
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	922	371
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	970	415
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	975	419
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений	981	423

Карточка	Связи	Стр.
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	982	426
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	998	444
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1002	448
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1007	451
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1025	466
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1037	478
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1049	488
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1061	494
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1080	509
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1089	514
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1112	525
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1168	564
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	1172	568
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1185	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1192	577

Компонент · AISAFE2-COMP-016

Адаптер внешнего AI-сервиса

Краткое описание

Передаёт запросы внешнему AI-сервису, нормализует ответы и обеспечивает управляемое поведение маршрута при ошибках зависимости.

Архитектурная область

Представляет технический узел на пути обращения к модели: клиент, прокси или адаптер, который выбирает разрешённую конечную точку, преобразует данные и обрабатывает ответ. Абстрактная граница доверия, административная панель провайдера и сама внешняя модель не являются этим компонентом.

На схеме

Периферийная область

Что входит

- Клиенты, прокси и адаптеры внешних AI API
- Маршрутизация к разрешённым конечным точкам

- Преобразование запросов и нормализация ответов
- Тайм-ауты, повторные обращения и обработка отказа провайдера

Что не входит

- Учётная запись и административный контур AI-провайдера
- Исполняемая внешним провайдером модель
- Обычные прикладные инструменты и коннекторы

Положение на архитектурной схеме

Адаптер находится на внешней границе маршрута к модели и пересекается с поверхностью инференса, не заменяя её.

- Внешний AI-провайдер: Периферийная область

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 44** · Площадка реализации или атаки
>
> Адаптер внешнего AI-сервиса обозначает один технический путь, на котором может происходить передача персональных данных в третью страну.
>
> **Граница связи:** Наличие адаптера не означает, что такая передача есть или законна; другие маршруты проверяются отдельно.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Установлена фактическая передача персональных данных в третью страну и территориальная применимость GDPR.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	63	188
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	68	194
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	73	199
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	91	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	102	214
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	133	237
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	148	240

Карточка	Связи	Стр.
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	171	254
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей	182	262
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения	206	284
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	232	299
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	728	97
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	732	100
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	751	106
AISAFE2-IMP-012 · Кража модели	765	111
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	771	113
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	794	118
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	801	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	806	121
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	826	127
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	836	130
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	878	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	897	349
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	917	368
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	936	383
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	963	410
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	990	434
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	995	441
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1014	456
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1020	462
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1026	466
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1038	478
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1045	485
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1056	492

Карточка	Связи	Стр.
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1081	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1104	519
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	1118	531
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1146	550
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1154	554
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1175	571

Компонент · AISAFE2-COMP-017

Серверная часть приложения и бизнес-логика

Краткое описание

Реализует прикладной сценарий, бизнес-правила и координацию AI-возможностей для агентных и неагентных приложений.

Архитектурная область

Обрабатывает запрос после внешнего интерфейса, применяет правила предметной области, вызывает нужные AI- и не-AI-компоненты и формирует прикладной результат. Эта граница применима и к приложениям без агента.

На схеме

Сквозная область

Что входит

- Сервисный и прикладной слой приложения
- Бизнес-правила и рабочие процессы
- Неагентная координация вызовов модели и данных
- Фоновые прикладные задания и формирование результата

Что не входит

- Публичный интерфейс и внешний API приложения
- Ядро агента и управление его шагами
- Операционные хранилища, очереди и кэши
- Общий контур исполнения политик для AI-запросов

Положение на архитектурной схеме

Серверная часть и бизнес-логика координируют показанные узлы, но отсутствуют в исходном эскизе как отдельный блок.

- Прикладное исполнение и операционные данные: Периферийная область

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	7	146
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	13	148
AISAFE2-WEAK-051 · Выдача результата без завершения обязательных стадий вычислений	129	235
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	134	237
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	149	240
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель	158	245
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	172	254
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	176	257
AISAFE2-WEAK-063 · Недостаточные пользовательские средства управления расходами и прозрачность тарификации	180	260
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей	183	262
AISAFE2-WEAK-073 · Отсутствие независимого проверочного материала для AI-результата	190	274
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке	202	282
AISAFE2-WEAK-081 · Неаутентифицированный внешний вход в AI-приложение	211	289
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	237	303
AISAFE2-WEAK-093 · Неверная атрибуция тарифицируемого потребления между арендаторами	247	311
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	720	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	723	93
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	729	97
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	733	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	737	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	743	104
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	752	106
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	777	114

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	795	118
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	802	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	807	121
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	811	123
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	814	125
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	818	126
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	827	127
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	848	139
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	857	316
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	864	320
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	879	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	898	349
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	927	375
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	943	389
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	949	392
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	964	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	976	419
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	986	429
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1021	462
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1050	488
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1057	492
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	1067	507
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1082	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1105	519
AISAFE2-MIT-058 · Аутентификация на внешнем входе AI-приложения	1121	535
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	1160	561

Карточка	Связи	Стр.
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1186	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1193	577

Компонент · AISAFE2-COMP-018

Операционные данные, очереди и кэши

Краткое описание

Хранит и передаёт рабочие данные приложения, промежуточное состояние сессий и задания, не относящиеся к специализированным AI-хранилищам.

Архитектурная область

Охватывает долговременные и временные прикладные данные, через которые серверная часть связывает этапы обработки и выполняет фоновые задания.

На схеме

Сквозная область

Что входит

- Операционные реляционные и документные базы данных
- Хранилища входов и промежуточного состояния обработки
- Очереди сообщений, задания и событийные шины
- Кэши и хранилища серверных сессий

Что не входит

- RAG-корпусы и векторные индексы
- Долговременная память пользователя или агента
- Наборы данных для обучения и оценки
- Переиспользуемые результаты AI-системы, артефакты моделей и телеметрия

Положение на архитектурной схеме

Операционные данные, очереди и кэши поддерживают прикладное исполнение и не смешиваются со специализированной памятью или RAG-хранилищем.

- Прикладное исполнение и операционные данные: Периферийная область

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-023 · Незашифрованное хранение	58	186
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	92	209

Карточка	Связи	Стр.
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	103	214
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	110	216
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	135	237
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	150	240
AISAFE2-WEAK-076 · Отсутствие ресурсной изоляции арендаторов в разделяемой среде	197	280
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке	203	282
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	744	104
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	753	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	757	107
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	778	114
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	785	116
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	812	123
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	828	127
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	849	139
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	880	333
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	887	342
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	899	349
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	923	371
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	965	410
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1058	492
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1083	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1106	519
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1131	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1139	547

Карточка	Связи	Стр.
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1147	550
AISAFE2-MIT-065 · Защищённое сохранение материалов расследования	1157	558

Компонент · AISAFE2-COMP-019

Контур исполнения политик и защиты AI-запросов

Краткое описание

Проверяет и принудительно применяет утверждённые правила безопасности до, во время и после обработки AI-запроса.

Архитектурная область

Размещает на продуктивном пути шлюзы, защитные правила и механизмы разрешения, блокировки или преобразования входов, выходов и действий. В частности, решает задачи санитизации и валидации данных, на формирование которых может влиять потенциальный атакующий. Компонент исполняет уже утверждённую политику в реальном времени; разработка правил, базовая IAM-авторизация и последующее обнаружение событий остаются отдельными задачами.

На схеме

Сквозная область

Что входит

- AI-шлюзы и прокси принудительного контроля
- Защитные правила для входов, выходов и действий
- DLP, модерация и фильтрация чувствительных данных
- Ограничение частоты, ресурсов, источников и маршрутов AI-запросов

Что не входит

- Разработка, тестирование и выпуск политик
- Каталоги идентичностей, IAM и хранилища секретов
- Телеметрия, корреляция и обнаружение после события
- Прикладные бизнес-правила приложения

Положение на архитектурной схеме

Компонент соответствует детерминированному исполнению политик и защите AI-запросов на нескольких этапах рабочего пути.

- Исполнение политик и защита запросов: Сквозная область

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.5.3 · Площадка реализации или атаки

Контур исполнения политик обозначает возможную независимую точку применения решения о доступе к AI-действию.

Граница связи: Базовая IAM и прикладные правила вне его границы; конкретную политику доступа нужно подтвердить.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения требует решения о доступе в логике приложения или независимом механизме политики.
Условие связи с карточкой	В этом компоненте настроена независимая от модели политика доступа к AI-действию.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	8	146
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	69	194
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	93	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	104	214
AISAFE2-WEAK-051 · Выдача результата без завершения обязательных стадий вычислений	130	235
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	151	240
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	177	257
AISAFE2-WEAK-072 · Отсутствие обязательного технического контроля межзонных потоков данных	188	273
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	209	286
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	238	303
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	738	101
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	796	118
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	829	127

Карточка	Связи	Стр.
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	837	130
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	858	316
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	865	320
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	870	328
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	881	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	900	349
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	928	375
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	944	389
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	966	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	977	419
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	987	429
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	996	441
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1022	462
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1046	485
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	1068	507
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1084	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1107	519
AISAFE2-MIT-056 · Исполнение версионируемой политики допустимости содержимого	1114	529
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	1161	561
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1187	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1194	577

Компонент · AISAFE2-COMP-020

Контур управления внешним AI-провайдером

Краткое описание

Управляет учётными записями, разрешёнными моделями, регионами, квотами и настройками обработки данных внешнего AI-провайдера.

Архитектурная область

Охватывает административную плоскость стороннего AI-сервиса: организацию, проекты, роли, доступные модели и параметры провайдера, влияющие на хранение данных, использование запросов для обучения моделей, стоимость запросов и географию обработки. Он отделён от пути обработки запросов AI-системой и от внутреннего хранения выданных учётных данных.

На схеме

Периферийная область

Что входит

- Организации, проекты и административные API провайдера
- Разрешение моделей, регионов и конечных точек
- Настройки хранения данных и использования запросов провайдером
- Квоты на вычислительные ресурсы, бюджетные ограничения и выпуск учётных данных

Что не входит

- Передача конкретного запроса через адаптер внешнего AI-сервиса
- Внутреннее хранилище ключей и секретов
- Договорная и юридическая оценка условий провайдера

Положение на архитектурной схеме

Учётные записи, модели, регионы, квоты и настройки обработки данных поставщика находятся во внешнем контуре управления.

- Внешний AI-провайдер: Периферийная область

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 44** · Площадка реализации или атаки
>
> Настройки внешнего AI-провайдера помогают определить регион обработки и маршрут возможной передачи персональных данных в третью страну или международную организацию.
>
> **Граница связи:** Настройка региона не доказывает фактическое место обработки и не заменяет правовое основание, гарантии и проверку последующей передачи по главе V.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Установлена фактическая передача персональных данных в третью страну и территориальная применимость GDPR.
Условие связи с карточкой	Через поставщика действительно передаются персональные данные в третью страну или международную организацию; проверить документы поставщика и действующие настройки маршрута.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	119	223
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	152	240
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	178	257
AISAFE2-WEAK-075 · Отсутствие совокупных пределов масштабирования и расходов инфраструктуры	194	277
AISAFE2-WEAK-077 · Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку	199	281
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения	207	284
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	213	290
AISAFE2-IMP-012 · Кража модели	766	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	786	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	797	118
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	803	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	808	121
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	815	125
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	830	127
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	882	333
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	918	368
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	937	383
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	967	410
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1023	462
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1027	466
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1039	478
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1047	485
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1085	509
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1090	514

Карточка	Связи	Стр.
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1108	519
AISAFE2-MIT-059 · Усиленная аутентификация привилегированных человеческих учётных записей	1123	538
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1155	554

Компонент · AISAFE2-COMP-021

Межагентское взаимодействие и реестр агентов

Краткое описание

Обеспечивает обнаружение и подключение агентов к AI-системе, а также обмен запросами и ответами с контролируемым делегированием между независимыми агентными контурами.

Архитектурная область

Охватывает межкомпонентные протоколы и инфраструктуру взаимодействия агентов. Один агент обнаруживает и аутентифицирует другого, проверяет его возможности и доверие, передаёт задачу в защищённом конверте, связывает ответ с делегированием и блокирует повтор сообщений. Наряду с A2A возможны другие протоколы и субагенты. Как и инструмент, агент требует санитизации запроса и валидации результата. Разработчик отвечает за выбор агента, раскрываемые данные, полномочия и результат. Внутренние шаги одного агента и обычные инструменты рассматриваются отдельно.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- Реестры агентов и описание их возможностей
- A2A-шлюзы, брокеры и другие протоколы обмена
- Конверты сообщений, корреляция, аутентификация, проверка доверия и защита от повтора
- Делегирование задач и возврат результата между агентами и субагентами

Что не входит

- Шаги и состояние внутри одного ядра агента
- MCP-инструменты и прикладные коннекторы
- Общие очереди сообщений без агентной семантики
- Первичный выпуск идентичностей и ключей

Положение на архитектурной схеме

Межагентские сообщения, обнаружение и делегирование представлены отдельной A2A-границей слоя межкомпонентного взаимодействия.

- Межагентское взаимодействие: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI07** · Площадка реализации или атаки
>
> Компонент межагентского взаимодействия обозначает путь сообщений, на котором нужно проверять отправителя, адресата и задачу.
>
> **Граница связи:** Наличие реестра не подтверждает защиту сообщений, отказ при ошибке и охват обходных каналов связи.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агенты обмениваются сообщениями, для которых значимы отправитель, получатель, задача и защита от повтора.
Условие связи с карточкой	Агенты обмениваются рабочими сообщениями через реестр или коммуникационный компонент.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	64	188
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	127	232
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	153	240
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	739	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	745	104
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	798	118
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	819	126
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	831	127
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	838	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	850	139
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	859	316
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	883	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	901	349

Карточка	Связи	Стр.
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	938	383
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	950	392
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	968	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	978	419
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1059	492
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1064	497
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1086	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1109	519
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1148	550
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1188	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1195	577

Компонент · AISAFE2-COMP-022

Программные артефакты и цепочка поставки

Краткое описание

Управляет происхождением, составом, целостностью и распространением программных зависимостей и сборочных артефактов AI-системы.

Архитектурная область

Представляет программные активы, поступающие в сборку и среду выполнения: библиотеки, пакеты, образы, плагины и инфраструктурный код вместе с манифестами и подтверждениями происхождения.

На схеме

Периферийная область

Что входит

- Сторонние библиотеки, пакеты и зафиксированные зависимости
- Контейнерные образы, плагины и расширения
- Инфраструктурный код и сборочные заготовки
- Реестры программных артефактов, SBOM, подписи и аттестации

Что не входит

- Артефакты модели и их реестр
- Наборы данных, RAG-корпусы и долговременная память
- Рабочий процесс CI/CD, тестирование и решение о выпуске

Положение на архитектурной схеме

Программные зависимости и сборочные артефакты относятся только к цепочке поставки и не смешиваются с рабочими результатами AI-системы.

- Разработка, выпуск и цепочка поставки: Периферийная область

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM04:2026** · Площадка реализации или атаки
>
> Программные зависимости и сборочные артефакты образуют часть цепочки поставки, через которую подмена может попасть в AI-систему.
>
> **Граница связи:** Карточка не охватывает происхождение весов и обучающих данных и не доказывает, что зависимость была проверена.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	31	167
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	125	227
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	154	240
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	779	114
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	787	116
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	913	364
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	924	371
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	979	419
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1008	451
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1087	509
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1091	514

Компонент · AISAFE2-COMP-023

Подсистема формирования базы знаний

Краткое описание

Допускает источники, подготавливает и преобразует данные, строит индексы и публикует защищённую базу знаний для RAG и других инструментов.

Архитектурная область

Охватывает жизненный цикл подготовки базы знаний: проверку и допуск источников, очистку и преобразование, разделение по конфиденциальности и доступу, индексирование, версионирование и публикацию. Граница заканчивается доступной для поиска базой и не включает запросное извлечение, исходные наборы источников или данные обучения.

На схеме

Слой межкомпонентного взаимодействия

Что входит

- Правила допуска и регистрации источников базы знаний
- Очистка, разбиение, преобразование и обогащение данных
- Разделение по конфиденциальности и правилам доступа
- Построение и обновление поисковых и векторных индексов
- Версионирование и публикация подготовленной базы знаний

Что не входит

- Запросное извлечение, ранжирование и возврат фрагментов
- Исходные внешние и внутренние наборы данных для инструментов
- Данные для обучения, дообучения и независимой оценки модели

Положение на архитектурной схеме

Компонент соответствует допуску источников, подготовке, защите, индексированию и публикации базы знаний.

- Формирование базы знаний: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI06** · Площадка реализации или атаки
>
> Подсистема формирования базы знаний обозначает путь, по которому вредный материал может попасть в повторно используемый контекст агента.
>
> **Граница связи:** Сам компонент не подтверждает проверку источника, обнаружение отравления или защиту иных видов памяти.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.
Условие связи с карточкой	Допущенный материал сохраняется и позже повторно используется агентом.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	34	171
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	38	173
AISAFE2-WEAK-023 · Незашифрованное хранение	59	186
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	94	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	105	214
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	155	240
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	225	294
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	229	296
AISAFE2-WEAK-086 · Утрата атрибутов доступа исходного материала при индексировании базы знаний	230	298
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	233	299
AISAFE2-IMP-011 · Компрометация баз знаний	760	109
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	788	116
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	854	141

Карточка	Связи	Стр.
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	884	339
AISAFE2-MIT-009 · Версионирование и откат базы знаний	888	346
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	902	349
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	905	356
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	939	383
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	969	410
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	999	444
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1015	456
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1088	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1110	519
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	1126	541
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1132	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1140	547
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1149	550
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1169	564
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	1173	568
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1176	571

Компонент · AISAFE2-COMP-024

Артефакты работы AI-системы

Краткое описание

Хранит и передаёт результаты AI-системы для повторного использования; они остаются недоверенными и требуют проверки перед применением для решения бизнес-задач.

Архитектурная область

Охватывает созданные AI-системой артефакты различной природы и структуры: в частности, код, изображения, документы и текстовые результаты, предназначенные для повторного использования, передачи или исполнения. Программное происхождение не делает их доверенными ни с точки зрения семантического содержания, ни с точки зрения структуры: перед применением требуются проверка и безопасная обработка. Злоумышленник может пытаться использовать артефакты AI-системы для размещения вредоносной нагрузки.

На схеме

Слой данных

Что входит

- Сгенерированный код и конфигурации
- Сгенерированные изображения и другие медиарезультаты
- Сгенерированные документы и сохраняемые текстовые результаты
- Хранилища, передача и выдача переиспользуемых AI-результатов

Что не входит

- Веса, конфигурации и реестры артефактов модели
- Операционное состояние, очереди, кэши и серверные сессии
- Программные зависимости, контейнерные образы и сборочные артефакты

Положение на архитектурной схеме

Компонент соответствует недоверенным переиспользуемым результатам работы AI-системы, требующим проверки перед бизнес-применением.

- Артефакты работы AI-системы: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM10:2026** · Площадка реализации или атаки
>
> Повторно используемый AI-артефакт остаётся недоверенным выводом и может передать опасное содержимое следующему обработчику.
>
> **Граница связи:** Хранение артефакта само по себе не проверяет его структуру, смысл или безопасное применение у конкретного получателя.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется получателем, инструментом или приложением для последующего действия.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	9	146
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	851	139
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	866	320
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	1133	544
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1141	547

Компонент · AISAFE2-COMP-025

Источники данных для инструментов

Краткое описание

Предоставляет инструментам и RAG исходные внутренние и внешние данные, включая API и общедоступные источники, с учётом их состава, целостности, происхождения, правил доступа и конфиденциальности.

Архитектурная область

Охватывает исходные наборы данных, API, файлы, хранилища, общедоступные источники и иные поверхности, из которых инструменты или RAG получают данные. Для каждого источника значимы возможность влияния атакующего на состав и целостность, доказуемость происхождения, правила доступа и риск раскрытия. Подготовленная база знаний, обучающие и оценочные данные имеют отдельные жизненные циклы и компоненты.

На схеме

Слой данных

Что входит

- Внутренние наборы данных или файлы и прикладные хранилища
- Внешние API, файлы, потоки, общедоступные источники и данные, предоставляемые вендором или другими командами
- Метаданные происхождения и целостности исходных данных
- Правила доступа и конфиденциальности источника данных

Что не входит

- Подготовленная и опубликованная база знаний
- Запросное извлечение и ранжирование данных с помощью RAG
- Наборы данных для обучения, дообучения и оценки модели

Положение на архитектурной схеме

Компонент соответствует исходным внутренним и внешним поверхностям данных, к которым обращаются инструменты и RAG.

- Источники данных для инструментов: Прямое соответствие

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI06** · Площадка реализации или атаки
>
> Источники данных инструментов могут внести вредный материал в память или повторно извлекаемый контекст агента.
>
> **Граница связи:** Связь требует сохранения и последующего использования материала; разовый ответ инструмента сам по себе не образует ASI06.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.
Условие связи с карточкой	Данные инструмента или RAG-источника сохраняются в памяти агента либо извлекаемом контексте.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	35	171
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	86	207
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	789	116
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	903	349
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1177	571

Воздействия

Воздействие · AISAFE2-IMP-001

Выполнение действий, явно предложенных AI-системой

Краткое описание

Потребитель выполняет действие, прямо предложенное AI-системой, и тем самым создаёт для себя вредное последствие.

Подробное описание

Воздействие соответствует ситуации, когда происходит прямое следование предложению AI-системы, а не самостоятельное решение, принятое с учётом результата AI-системы. Типичный пример — запуск недоверенного кода, который AI-система предложила выполнить, повлекший какой-либо ущерб, например, удаление файлов с критичными бизнес-данными.

Класс воздействия

- I.1 Негативные последствия для потребителя за счет использования результатов AI-системы

Наблюдаемые признаки

- В истории взаимодействия сохранено явное предложение AI-системы, а запись об исполнении подтверждает соответствующее действие потребителя.
- Материалы инцидента связывают наступившее вредное последствие именно с выполнением этого предложения, а не только с фактом получения ответа.

Области возможного ущерба

- Человек
- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM07:2026** · Возможное последствие
>
> Прямое выполнение недостоверного совета AI-системы может превратить ложный ответ в вредное действие пользователя.
>
> **Граница связи:** Карточка фиксирует последствие следования совету; его ложность и источник ошибки устанавливаются отдельно.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется как фактическое основание ответа, рекомендации или решения.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	250	143
AISAFE2-WEAK-002 · Отсутствие привязки утверждений AI-ответа к источникам	254	145
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	256	146
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе	266	151
AISAFE2-WEAK-006 · Отсутствие раскрытия факта взаимодействия с AI	270	152
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели	409	242
AISAFE2-WEAK-073 · Отсутствие независимого проверочного материала для AI-результата	442	274
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	718	35
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	719	65
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	720	72

Воздействие · AISAFE2-IMP-002

Последствия принятия решений на основе результатов AI-системы

Краткое описание

Потребитель или организация принимает решение, существенно опираясь на результат AI-системы, и сталкивается с негативным последствием.

Подробное описание

В отличие от буквального выполнения предложенного действия в данном воздействии рассматривается ситуация, когда пользователь принимает решение на основе сведений, сформированных как результат работы AI-системы. Воздействие относится к последствиям для потребителя, поскольку вред возникает при практическом использовании результата. Примеры — покупка акций без оценки рисков по рекомендации AI-системы или отмена заказов из-за неверного ответа бота поддержки.

Класс воздействия

- I.1 Негативные последствия для потребителя за счет использования результатов AI-системы

Наблюдаемые признаки

- Материалы процесса показывают, что результат AI-системы был существенным основанием выбора, а журнал подтверждает исполнение решения.
- После исполнения зарегистрировано вредное последствие; одной ошибки в ответе без решения и последующего события недостаточно.

Области возможного ущерба

- Человек
- Организация
- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM07:2026** · Возможное последствие
>
> Недостоверный ответ AI-системы может стать существенным основанием решения, которое приводит к ущербу.
>
> **Граница связи:** Вредное решение возможно и при верном ответе; карточка не доказывает недостоверность или степень влияния модели.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется как фактическое основание ответа, рекомендации или решения.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	251	143
AISAFE2-WEAK-002 · Отсутствие привязки утверждений AI-ответа к источникам	255	145

Карточка	Связи	Стр.
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе	267	151
AISAFE2-WEAK-006 · Отсутствие раскрытия факта взаимодействия с AI	271	152
AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария	375	217
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели	410	242
AISAFE2-WEAK-073 · Отсутствие независимого проверочного материала для AI-результата	443	274
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	491	331
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	492	466
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	493	488
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	721	35
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	722	65
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	723	72

Воздействие · AISAFE2-IMP-003

Вовлечение пользователей в злонамеренные схемы

Краткое описание

Общение через AI-систему склоняет пользователя к участию в мошеннической, обманной или иной злонамеренной схеме.

Подробное описание

Отличительный признак — намеренное вовлечение человека в схему, а не просто вредная рекомендация или ошибочное решение. Воздействие относится к классу последствий для потребителя, потому что атакующий добивается изменения его поведения. Пример описываемого воздействия — перевод денег на счёт атакующего после общения с AI-ассистентом.

Класс воздействия

- I.1 Негативные последствия для потребителя за счет использования результатов AI-системы

Наблюдаемые признаки

- История диалога содержит последовательность убеждающих сообщений, после которой пользователь совершил требуемое злонамеренной схемой действие.
- Обращение пользователя, платёжная запись или расследование связывает участие в схеме атакующего с конкретным взаимодействием через AI-систему.

Области возможного ущерба

- Человек
- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI09** · Возможное последствие
>
> Манипулятивная подача агента может склонить пользователя к участию в мошеннической или иной злонамеренной схеме.
>
> **Граница связи:** Для ASI09 нужно установить роль агента и влияние его сообщения на решение человека; сам факт мошенничества недостаточен.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Человек принимает решение на основе представления, сформированного агентом.
Условие связи с карточкой	Убедительная или сфабрикованная подача агента склоняет человека участвовать во вредоносной схеме.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	252	143
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	257	146
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	259	148
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе	268	151
AISAFE2-WEAK-006 · Отсутствие раскрытия факта взаимодействия с AI	272	152
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	274	154
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	300	167
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	311	173
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	494	405
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	495	462
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	724	35

Воздействие · AISAFE2-IMP-004

Передача потребителем чувствительных данных

Краткое описание

Потребитель передаёт AI-системе или через неё атакующему секреты, персональные или иные чувствительные сведения, не предназначенные для такой обработки или передачи.

Подробное описание

Воздействие возникает, когда потребитель вследствие результата работы AI-системы сам передаёт чувствительные данные AI-системе либо через неё атакующему, не проверив получателя и основание обработки. Оно отличается от извлечения уже хранимых данных и от последующей передачи в недоверенный контур. Причиной могут стать фишинг или социальная инженерия, включая внедрение в систему инструмента атакующего, который запрашивает сведения.

Класс воздействия

- I.1 Негативные последствия для потребителя за счет использования результатов AI-системы

Наблюдаемые признаки

- В журнале диалога или вызова инструмента обнаружено значение, классифицированное как секрет, персональные либо иные чувствительные сведения.
- Происхождение значения указывает на пользовательский ввод или файл; предшествующий ответ AI-системы или запрос подключённого инструмента побудил потребителя передать его, а для такой передачи отсутствует утверждённое основание обработки.

Области возможного ущерба

- Человек
- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Возможное последствие
>
> Когда пользователь по ответу системы передаёт секреты или персональные данные неуполномоченному получателю, возникает один из путей раскрытия сведений.
>
> **Граница связи:** Обычный разрешённый ввод данных в AI-сервис не равен утечке; получатель и основание передачи должны быть установлены.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	253	143
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	260	148
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе	269	151
AISAFE2-WEAK-006 · Отсутствие раскрытия факта взаимодействия с AI	273	152
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	275	154
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	496	349
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	497	410
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	725	35
AISAFE2-COMP-003 · Управление контекстом общения с моделью	726	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	727	44
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	728	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	729	72

Воздействие · AISAFE2-IMP-005

Нарушение этических норм по отношению к пользователю AI-системы

Краткое описание

Ответ, решение или обращение AI-системы с пользователем нарушает утверждённые для сценария этические нормы и причиняет ему вред.

Подробное описание

Это непосредственное пользовательское последствие, а не высокоуровневый этический вывод для организации или общества. Пример — оскорбление пользователя моделью по признаку принадлежности к определённой социальной группе. Этическая оценка сама по себе не устанавливает нарушение применимого права.

Класс воздействия

- I.1 Негативные последствия для потребителя за счет использования результатов AI-системы

Наблюдаемые признаки

- Проверка ответа или решения по утверждённым этическим критериям фиксирует нарушение, непосредственно затрагивающее пользователя.
- Обращения или результаты оценки позволяют определить затронутую группу и вид вреда; одного несогласия с ответом недостаточно.

Области возможного ущерба

- Человек
- Этические последствия

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	258	146
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	281	158
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	290	162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	295	164
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	452	286
AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы	473	301
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	498	415
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	499	488

Воздействие · AISAFE2-IMP-006

Неожидаемые (для разработчика AI-системы) ответы моделей

Краткое описание

Модель выдаёт ответ, выходящий за предусмотренные разработчиком границы содержания, формата, качества или поведения.

Подробное описание

Это воздействие относится к классу нарушений заложенных бизнес-сценариев и описывает поведение самой модели: например, успешный обход её встроенных ограничений или непредсказуемый результат, включая галлюцинацию. Оно отличается от обхода логики Agent Core, где существенны алгоритмы ядра и механизмы управления контекстом, и от сбоя инструментов. Причины воздействия кроются в поведении или свойствах используемой модели.

Класс воздействия

- I.2 Нарушение предусмотренных сценариев работы AI-системы

Наблюдаемые признаки

- Журнал вывода содержит ответ, нарушивший принятые ограничения содержания, формата, качества или поведения и переданный для дальнейшей обработки AI-системой.
- Трассировка локализует отклонение именно на выходе модели, а не в последующем выборе шага Agent Core или работе инструмента.

Области возможного ущерба

- Информационные активы

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	276	154
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	282	158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций	286	160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	291	162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	296	164
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	301	167
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	310	171
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	312	173
AISAFE2-WEAK-015 · Переобучение модели	315	175

Карточка	Связи	Стр.
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	317	176
AISAFE2-WEAK-017 · Отсутствие проверяемых критериев выбора модели	322	179
AISAFE2-WEAK-037 · Непропорциональный вклад отдельных обучающих записей в результат обучения	366	211
AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария	376	217
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции	377	220
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	381	222
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели	411	242
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	500	316
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	501	320
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	502	324
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	503	328
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	504	456
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	505	469
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	506	494
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	507	507
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	730	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	731	37
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	732	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	733	72

Воздействие · AISAFE2-IMP-007

Обход ограничений или предусмотренной логики работы Agent Core

Краткое описание

Сочетание модели, собранного контекста и алгоритмов Agent Core приводит к обходу предусмотренной бизнес-логики или ограничений.

Подробное описание

Воздействие нарушает сценарий обработки запроса на уровне формирования контекста и оркестрации других компонентов AI-системы, а не сводится к неожиданному тексту модели или самостоятельному поведению инструмента. Атака может полностью или частично

переопределить поведение Agent Core: манипулировать подцелью, выбором инструмента, аргументами или последовательностью шагов либо исказить интерпретацию результата. Меры применяются к контексту, логике Agent Core и ограничениям заданных бизнес-сценариев.

Класс воздействия

- I.2 Нарушение предусмотренных сценариев работы AI-системы

Наблюдаемые признаки

- Трасса Agent Core содержит подцель, инструмент, аргументы, последовательность или переход состояния вне разрешённой для запроса логики.
- Сравнение с моделью допустимого процесса показывает, что отклонение возникло при планировании, действии или обработке результата шага.

Области возможного ущерба

- Информационные активы
- Работа сервиса

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI01** · Возможное последствие
>
> Обход предусмотренной логики Agent Core может быть результатом подмены цели или последовательности действий агента.
>
> **Граница связи:** Само отклонение плана не доказывает ASI01: нужен путь от недоверенного материала к изменённому решению агента.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент получает недоверенный ввод или сообщение другого агента, способное изменить его разрешённую цель.
Условие связи с карточкой	Инструкция атакующего меняет план или действие Agent Core.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	261	148
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	277	154
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	283	158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций	287	160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	292	162

Карточка	Связи	Стр.
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	297	164
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	302	167
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	313	173
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции	378	220
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	396	232
AISAFE2-WEAK-051 · Выдача результата без завершения обязательных стадий вычислений	400	235
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель	412	245
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	413	246
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	417	248
AISAFE2-WEAK-069 · Неопределённый режим переполнения и недоступности человеческой проверки	437	269
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	475	303
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	508	320
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	509	328
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	510	375
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	511	389
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	512	392
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	513	429
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	514	488
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	515	497
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	516	507
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	734	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	735	44
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	736	61
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	737	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	738	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	739	82

Воздействие · AISAFE2-IMP-008

Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)

Краткое описание

Инструмент, RAG или сторонний агент возвращает непредусмотренный результат либо выполняет действие за пределами ожидаемого контракта.

Подробное описание

Воздействие относится к нарушению сценария на границе инструмента и отличается от выбора неверного вызова самим Agent Core и от ответа модели. Примеры включают неожидаемые данные из базы знаний, деструктивное действие из-за избыточной автономности, появление незаявленного инструмента, взаимодействие с вредоносным агентом и отравленные данные в базе знаний. Меры противодействия сосредоточены на улучшении контракта или снижении доверия к используемому инструменту.

Класс воздействия

- I.2 Нарушение предусмотренных сценариев работы AI-системы

Наблюдаемые признаки

- Журнал инструмента, RAG или стороннего агента фиксирует результат, действие либо состав доступных инструментов, не соответствующий ожидаемому контракту.
- Трассировка подтверждает передачу этого результата или внешний эффект действия; простая ошибка доступности без такого эффекта исключается.

Области возможного ущерба

- Информационные активы
- Работа сервиса

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI02, Example Attack Scenarios, 1. Tool Poisoning** · Возможное последствие
>
> Подмена описания, схемы или маршрута легитимного инструмента во время работы может привести к результату или действию, не соответствующему объявленному контракту.
>
> **Граница связи:** Связь относится к подмене интерфейса инструмента. Ошибочный выбор неизменённого инструмента агентом сам по себе не означает это воздействие.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент вызывает легитимный инструмент и может применить его вне ожидаемого контракта.
Условие связи с карточкой	Нарушитель подменяет описание, схему или маршрутизацию легитимного инструмента, и фактический результат либо действие расходится с объявленным контрактом.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	262	148
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	303	167
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	314	173
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	397	232
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	414	246
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	517	324
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	518	339
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	519	375
AISAFE2-MIT-021 · Управление доверием и сроком жизни долговременной памяти	520	396
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	521	497
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	740	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	741	48
AISAFE2-COMP-007 · Долговременная память	742	50
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	743	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	744	75
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	745	82

Воздействие · AISAFE2-IMP-009

Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Краткое описание

Ответ инструмента, RAG или связанного источника раскрывает либо позволяет реконструировать чувствительные данные неуполномоченному получателю.

Подробное описание

Воздействие относится к классу раскрытия чувствительных данных и охватывает сведения, участвовавшие в формировании ответа инструмента: документы базы знаний, результаты поиска, память, телеметрию или журналы. Оно отличается от раскрытия инструкций, обучающих данных и компрометации всей базы знаний. Пример — выдача конфиденциального фрагмента внутренней документации, попавшего в базу знаний для RAG. Инструменты могут использовать публичные данные для формирования своих ответов; в этом случае их раскрытие не считается негативным воздействием.

Класс воздействия

- I.3 Раскрытие чувствительных данных

Наблюдаемые признаки

- В ответе обнаружен чувствительный фрагмент или устойчиво восстановлен его атрибут, а трассировка происхождения ведёт к инструменту, RAG, памяти или журналу.
- Проверка полномочий подтверждает, что получатель не имел права на сведения; одного чтения источника без выдачи недостаточно.

Области возможного ущерба

- Информационные активы

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	278	154
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	284	158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций	288	160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	293	162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	298	164
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	361	207
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	363	209
AISAFE2-WEAK-039 · Выдача инструментом чувствительных или избыточных данных	368	212

Карточка	Связи	Стр.
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	370	214
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	421	251
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели	444	276
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	453	286
AISAFE2-WEAK-086 · Утрата атрибутов доступа исходного материала при индексировании базы знаний	470	298
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	522	342
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	523	349
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	524	441
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	746	35
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	747	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	748	48
AISAFE2-COMP-007 · Долговременная память	749	50
AISAFE2-COMP-013 · Телеметрия и обнаружение	750	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	751	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	752	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	753	75

Воздействие · AISAFE2-IMP-010

Раскрытие содержимого системного промпта и инструкций инструментов

Краткое описание

Неуполномоченный субъект получает системный промпт или инструкции, описывающие поведение подключённых инструментов.

Подробное описание

Воздействие выделено по типу раскрываемых сведений: раскрывается системный промпт, инструкция разработчика или инструмента, определяющая поведение модели либо выбор инструментов, а не пользовательские или обучающие данные. Получение инструкции не всегда нарушает бизнес-логику и само по себе не доказывает обход ограничений. Однако выдача текста или его содержательно полной реконструкции субъекту, для которого он был закрыт, упрощает подготовку дальнейших атак.

Класс воздействия

- I.3 Раскрытие чувствительных данных

Наблюдаемые признаки

- В ответе обнаружен дословный или содержательно полный фрагмент системного промпта, инструкции разработчика либо инструмента.
- Политика назначения подтверждает, что получатель не должен был видеть этот текст; предположение о поведении без содержания инструкции исключается.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

> **AML.T0056** · Возможное последствие
>
> Получение закрытого системного промпта неуполномоченным адресатом соответствует воздействию успешного извлечения по `AML.T0056`.
>
> **Граница связи:** Карта не устанавливает способ извлечения и охватывает также инструкции инструментов.
>
Субъект	Аналитик техник атак на AI-систему
Условие применения	Закрытая системная инструкция доступна модели либо хранится в конфигурации, из которой её может получить неуполномоченный субъект.
>
> Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM08:2026** · Возможное последствие
>
> Раскрытие системного промпта или инструкций инструментов представляет конкретный исход утечки скрытого контекста.
>
> **Граница связи:** Не каждый скрытый фрагмент является инструкцией; само раскрытие не доказывает обход ограничений или последующую атаку.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Модель имеет доступ к скрытым инструкциям, контексту или описаниям инструментов, которые могут раскрыться.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	279	154
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	285	158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций	289	160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	294	162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	299	164
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	340	190
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	373	216
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции	379	220
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели	445	276
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	454	286
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	456	290
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	525	328
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	526	371
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	754	35
AISAFE2-COMP-003 · Управление контекстом общения с моделью	755	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	756	44
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	757	75

Воздействие · AISAFE2-IMP-011

Компрометация баз знаний

Краткое описание

Неуполномоченный субъект получает существенную часть базы знаний, её индекса или возможность воспроизвести их содержимое и свойства.

Подробное описание

Воздействие относится к раскрытию чувствительных данных, но выделяет компрометацию базы как целостного ценного актива. Оно применимо и тогда, когда отдельные документы не признаны чувствительными, но закрытый корпус, его структура или индекс как совокупность составляют чувствительный информационный актив: их копирование может лишить сервис конкурентной ценности. В отличие от единичной выдачи чувствительной информации через RAG здесь подтверждается доступ, экспорт или реконструкция существенной части базы.

Класс воздействия

- I.3 Раскрытие чувствительных данных

Наблюдаемые признаки

- Журнал доступа или экспорта показывает получение неуполномоченным субъектом существенной части документов, метаданных либо индекса базы знаний.
- Серия запросов позволяет воспроизвести содержимое или свойства закрытого корпуса, а контрольная проверка подтверждает совпадение.

Области возможного ущерба

- Информационные активы

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	304	167
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	327	184
AISAFE2-WEAK-023 · Незашифрованное хранение	333	186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	336	188
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	341	190
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	345	194
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	352	199
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	356	204
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	383	223
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	457	290
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	461	292
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	471	299
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	527	356
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	528	444
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	758	48
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	759	61
AISAFE2-COMP-023 · Подсистема формирования базы знаний	760	86

Воздействие · AISAFE2-IMP-012

Кража модели

Краткое описание

Неуполномоченный субъект получает веса или связанные артефакты модели либо воспроизводит её ценное поведение без разрешения владельца.

Подробное описание

Воздействие отличается от раскрытия архитектуры и свойств модели тем, что субъект получает сам артефакт или функционально эквивалентный результат. Прямой пример — доступ атакующего к весам; извлечение по ответам интерфейса учитывается, если воспроизведение подтверждено.

Класс воздействия

- I.3 Раскрытие чувствительных данных

Наблюдаемые признаки

- Аудит доступа фиксирует получение весов или связанного артефакта субъектом, отсутствующим среди утверждённых получателей.
- Для извлечения по интерфейсу независимая проверка подтверждает воспроизведение ценного поведения, а не только наблюдение отдельных ответов.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

> **AML.T0024.002** · Возможное последствие
>
> Подтверждённое воспроизведение ценного поведения модели по ответам API является одним из проявлений кражи модели.
>
> **Граница связи:** Карта включает и прямое получение артефактов; путь извлечения через API нужно подтвердить отдельно.

Субъект	Аналитик техник атак на AI-систему
Условие применения	Защищаемая модель доступна через API инференса, позволяющий многократно получать ответы для обучения её функциональной копии.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	305	167
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	318	176
AISAFE2-WEAK-018 · Избыточная наблюдаемость и воспроизводимость поведения модели через интерфейс	323	180
AISAFE2-WEAK-019 · Избыточное раскрытие внутренних оценок модели	325	182
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	328	184
AISAFE2-WEAK-023 · Незашифрованное хранение	334	186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	337	188
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	342	190
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	346	194
AISAFE2-WEAK-027 · Отсутствие экономического ограничения массовых запросов	350	196
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	353	199
AISAFE2-WEAK-031 · Недостаточная защита модели при поставке в недоверенную среду	355	203
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	357	204
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	384	223
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	406	240
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	458	290
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	462	292
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	485	308
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	529	481
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	530	485
AISAFE2-COMP-002 · Модель и интерфейс инференса	761	37
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	762	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	763	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	764	61
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	765	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	766	79

Воздействие · AISAFE2-IMP-013

Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Краткое описание

Модель или связанные артефакты раскрывают либо позволяют реконструировать чувствительные записи и атрибуты данных обучения или дообучения.

Подробное описание

Источник раскрываемых сведений отличает данное воздействие от раскрытия данных инструментов и системных инструкций. Воздействие подразумевает раскрытие информации, использованной при обучении. Иллюстративный путь — успешная атака восстановления данных модели; конкретная техника должна подтверждаться испытанием. Для подтверждения воздействия также важно подтвердить, что раскрытые данные не являются публичными.

Класс воздействия

- I.3 Раскрытие чувствительных данных

Наблюдаемые признаки

- В ответе модели или связанном артефакте обнаружена чувствительная запись либо её атрибуты, совпадение с которыми подтверждено владельцем датасета.
- Испытание на участие записи или реконструкцию даёт воспроизводимый результат для индивидуальных данных, а не только для агрегированной статистики.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM02:2026 · Возможное последствие

Воспроизведение или реконструкция непубличных обучающих записей через модель является конкретным исходом раскрытия чувствительной информации.

Граница связи: Воздействие касается источника в обучении; текущий контекст, журналы и инструменты остаются другими путями утечки.

Субъект	Аналитик угроз AI-приложения
Условие применения	AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-015 · Переобучение модели	316	175
AISAFE2-WEAK-019 · Избыточное раскрытие внутренних оценок модели	326	182
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	347	194
AISAFE2-WEAK-033 · Раскрытие сведений об отдельных обучающих записях через модель	360	206
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	362	207
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	364	209
AISAFE2-WEAK-037 · Непропорциональный вклад отдельных обучающих записей в результат обучения	367	211
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	371	214
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели	424	253
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	531	349
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	532	478
AISAFE2-COMP-002 · Модель и интерфейс инференса	767	37
AISAFE2-COMP-008 · Данные для подсистемы обучения	768	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	769	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	770	56
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	771	69

Воздействие · AISAFE2-IMP-014

Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных

Краткое описание

Неуполномоченный субъект узнаёт закрытые сведения об архитектуре, технологиях, алгоритмах AI-системы либо происхождении и свойствах её данных.

Подробное описание

Воздействие относится к раскрытию технических сведений об AI-системе, а не самих моделей, баз знаний или индивидуальных записей о запросах пользователей. Важная для функционирования AI-системы информация может быть прочитана напрямую или

реконструирована по поведению системы. Примеры — раскрытие архитектуры модели и вывод свойств датасета, если владелец действительно ограничивал доступ к этим сведениям.

Класс воздействия

- I.3 Раскрытие чувствительных данных

Наблюдаемые признаки

- Неуполномоченному субъекту выдана закрытая конфигурация, архитектурное описание, алгоритм или сведения о происхождении данных.
- По серии ответов восстановлено закрытое свойство модели либо датасета, и владелец информации подтверждает совпадение.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM08:2026** · Возможное последствие
>
> Раскрытие закрытых сведений об архитектуре или правилах работы из скрытого контекста даёт атакующему детали для дальнейших попыток воздействия.
>
> **Граница связи:** Карточка охватывает и технические сведения вне контекста модели; их раскрытие из другого источника не относится к LLM08.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Модель имеет доступ к скрытым инструкциям, контексту или описаниям инструментов, которые могут раскрыться.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	306	167
AISAFE2-WEAK-018 · Избыточная наблюдаемость и воспроизводимость поведения модели через интерфейс	324	180
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	425	254
AISAFE2-WEAK-061 · Избыточно открытая внутренняя телеметрия	426	255
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	533	371
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	772	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	773	37

Карточка	Связи	Стр.
AISAFE2-COMP-003 · Управление контекстом общения с моделью	774	40
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	775	48
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	776	56
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	777	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	778	75
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	779	84

Воздействие · AISAFE2-IMP-015

Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Краткое описание

Субъект успешно читает или изменяет хранимые данные AI-системы вне предоставленных ему полномочий, в том числе после повышения привилегий.

Подробное описание

Воздействие относится к классу компрометации компонентов AI-системы и фиксирует несанкционированный контроль над хранимыми данными, а не выполнение функции инфраструктуры. Оно охватывает базы знаний, датасеты, параметры, конфигурацию и сохранённый пользовательский контекст. Примеры — чтение контекста другого пользователя или неразрешённое изменение параметров модели.

Класс воздействия

- I.4 Компрометация компонента системы, получения контроля над ним

Наблюдаемые признаки

- Журнал хранилища содержит успешное чтение или изменение объекта субъектом, которому политика не предоставляла такую операцию.
- Для изменения контроль целостности фиксирует новое состояние объекта, связанное с несанкционированной операцией или повышением привилегий.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI03** · Возможное последствие
>
> Злоупотребление идентичностью или унаследованными правами агента может закончиться несанкционированным доступом к хранимым данным.
>
> **Граница связи:** Карточка фиксирует результат; цепочку делегирования, источник права и конкретный обход нужно подтвердить отдельно.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент использует собственную, делегированную или унаследованную идентичность и связанные права.
Условие связи с карточкой	Для чтения или изменения данных используется идентичность агента, делегированный секрет или унаследованное право за пределами полномочий исходного субъекта.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	263	148
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	307	167
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	319	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	329	184
AISAFE2-WEAK-023 · Незашифрованное хранение	335	186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	338	188
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	343	190
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	358	204
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	374	216
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	385	223
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	388	227
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	407	240
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	415	246
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	422	251
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	459	290

Карточка	Связи	Стр.
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	463	292
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	466	294
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	481	306
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	486	308
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	534	359
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	535	448
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	536	500
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	537	509
AISAFE2-COMP-007 · Долговременная память	780	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	781	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	782	56
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	783	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	784	64
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	785	75
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	786	79
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	787	84
AISAFE2-COMP-023 · Подсистема формирования базы знаний	788	86
AISAFE2-COMP-025 · Источники данных для инструментов	789	90

Воздействие · AISAFE2-IMP-016

Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Краткое описание

Субъект вызывает функцию инструмента или инфраструктурного компонента, на которую у него нет полномочий, включая доступ после повышения привилегий.

Подробное описание

Это воздействие класса компрометации компонентов AI-системы описывает несанкционированно полученную возможность выполнить функцию, а не прочитать или изменить хранимый объект. Примеры — вызов инструмента с более высокими правами,

действие от имени другого пользователя или удалённое выполнение кода в скомпрометированном компоненте. Для подтверждения воздействия фиксируется факт неразрешённого выполнения операции.

Класс воздействия

- I.4 Компрометация компонента системы, получения контроля над ним

Наблюдаемые признаки

- Журнал инструмента или инфраструктуры фиксирует успешное выполнение операции от имени субъекта, которому она не была разрешена.
- Аудит связывает операцию с повышенными правами, чужой учётной записью или скомпрометированным компонентом; одной уязвимости недостаточно.

Области возможного ущерба

- Информационные активы

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	264	148
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	308	167
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	330	184
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	339	188
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	344	190
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	359	204
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	389	227
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	408	240
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	416	246
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	423	251
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	460	290
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	464	292
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	467	294
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	476	303
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	482	306
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	487	308
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	538	359

Карточка	Связи	Стр.
AISAFE2-MIT-017 · Изолированное исполнение инструментов	539	379
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	540	383
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	541	434
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	542	438
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	790	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	791	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	792	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	793	61
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	794	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	795	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	796	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	797	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	798	82

Воздействие · AISAFE2-IMP-017

Неограниченное потребление токенов при обращении к языковым моделям

Краткое описание

Запрос, сессия или агентная цепочка продолжает расходовать токены языковой модели без действующего верхнего ограничения.

Подробное описание

Потребление измеряется токенами модели. Оно отличается от стоимости платных инструментов, загрузки собственных вычислений и расходов конечного пользователя. Неограниченный вход, вывод, повторные обращения или внутренние шаги могут исчерпать квоту и вытеснить другие запросы независимо от того, оплачиваются ли токены отдельно.

Класс воздействия

- I.5 Повышенное потребление (ресурсов, токенов)

Наблюдаемые признаки

- Настройки не задают верхний предел расхода токенов для запроса, сессии или цепочки либо трасса показывает, что заданный предел не сработал.
- Телеметрия квоты связывает исчерпание доступного остатка или вытеснение других запросов именно с токенным расходом этой обработки.

Области возможного ущерба

- Работа сервиса

- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Возможное последствие
>
> Неограниченный расход токенов языковой модели может исчерпать квоту и вытеснить другие запросы.
>
> **Граница связи:** Воздействие измеряется расходом токенов; платные инструменты, собственные вычисления и расходы пользователя учитываются отдельно.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	401	237
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	428	257
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	543	333
AISAFE2-COMP-002 · Модель и интерфейс инференса	799	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	800	42
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	801	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	802	72
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	803	79

Воздействие · AISAFE2-IMP-018

Незапланированные расходы на использование платных инструментов

Краткое описание

AI-система вызывает тарифицируемые инструменты или внешние сервисы и создаёт расходы, не предусмотренные бюджетом владельца.

Подробное описание

Воздействие относится к повышенному потреблению, но объектом учёта служат платные вызовы инструментов. Оно отличается от чрезмерного потребления токенов модели, собственных вычислительных ресурсов и неожиданных расходов пользователя AI-системы. Расходы могут возникнуть из-за повторных, неожиданных или неправильно спланированных обращений к инструментам.

Класс воздействия

- I.5 Повышенное потребление (ресурсов, токенов)

Наблюдаемые признаки

- В детализации начислений появляются вызовы инструментов или внешних сервисов, отсутствующие в утверждённом бюджете владельца, либо объём расходов на инструменты превышает запланированное значение.
- Трасса обработки связывает начисления с конкретными повторными, неожиданными или неправильно запланированными платными обращениями.

Области возможного ущерба

- Работа сервиса
- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Возможное последствие
>
> Повторные вызовы платных инструментов могут превратить неконтролируемую агентную цепочку в непредвиденные расходы владельца.
>
> **Граница связи:** Карточка касается цены внешних вызовов, а не расхода токенов или вычислений; фактическая стоимость зависит от тарифа и вызовов.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	402	237
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	418	248
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	429	257

Карточка	Связи	Стр.
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей	432	262
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	477	303
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	544	333
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	804	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	805	44
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	806	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	807	72
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	808	79

Воздействие · AISAFE2-IMP-019

Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Краткое описание

AI-система использует вычисления, память, сеть или ускорители своей среды сверх установленного операционного бюджета.

Подробное описание

Измеряется загрузка собственных или выделенных вычислительных мощностей и сокращение ёмкости для других задач, а не потребление токенов или платных инструментов. В облаке такое потребление может отражаться в счёте провайдера, а в собственном контуре — в нагрузке, энергопотреблении или износе оборудования.

Класс воздействия

- I.5 Повышенное потребление (ресурсов, токенов)

Наблюдаемые признаки

- Метрики вычислений, памяти, сети или ускорителей превышают операционный бюджет среды во время выполнения AI-цепочки.
- Одновременно сокращается ёмкость для других задач, растёт очередь или фиксируется вытеснение нагрузки; одного повышенного счёта недостаточно.

Области возможного ущерба

- Работа сервиса
- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Возможное последствие
>
> Чрезмерная нагрузка на память, сеть или ускорители AI-системы может лишить другие задачи доступной вычислительной ёмкости.
>
> **Граница связи:** Воздействие измеряется нагрузкой на собственные или выделенные ресурсы; причина всплеска и отдельные начисления за токены или инструменты устанавливаются отдельно.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	348	194
AISAFE2-WEAK-027 · Отсутствие экономического ограничения массовых запросов	351	196
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	354	199
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	403	237
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	430	257
AISAFE2-WEAK-075 · Отсутствие совокупных пределов масштабирования и расходов инфраструктуры	447	277
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	465	292
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	483	306
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	545	333
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	546	392
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	547	405
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	809	42
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	810	58
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	811	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	812	75

Воздействие · AISAFE2-IMP-020

Неожиданно высокие расходы пользователей AI-системы

Краткое описание

Использование AI-системы приводит к расходам потребителя, существенно превышающим показанную или разумно ожидаемую стоимость операции.

Подробное описание

Воздействие относится к классу повышенного потребления и актуально, когда вычисления AI-системы оплачивает пользователь, а не владелец AI-системы. Оно отличается от расходов на платные инструменты и собственную инфраструктуру владельца. Конкретный порог и модель оплаты задаются сервисом и должны быть известны пользователю до совершения операции.

Класс воздействия

- I.5 Повышенное потребление (ресурсов, токенов)

Наблюдаемые признаки

- В счёте потребителя стоимость AI-сервиса или последующих операций существенно превышает показанную до запуска цену или оценку либо, если её не показывали, обоснованное ожидание по действующему тарифу и сопоставимому объёму операции.
- Журнал сценария не содержит подтверждения потребителем более высокой стоимости или её верхней границы до начисления; расходы владельца системы отделены от суммы, начисленной потребителю.

Области возможного ущерба

- Работа сервиса
- Финансы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM06:2026 · Возможное последствие

Неконтролируемое потребление AI-сервиса может привести к счёту пользователя выше объявленной или ожидаемой стоимости операции.

Граница связи: Эффект зависит от модели оплаты и предупреждений сервиса; расходы владельца и атака на доступность являются другими исходами.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	419	248
AISAFE2-WEAK-063 · Недостаточные пользовательские средства управления расходами и прозрачность тарификации	431	260
AISAFE2-WEAK-093 · Неверная атрибуция тарифицируемого потребления между арендаторами	488	311
AISAFE2-WEAK-095 · Недостаточно ограниченные делегированные платёжные полномочия	489	314
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	548	333
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	813	35
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	814	72
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	815	79

Воздействие · AISAFE2-IMP-021

Зацикливание AI-системы при обработке запросов

Краткое описание

Обработка запроса не достигает конечного состояния и повторяет шаги, вызовы или самокоррекцию, удерживая ресурсы системы.

Подробное описание

Это воздействие является представителем класса отказа в обслуживании. Оно отличается от общего отсутствия результата тем, что наблюдаемой причиной является зацикливание при обработке пользовательского запроса, которое может происходить в вызовах модели или на других шагах обработки. Примеры — эксплуатация цикла принятия решений или цикла самоулучшения, при которых шаги повторяются без нового состояния и завершения.

Класс воздействия

- I.6 Отказ в обслуживании

Наблюдаемые признаки

- Один запрос повторяет уже пройденную последовательность шагов или одинаковые вызовы без нового состояния.
- Счётчик шагов, длительность или расход квоты растут до внешнего ограничения, тогда как штатное конечное состояние не достигается.

Области возможного ущерба

- Работа сервиса

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI08** · Возможное последствие
>
> Зацикливание может стать наблюдаемой частью каскадного сбоя, когда ошибочные шаги повторяются и распространяются между агентами.
>
> **Граница связи:** Один локальный цикл не устанавливает ASI08; нужно подтвердить исходный сбой и его распространение за пределы одного пути.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Взаимодействующие агенты и зависимые операции могут усилить исходный сбой или повтор.
Условие связи с карточкой	Сбой запускает повторы или петли обратной связи между агентами, сессиями либо процессами.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	280	154
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	404	237
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	549	333
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	550	392
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	816	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	817	44
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	818	72
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	819	82

Воздействие · AISAFE2-IMP-022

Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Краткое описание

Потребитель не может использовать AI-систему или не получает результат в установленный для операции срок.

Подробное описание

Это негативное воздействие класса отказа в обслуживании фиксируется, когда потребитель не получает доступ или результат к установленному сроку. Оно может происходить по разным причинам, например из-за отказа ключевого инфраструктурного компонента или повторяющейся ошибки инструмента. В зависимости от назначения AI-системы воздействия такого класса могут приводить к бизнес-последствиям как низкой, так и высокой критичности.

Класс воздействия

- I.6 Отказ в обслуживании

Наблюдаемые признаки

- Проверка доступности или журнал запроса фиксирует отсутствие доступа либо результата к установленному для операции сроку.
- Трассировка показывает отказ зависимости, тайм-аут или повторяющуюся ошибку; задержка до истечения установленного срока исключается.

Области возможного ущерба

- Работа сервиса

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI08** · Возможное последствие
>
> Распространение ошибки между агентами может привести к недоступности AI-системы для потребителя.
>
> **Граница связи:** Недоступность сама по себе не показывает каскад; нужны свидетельства передачи и усиления исходного сбоя.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Взаимодействующие агенты и зависимые операции могут усилить исходный сбой или повтор.
Условие связи с карточкой	Первичный сбой агента распространяется на зависимые операции и вызывает более широкий отказ сервиса.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	349	194
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	382	222
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	390	227

Карточка	Связи	Стр.
AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление	393	229
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	405	237
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей	433	262
AISAFE2-WEAK-069 · Неопределённый режим переполнения и недоступности человеческой проверки	438	269
AISAFE2-WEAK-076 · Отсутствие ресурсной изоляции арендаторов в разделяемой среде	448	280
AISAFE2-WEAK-077 · Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку	449	281
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке	450	282
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения	451	284
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	551	333
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	552, 553	410
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	554	492
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	820	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	821	37
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	822	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	823	48
AISAFE2-COMP-007 · Долговременная память	824	50
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	825	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	826	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	827	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	828	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	829	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	830	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	831	82

Воздействие · AISAFE2-IMP-023

Перегрузка контура человеческой проверки

Краткое описание

Поток заданий на проверку или принятие решения превышает пропускную способность доступных специалистов.

Подробное описание

Это самостоятельный представитель класса отказа в обслуживании: ограничивающим ресурсом становится человек, а не вычислительная среда. Перегрузка может накопить просрочки, вызвать усталость или сделать ручной контур недоступным, тем самым усиливая доверие к непроверенному результату вычислений AI-системы.

Класс воздействия

- I.6 Отказ в обслуживании

Наблюдаемые признаки

- Длина очереди и время ожидания растут сверх утверждённых порогов при полной загрузке доступных компетентных участников HITL.
- В операционных записях появляются просрочки, поверхностные подтверждения, возвраты или переход к незапланированному режиму деградации.

Области возможного ущерба

- Человек
- Работа сервиса

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-065 · Избыточная маршрутизация задач на человеческую проверку	434	265
AISAFE2-WEAK-066 · Недостаточная мощность и резервирование человеческой проверки	435	266
AISAFE2-WEAK-067 · Недостаточная инструментальная поддержка оператора HITL	436	267
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	555	459
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	832	65

Воздействие · AISAFE2-IMP-024

Утечка чувствительных данных в недоверенный контур

Краткое описание

Чувствительные данные пересекают утверждённую границу доверия и поступают модели, инструменту или иному получателю, которому их передавать нельзя.

Подробное описание

Утечку определяют направление потока и доверенность получателя, а не источник или способ реконструкции данных. Недоверенность определяется архитектурой и политикой. Примеры — передача данных недоверенной модели или отправка сведений, включая учётные данные,

чужому инструменту при его использовании. Перехват межкомпонентного трафика относится к этому воздействию только тогда, когда чувствительные данные в результате фактически получает неуполномоченный адресат.

Класс воздействия

- I.7 Утечка чувствительных данных в недоверенный контур

Наблюдаемые признаки

- Контроль исходящего потока фиксирует чувствительное значение в запросе к модели, инструменту или иной среде за утверждённой границей доверия.
- Для получателя отсутствует действующее разрешение на этот класс данных и назначение обработки; разрешённый переход даёт контрпример.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

AML.T0086 · Возможное последствие

Передача чувствительных данных через аргументы вызова инструмента недоверенному получателю является одним из путей утечки за границу доверия.

Граница связи: Для AML.T0086 нужны инструмент записи и получатель под контролем нарушителя; карточка охватывает и другие пути утечки.

Субъект	Аналитик техник атак на AI-систему
Условие применения	Агент может передавать данные через аргументы инструмента записи получателю под контролем нарушителя.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM02:2026 · Возможное последствие

Передача чувствительных данных модели или коннектору, которому их передавать нельзя, является конкретной утечкой через AI-путь.

Граница связи: Карточка не устанавливает, были ли данные чувствительными и разрешён ли получатель; это определяется для фактического потока.

Субъект	Аналитик угроз AI-приложения
Условие применения	AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	265	148
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	309	167
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	365	209
AISAFE2-WEAK-039 · Выдача инструментом чувствительных или избыточных данных	369	212
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	372	214
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции	380	220
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	398	232
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	420	248
AISAFE2-WEAK-061 · Избыточно открытая внутренняя телеметрия	427	255
AISAFE2-WEAK-070 · Отсутствие управляемой политики допустимого использования AI-сервисов	439	270
AISAFE2-WEAK-071 · Отсутствие явной модели доверенных зон и потоков данных	440	271
AISAFE2-WEAK-072 · Отсутствие обязательного технического контроля межзонных потоков данных	441	273
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели	446	276
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	455	286
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	472	299
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	478	303
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	484	306
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	556	328
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	557	349
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	558	410
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	559	441
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	560	500
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	561	507

Карточка	Связи	Стр.
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	562	519
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	833	35
AISAFE2-COMP-003 · Управление контекстом общения с моделью	834	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	835	44
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	836	69
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	837	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	838	82

Воздействие · AISAFE2-IMP-025

Прямые финансовые потери

Краткое описание

Организация или потребитель несёт подтверждённый денежный убыток, непосредственно возникший вследствие технического воздействия, связанного с AI-системой.

Подробное описание

Это высокоуровневое последствие, к которому могут привести разные пути эксплуатации недостатков AI-системы. Например, чрезмерная нагрузка на вычислительный узел ускоряет износ оборудования, что может потребовать его досрочной замены; неконтролируемые обращения к платным сервисам могут привести к незапланированному увеличению расходов на оплату сторонних ресурсов. Упущенная выгода и предположительная будущая стоимость не относятся к этому воздействию.

Класс воздействия

- Высокоуровневые бизнес-последствия

Наблюдаемые признаки

- Финансовые документы фиксируют списание, платёж, ремонт или замену актива, которых не было бы без соответствующего события AI-системы.
- Расследование подтверждает сумму и причинный путь от технического воздействия к расходу; одного роста потребления вычислительных мощностей без зафиксированного убытка недостаточно.

Области возможного ущерба

- Финансы

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-095 · Недостаточно ограниченные делегированные платёжные полномочия	490	314

Воздействие · AISAFE2-IMP-026

Нарушение операционной деятельности, организационных процессов или бизнес-процессов

Краткое описание

Работа AI-системы приводит к остановке, задержке или неверному выполнению операции, организационного процесса либо бизнес-процесса.

Подробное описание

Это высокоуровневое последствие, к которому могут вести разные технические события и которое зависит от области применения AI-системы. Иллюстративные, неисчерпывающие сценарии: ошибка AI-поддерживаемой логистики приводит к столкновению транспорта и экологическому ущербу; неверное решение, принятое на основе рекомендаций AI-системы в процессе здравоохранения, нарушает оказание помощи и создаёт угрозу жизни.

Класс воздействия

- Высокоуровневые бизнес-последствия

Наблюдаемые признаки

- Журнал процесса фиксирует прерванный, задержанный или неверно выполненный этап и конкретное отклонение от целевого состояния.
- В ходе расследования инцидента трассировка подтверждает участие AI-системы в причинном пути; локальный сбой без влияния на операцию или процесс исключается.

Области возможного ущерба

- Работа сервиса
- Организация

Воздействие · AISAFE2-IMP-027

Несоответствие требованиям регуляторов

Краткое описание

AI-система или связанный с ней процесс не выполняет применимое обязательное требование для конкретного сценария, конфигурации и юрисдикции.

Подробное описание

Это высокоуровневое последствие может наступать вследствие различных ситуаций, вызванных ошибками конфигурации AI-системы или сопутствующих процессов, а также действиями атакующего. Иллюстративный сценарий: обработка персональных данных не соответствует применимым обязательным требованиям. Применимость конкретных требований регуляторов и факт нарушения требуют отдельной правовой оценки.

Класс воздействия

- Высокоуровневые бизнес-последствия

Наблюдаемые признаки

- Результат уполномоченной проверки или решение о выпуске фиксирует несоответствие конкретному применимому обязательному требованию.
- Материалы оценки связывают нарушение с конкретной AI-системой, её конфигурацией и рассматриваемым периодом; одной нехватки внутренних документов недостаточно.

Области возможного ущерба

- Правовые и нормативные обязательства

Воздействие · AISAFE2-IMP-028

Нарушение этических норм

Краткое описание

Применение AI-системы приводит к последствию, противоречащему утверждённым этическим нормам для организации, общества или иных затронутых субъектов.

Подробное описание

Нарушение этических норм шире непосредственного этического вреда пользователю и может возникнуть по разным техническим путям. Иллюстративный, неисчерпывающий сценарий: автоматизированное решение систематически ставит социальную группу в неблагоприятное положение, что подтверждено оценкой по заранее утверждённым этическим критериям.

Класс воздействия

- Высокоуровневые бизнес-последствия

Наблюдаемые признаки

- Оценка по утверждённым этическим критериям фиксирует нарушение и называет затронутых субъектов, масштаб и фактическое последствие.
- Материалы надзора или исследования последствий подтверждают устойчивый эффект на уровне организации либо общества; однако правовой вывод отсюда автоматически не следует.

Области возможного ущерба

- Этические последствия

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы	474	301

Воздействие · AISAFE2-IMP-029

Репутационный ущерб

Краткое описание

Связанное с AI-системой событие измеримо снижает доверие к организации, продукту или сервису у значимой аудитории.

Подробное описание

Это высокоуровневое последствие не следует автоматически из какого-либо инцидента, однако может возникнуть при разных сценариях атак или непосредственно из-за недостатков AI-системы. Иллюстративный, неисчерпывающий сценарий: публично известный неверный ответ AI-сервиса связывают с владельцем продукта, после чего клиенты отказываются от сервиса, а партнёры пересматривают сотрудничество.

Класс воздействия

- Высокоуровневые бизнес-последствия

Наблюдаемые признаки

- После огласки события ухудшается заранее выбранная метрика доверия у определённой аудитории по сравнению с исходным периодом.
- Фиксируется измеримый отказ от продукта или ухудшение отношений с клиентами либо партнёрами, связанное с этим событием.

Области возможного ущерба

- Репутация

Воздействие · AISAFE2-IMP-030

Утеря конкурентного преимущества

Краткое описание

Организация утрачивает исключительную ценность AI-актива, знаний или продукта, и её конкурентное положение измеримо ухудшается.

Подробное описание

Это высокоуровневое последствие разных путей эксплуатации недостатков, не являющееся синонимом раскрытия какого-то конкретного актива. Например, копирование базы знаний позволяет конкуренту воспроизвести ценную функцию сервиса; кража модели сокращает преимущество, которое давало её уникальное поведение. Для фиксации данного воздействия требуется подтверждённое влияние на конкурентную позицию.

Класс воздействия

- Высокоуровневые бизнес-последствия

Наблюдаемые признаки

- После утраты актива его существенная функция или знание появляется у конкурента либо в конкурирующем продукте.

- Заранее выбранный показатель конкурентной позиции ухудшается относительно исходного периода, и анализ связывает изменение с утратой исключительности.

Области возможного ущерба

- Финансы
- Конкурентная позиция

Воздействие · AISAFE2-IMP-031

Утрата датасета в результате атаки

Краткое описание

Атака приводит к уничтожению или невосстановимому повреждению датасета, необходимого для обучения, тестирования либо валидации AI-системы.

Подробное описание

Атакующий уничтожает, шифрует, перезаписывает или необратимо повреждает датасет и доступные для восстановления копии, из-за чего набор нельзя использовать по назначению. Объектом являются данные обучения, тестирования или валидации; причинная связь с атакой обязательна.

Класс воздействия

- I.4 Компрометация компонента системы, получения контроля над ним

Наблюдаемые признаки

- Контроль целостности фиксирует уничтожение, шифрование, перезапись или необратимое повреждение датасета после событий, классифицированных расследованием как атака.
- Проверка восстановления показывает, что доступные копии также непригодны или отсутствуют, поэтому набор нельзя использовать для обучения, тестирования либо валидации.

Области возможного ущерба

- Информационные активы
- Работа сервиса

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(c) · Возможное последствие

Невосстановимая утрата датасета с персональными данными в результате атаки показывает сценарий, который должна учитывать проверка возможности вернуть доступ к данным после инцидента.

Граница связи: Карточка относится только к набору обучения или оценки; состав данных, применимость GDPR и требуемый срок восстановления устанавливаются отдельно.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Утраченный набор содержит персональные данные в области GDPR, и другого пригодного пути восстановления нет.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	320	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	331	184
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	386	223
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	391	227
AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление	394	229
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	468	296
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	479	303
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных	563	503
AISAFE2-COMP-008 · Данные для подсистемы обучения	839	52

Воздействие · AISAFE2-IMP-032

Несанкционированное смешение артефактов между изолированными контекстами

Краткое описание

Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Подробное описание

Изолированным контекстом может быть арендатор, пользователь, домен безопасности, задача или сессия. Воздействие является разновидностью несанкционированного доступа, но чтение данных другого пользователя остаётся отдельным воздействием AISAFE2-IMP-015, а передача конфиденциальных сведений получателю за пределами разрешённой границы доверия — AISAFE2-IMP-024. Оно может происходить как из-за намеренных действий атакующего, так и непосредственно из-за ошибки конфигурации компонентов AI-системы. Категория утечки чувствительных данных применима только тогда, когда подмешанный артефакт содержит такие сведения и они выданы получателю вне разрешённой границы доверия; смешение остальных артефактов остаётся нарушением изоляции.

Класс воздействия

- I.4 Компрометация компонента системы, получения контроля над ним
- I.7 Утечка чувствительных данных в недоверенный контур

Наблюдаемые признаки

- В ответе, кэше или собранном контексте обнаружен артефакт с идентификатором другого арендатора, пользователя, домена безопасности, задачи либо сессии.
- Трасса маршрутизации показывает выдачу или подмешивание чужого запроса, ответа либо промежуточного состояния без разрешения; отдельное чтение сохранённого объекта по неверному ACL учитывается отдельно.

Области возможного ущерба

- Информационные активы

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Возможное последствие
>
> Подмешивание ответа, сессии или состояния одного арендатора в контекст другого может раскрыть сведения через общий AI-контур.
>
> **Граница связи:** Смешение артефактов без чувствительного или закрытого содержимого не доказывает раскрытие защищённой информации.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-050 · Остаточные данные в общей инфраструктуре исполнения моделей	399	234
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	564	349
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	565	405
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	566	410
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	567	500
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	840	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	841	37
AISAFE2-COMP-003 · Управление контекстом общения с моделью	842	40
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	843	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	844	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	845	48
AISAFE2-COMP-007 · Долговременная память	846	50
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	847	58
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	848	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	849	75
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	850	82
AISAFE2-COMP-024 · Артефакты работы AI-системы	851	88

Воздействие · AISAFE2-IMP-033

Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Краткое описание

Атака уничтожает или необратимо повреждает критичные данные либо артефакты AI-системы, не являющиеся датасетами, и не оставляет пригодной версии для восстановления.

Подробное описание

Объектом может быть точная версия модели или её весов, подготовленная база знаний и индекс либо критичная конфигурация, необходимая для восстановления системы. Воздействие наступает, когда атака уничтожает, шифрует, перезаписывает или необратимо повреждает рабочий объект и все доступные пригодные версии. Причинная связь с атакой и безуспешная проверка восстановления обязательны.

Класс воздействия

- I.4 Компрометация компонента системы, получения контроля над ним

Наблюдаемые признаки

- Расследование связывает уничтожение, шифрование, перезапись или необратимое повреждение точной версии критичного объекта, не являющегося датасетом, с атакой.
- Проверка восстановления подтверждает, что доступные версии, копии и доверенный рецепт отсутствуют или непригодны, поэтому требуемое состояние AI-системы восстановить нельзя.

Области возможного ущерба

- Информационные активы
- Работа сервиса

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(c) · Возможное последствие

Утрата в результате атаки базы знаний или индекса с персональными данными при отсутствии пригодной копии показывает возможный отказ восстановления доступа после инцидента.

Граница связи: Потеря одних весов модели или конфигурации не равна потере персональных данных; состав объекта и срок восстановления проверяются отдельно.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Утраченный корпус или индекс содержит персональные данные либо необходим для доступа к ним, а пригодной восстанавливаемой версии нет.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	321	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	332	184
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	387	223
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	392	227
AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление	395	229
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	469	296
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	480	303
AISAFE2-MIT-009 · Версионирование и откат базы знаний	568	346
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	569	448
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	570	568
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	852	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	853	67
AISAFE2-COMP-023 · Подсистема формирования базы знаний	854	86

Недостатки

Недостаток · AISAFE2-WEAK-001

Избыточное доверие пользователя к AI-выводу

Краткое описание

Убедительный AI-вывод становится непропорционально значимым основанием решения и снижает независимую проверку со стороны пользователя, даже если система прямо не гарантирует его правильность.

Подробное описание

Недостаток проявляется, когда интерфейс или процесс позволяет действовать по беглому и уверенному выводу модели без сверки существенных фактов, исходных данных и альтернативных оснований. Независимая проверка требует времени, усилий и знаний; если они недоступны пользователю, убедительная форма ответа или приписывание системе человеческой компетентности дополнительно смещают решение в пользу AI-вывода.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-024 · Доверие человека и прозрачность интерфейса

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Избыточное доверие пользователя к AI-выводу может привести к выполнению пользователем предложенного системой вредоносного действия, потому что уверенная форма ответа и приписываемая системе компетентность воспринимаются как признак достоверности, и пользователь запускает предложенный код, команду или настройку без независимой проверки.

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Избыточное доверие пользователя к AI-выводу может привести к ущербу от решения, принятого на основе результата AI-системы, потому что пользователь доверяет системе как эксперту и принимает решение по уверенно сформулированному ответу без независимой перепроверки, превращая ошибку результата в ошибку решения.

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Избыточное доверие пользователя к AI-выводу может привести к вовлечению пользователя в злонамеренную схему, потому что пользователь тем охотнее следует указаниям, что доверяет уверенным ответам системы, и при воздействии злоумышленника легче вовлекается в мошеннические схемы.

AISAFE2-IMP-004 · Передача потребителем чувствительных данных

Избыточное доверие пользователя к AI-выводу может привести к передаче потребителем чувствительных данных, потому что пользователь доверяет системе как помощнику и добровольно передаёт секреты и персональные данные, когда их запрос выглядит естественной частью диалога.

Условия проявления

1. **Предпосылка.** AI-вывод недостоверен или сформирован под влиянием злоумышленника, но изложен уверенно и убедительно.
2. **Триггер.** Пользователь принимает на его основании значимое решение или выполняет предложенное действие.
3. **Разрешающий пробел.** Пользователь не проводит независимую проверку, и рабочий процесс её не требует.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI09** · Диагностический признак сбоя
>
> Избыточное доверие к убедительному ответу агента облегчает принятие решения без независимой проверки.
>
> **Граница связи:** Карточка описывает склонность доверять ответу. Целенаправленное манипулирование и влияние на конкретное решение человека устанавливаются отдельно.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Человек принимает решение на основе представления, сформированного агентом.
Условие связи с карточкой	Человек принимает рекомендацию агента без достаточной независимой проверки.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1	35
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	2	65
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	250	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	251	93
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	252	95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	253	97
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	571	488

Недостаток · AISAFE2-WEAK-002

Отсутствие привязки утверждений AI-ответа к источникам

Краткое описание

Для утверждений AI-ответа отсутствует проверяемая связь с фактически использованными источниками и статусом подтверждения.

Подробное описание

Оно существенно, когда пользователь принимает ответ за факт или использует его в решении, не имея возможности восстановить использованные источники и проверить основание утверждения. Связь может не фиксироваться при извлечении и обработке либо теряться до отображения или экспорта результата.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-024 · Доверие человека и прозрачность интерфейса

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Отсутствие привязки утверждений AI-ответа к источникам может привести к выполнению пользователем предложенного системой вредоносного действия, потому что без проверяемой связи утверждений с фактически использованными источниками пользователь не может восстановить основание рекомендации и действует по непроверяемому выводу, принятому за факт.

Ограничение: Только если отсутствие проверяемого источника повлияло на доверие к совету, пользователь выполнил предложенное действие и понёс вред

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Отсутствие привязки утверждений AI-ответа к источникам может привести к ущербу от решения, принятого на основе результата AI-системы, потому что утверждения ответа не сопровождаются надёжными ссылками на фактически использованные источники и статусом подтверждения, по которым решение можно было бы перепроверить.

Ограничение: Только если отсутствие проверяемого источника стало существенным для исполненного решения по AI-результату, после которого наступил вред

Условия проявления

1. **Предпосылка.** До применения у получателя нет иного способа проверить основание существенного утверждения.
2. **Триггер.** Получатель опирается на утверждение при принятии решения или выполнении действия.
3. **Разрешающий пробел.** Существенное утверждение AI-ответа не связано проверяемым образом с фактически использованным источником.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	3	35
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	4	48
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	254	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	255	93
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	572	331

Недостаток · AISAFE2-WEAK-003

Недостаточная проверка AI-вывода на границе доверия

Краткое описание

Между формированием AI-вывода и его потребителем — интерфейсом, API, хранилищем, другой моделью или агентом — нет обязательного рубежа, который проверяет соблюдение заданного контракта и формализуемых ограничений содержимого и отклоняет результат при их нарушении.

Подробное описание

Недостаток активируется, когда вывод становится разметкой, кодом, параметром действия, пользовательским содержимым или входом другого компонента без принудительной проверки и отклонения результата, нарушающего заданный контракт или формализуемое ограничение.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-013 · Проверка вывода и контекстная изоляция

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Недостаточная проверка AI-вывода на границе доверия может привести к выполнению пользователем предложенного системой вредоносного действия, потому что между формированием вывода и его выдачей нет обязательного рубежа, проверяющего допустимость содержания и отклоняющего недопустимый результат, и опасное предложение беспрепятственно достигает пользователя.

Ограничение: Только если опасный совет прошёл непроверенную границу выдачи, пользователь его выполнил и понёс вред

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Недостаточная проверка AI-вывода на границе доверия может привести к вовлечению пользователя в злонамеренную схему, потому что система не отфильтровывает убеждающий обманный текст на границе выдачи, и склоняющее к действиям в интересах злоумышленника содержимое беспрепятственно достигает пользователя.

Ограничение: Только при подтверждённом изменении поведения пользователя в интересах злонамеренной схемы; создание вредоносного текста без вовлечения человека не входит.

AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы

Недостаточная проверка AI-вывода на границе доверия может привести к нарушению этических норм по отношению к пользователю, потому что ответ не проходит обязательную проверку допустимости содержания перед выдачей, и оскорбительное или недопустимое обращение беспрепятственно достигает человека.

Ограничение: Только если пропущенный без проверки AI-вывод вызвал ответ, решение или обращение с пользователем, нарушившие утверждённые для сценария этические нормы и причинившие ему вред

Условия проявления

1. **Предпосылка.** Вывод модели может оказаться некорректным или опасным из-за вероятностной генерации или недоверенного контекста.
2. **Триггер.** Вывод передаётся потребителю, способному его отобразить, сохранить, исполнить или использовать в работе: браузеру, API, хранилищу, другой модели или агенту.
3. **Разрешающий пробел.** На пути к потребителю отсутствует контроль, который проверяет соблюдение заданного контракта и формализуемых ограничений содержимого и отклоняет результат при их нарушении.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM10:2026** · Диагностический признак сбоя
>
> Отсутствие обязательной проверки вывода перед следующим компонентом позволяет недопустимому результату модели стать действием или активным содержимым.
>
> **Граница связи:** Конкретный вред зависит от получателя и его контракта; карточка не доказывает исполнение опасного результата.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | Вывод модели используется получателем, инструментом или приложением для последующего действия. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	5	35
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	6	44
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	7	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	8	77
AISAFE2-COMP-024 · Артефакты работы AI-системы	9	88
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	256	92
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	257	95
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	258	98
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	573	320
AISAFE2-MIT-056 · Исполнение версионируемой политики допустимости содержимого	574	529

Недостаток · AISAFE2-WEAK-004

Отсутствие контекстного кодирования или изоляции

Краткое описание

Содержимое модели попадает в активный синтаксический или исполняемый контекст без соответствующего экранирования либо изоляции.

Подробное описание

Оно проявляется при переносе вывода между контекстами с разными правилами синтаксиса и исполнения, когда отсутствуют соответствующее экранирование, изолированная среда, предварительный просмотр или безопасный пробный запуск.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-013 · Проверка вывода и контекстная изоляция

Прямые причинные воздействия

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Недостаточность санитизации или изоляции вывода модели может привести к вовлечению пользователя в злонамеренную схему, потому что не нейтрализованная активная разметка из вывода модели исполняется в браузере пользователя (XSS) и в доверенном интерфейсе системы показывает ему обманное содержимое, формы и ссылки, уводящие в мошеннический сценарий.

Ограничение: Только если активный вывод создал обманное содержимое в доверенном интерфейсе и пользователь после этого вступил в злонамеренную схему; одно срабатывание XSS без вовлечения недостаточно

AISAFE2-IMP-004 · Передача потребителем чувствительных данных

Неизолированный активный вывод создаёт обманную форму в доверенном интерфейсе; пользователь фактически вводит чувствительное значение, которое затем попадает в модель либо инструмент без допустимого основания.

Ограничение: Только когда пользователь действительно вводит чувствительное значение в созданную активным выводом форму и это значение входит в контекст модели или инструмента без допустимого основания. Непосредственная XSS-эксфильтрация остаётся ветвью I024, а обман без передачи значения — I003.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Недостаточность санитизации или изоляции вывода модели может привести к обходу ограничений и предусмотренной логики агента, потому что вывод модели, разбираемый оркестратором как структурированный формат, шаблон или конфигурация, без нейтрализации позволяет внедрённым конструкциям вмешаться в дальнейший ход сценария.

AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)

Недостаточность санитизации или изоляции вывода модели может привести к неожидаемому поведению слоя инструментов, потому что сформированное моделью содержимое попадает во входные контексты инструментов и интерпретаторов без обработки по правилам этих контекстов, и внедрённые конструкции заставляют исправный компонент выполнить непредусмотренную операцию.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостаточность санитизации или изоляции вывода модели может привести к несанкционированному чтению или изменению хранимых данных, потому что внедрённые в вывод конструкции запросов и путей (SQL-инъекция, обход файловых путей) исполняются хранилищем и файловой системой с правами приложения, открывая чтение и изменение данных вне полномочий инициатора.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Сформированный моделью вывод попадает в командный или иной исполняемый контекст инструмента без обработки по правилам этого контекста и выполняется с полномочиями принимающего сервиса; это позволяет инициатору вызвать инфраструктурную функцию, на которую у него нет права.

Ограничение: Только при фактическом вызове неразрешённой инициатору инфраструктурной функции; отображение разметки или исполнение в браузере само по себе этого не доказывает.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Если ответ модели вставляет чувствительное значение в адрес автоматически загружаемого ресурса или вызывает переход по такому адресу, небезопасная отрисовка может отправить значение внешнему узлу.

Ограничение: Только если клиент действительно выполняет сетевой запрос к получателю, которому значение передавать нельзя. Обычная ссылка без перехода и отображение текста не доказывают утечку.

Условия проявления

1. **Предпосылка.** В вывод модели могут быть внедрены активные конструкции.
2. **Триггер.** Вывод переносится в контекст, где содержимое интерпретируется как разметка, запрос, команда, конфигурация или код.
3. **Разрешающий пробел.** Не применяются санитизация по правилам принимающего контекста, изолированная среда или безопасный пробный запуск.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM10:2026** · Диагностический признак сбоя
>
> Вывод без кодирования для HTML, скрипта, SQL или терминала может быть истолкован как управляющее содержимое получателя.
>
> **Граница связи:** Риск возникает при фактическом переносе в активный контекст; неправильный смысл обычного текста относится к иной категории.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется получателем, инструментом или приложением для последующего действия.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	10	35
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	11	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	12	58
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	13	72
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	259	95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	260	97
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	261	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	262	104
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	263	116

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	264	118
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	265	130
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	575	320
AISAFE2-MIT-017 · Изолированное исполнение инструментов	576	379

Недостаток · AISAFE2-WEAK-005

Недостаточная маркировка и предупреждения в интерфейсе

Краткое описание

Представление ответа не позволяет пользователю различить AI-генерацию, подтверждённый факт и результат, требующий проверки перед применением.

Подробное описание

Недостаток становится существенным, когда человек принимает решение, публикует, передаёт или запускает результат, не видя его происхождения, ограничений, статуса проверки и предупреждения о значимом риске.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-024 · Доверие человека и прозрачность интерфейса

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Недостаточность маркировки и предупреждений в интерфейсе может привести к выполнению пользователем предложенного системой вредоносного действия, потому что интерфейс не помечает результат как сгенерированный и требующий проверки и не создаёт точку осознанного решения перед значимым шагом, и человек запускает непроверенное предложение как готовое.

Ограничение: Только если отсутствие маркировки или предупреждения повлияло на выполнение опасного совета и действие причинило вред

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Недостаточность маркировки и предупреждений в интерфейсе может привести к ущербу от решения, принятого на основе результата AI-системы, потому что представление ответа не предупреждает об ограничениях, статусе проверки и рисках дальнейшего использования, и человек утверждает и передаёт непроверенный результат как готовый.

Ограничение: Только если неуказанный статус проверки или риск повлиял на исполненное решение и наступило вредное последствие

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Если злоумышленник подсовывает AI-результат как подтверждённое основание, отсутствие ясной маркировки и статуса проверки повышает доверие пользователя к непроверенным сведениям и может вовлечь его в обман.

Ограничение: Только при подтверждённом изменении поведения пользователя в интересах злонамеренной схемы; создание вредоносного текста без вовлечения человека не входит.

AISAFE2-IMP-004 · Передача потребителем чувствительных данных

Недостаточная маркировка и предупреждения в интерфейсе могут привести к передаче потребителем чувствительных данных: пользователь не видит границ сценария обработки и передаёт секреты или персональные сведения, не понимая, что они не нужны или не разрешены для этого сценария.

Ограничение: Только когда чувствительное значение из пользовательского ввода или файла действительно попадает в контекст без допустимого основания.

Условия проявления

1. **Предпосылка.** Интерфейс показывает AI-результат без признаков происхождения, ограничений и статуса проверки.
2. **Триггер.** Пользователь применяет результат в значимой операции: публикует, передаёт или запускает его.
3. **Разрешающий пробел.** Маршрут применения не содержит точки осознанного подтверждения риска.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	14	35
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	266	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	267	93
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	268	95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	269	97
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	577	331

Недостаток · AISAFE2-WEAK-006

Отсутствие раскрытия факта взаимодействия с AI

Краткое описание

Пользователь не получает явного сигнала, что взаимодействует с AI или видит созданный им материал.

Подробное описание

Пользователь не получает явного сигнала, что общается с AI-системой или видит созданный ею материал: сообщение модели внешне неотличимо от сообщения человека или уже проверенного результата сервиса. В результате пользователь неверно калибрует доверие: относится к сгенерированному содержимому как к проверенному людьми и принимает на его основе решения, а злоумышленнику проще выдавать AI-агент за человека в схемах обмана и социальной инженерии.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-024 · Доверие человека и прозрачность интерфейса

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Отсутствие раскрытия факта взаимодействия с AI может привести к выполнению пользователем предложенного системой вредоносного действия, потому что без явного сигнала о взаимодействии с AI пользователь относится к сгенерированному содержимому как к проверенному людьми и неверно калибрует доверие при принятии решения.

Ограничение: Только если пользователь не знал об AI-происхождении совета, из-за этого доверился ему и выполнил опасное действие с вредом

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Отсутствие раскрытия факта взаимодействия с AI может привести к ущербу от решения, принятого на основе результата AI-системы, потому что без явного сигнала о взаимодействии с AI пользователь относится к сгенерированному содержимому как к проверенному людьми и неверно калибрует доверие при принятии решения.

Ограничение: Только если сокрытие AI-происхождения результата повлияло на исполненное решение и наступил вред

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Отсутствие раскрытия факта взаимодействия с AI может привести к вовлечению пользователя в злонамеренную схему, потому что пользователь не осознаёт, что общается с управляемой AI-системой, и злоумышленнику проще выдавать агента за человека в схемах обмана и социальной инженерии.

Ограничение: Только если скрытое AI-происхождение общения способствовало вовлечению человека в конкретную злонамеренную схему с наблюдаемым изменением поведения

AISAFE2-IMP-004 · Передача потребителем чувствительных данных

Отсутствие раскрытия факта взаимодействия с AI может привести к передаче потребителем чувствительных данных, потому что не зная, что собеседник — управляемый AI-агент, пользователь сообщает ему сведения, которые не раскрыл бы в распознанном сценарии социальной инженерии злоумышленника.

Ограничение: Только если сокрытие AI-статуса побудило пользователя передать секреты или персональные данные этому агенту без допустимого основания

Условия проявления

1. **Предпосылка.** Сгенерированный материал или реплики AI-системы внешне неотличимы от сообщений человека или проверенного материала сервиса.
2. **Триггер.** Получатель выбирает уровень доверия и дальнейшие действия, считая источником человека.
3. **Разрешающий пробел.** До этого момента не предъявляется явный сигнал о взаимодействии с AI.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	15	35
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	270	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	271	93
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	272	95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	273	97

Недостаток · AISAFE2-WEAK-007

Отсутствие границы доверия между инструкциями и данными

Краткое описание

Доверенные правила и недоверенный контент объединяются в одном контексте без сохраняемой типизации и происхождения.

Подробное описание

Недостаток проявляется, когда недоверенный текст, включая данные ограниченного доступа, может быть воспринят моделью как инструкция и повлиять на политику, порядок действий, выбор инструмента или дальнейшее раскрытие контекста.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-007 · Типизация доверия в контексте

Прямые причинные воздействия

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к вовлечению пользователя в злонамеренную схему, потому что инструкции злоумышленника, внедрённые в документы и ответы инструментов, исполняются моделью как

команды и позволяют управлять ответами системы, склоняющими пользователя к нужным атакующему действиям.

Ограничение: Только когда человек фактически вовлечён в конкретную злонамеренную схему и его наблюдаемое поведение изменилось; одного обманного сообщения недостаточно.

AISAFE2-IMP-004 · Передача потребителем чувствительных данных

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к передаче потребителем чувствительных данных, потому что внедрённые в контекст инструкции атакующего заставляют систему запрашивать у пользователя секреты или направлять его в указанный внешний канал под видом естественного диалога.

Ограничение: Только если внедрённая инструкция побудила пользователя передать чувствительные сведения модели, инструменту или через систему внешнему адресату без допустимого основания

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к неожидаемым для разработчика ответам модели, потому что любой текст, попавший в контекст, потенциально исполняется как инструкция, и недоверенное содержимое документов и ответов инструментов преодолевает заложенные ограничения, меняя поведение модели.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к обходу ограничений и предусмотренной логики агента, потому что инструкция, проведённая через обрабатываемый документ или ответ инструмента, исполняется моделью как команда и меняет порядок действий, выбор инструментов и цель агента без изменения кода.

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что недоверенный текст контекста меняет объём раскрытия его содержимого, и внедрённая инструкция заставляет модель выдать закрытые данные, полученные от инструментов.

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к раскрытию системного промпта и инструкций инструментов, потому что внедрённая в контекст инструкция меняет объём раскрытия его содержимого, и модель пересказывает системные инструкции по команде из недоверенного текста.

AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов

Внедрённая через недоверенное содержимое инструкция может удерживать агентную обработку одного запроса в повторяющихся шагах, если завершение зависит от решения модели и жёсткий предел шагов отсутствует.

Ограничение: Только для повторяющегося выполнения одного запроса или цепочки, которое не завершается по штатному условию.

Условия проявления

1. **Предпосылка.** От поведения модели зависят ответы, выбор действий и инструментов или раскрытие содержимого контекста.
2. **Триггер.** В контекст модели попадает недоверенное содержимое: пользовательский ввод, документы, веб-страницы, ответы инструментов или записи памяти.
3. **Разрешающий пробел.** Сборка контекста не сохраняет роль и происхождение фрагментов, поэтому модель может исполнить содержащийся в них текст как инструкцию.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

AML.T0051 · Диагностический признак сбоя

Смешение недоверенного текста с инструкциями создаёт условие, при котором внедрённый промпт может изменить ответ или действие модели.

Граница связи: Карточка описывает границу доверия, а не успешную атаку `AML.T0051` в конкретной системе.

Субъект	Аналитик техник атак на AI-систему
Условие применения	LLM получает пользовательский ввод либо извлекает внешнее содержимое, в котором могут находиться инструкции нарушителя.

Открыть положение в источнике

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI01 · Диагностический признак сбоя

Когда происхождение фрагментов контекста не сохраняется, агент может принять внешние данные за инструкцию и изменить цель или порядок действий.

Граница связи: Это один путь перехвата цели. Полномочия агента, контроль цели и последствия его действий проверяются отдельно.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент получает недоверенный ввод или сообщение другого агента, способное изменить его разрешённую цель.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM01:2026 · Диагностический признак сбоя

Смешение недоверенных данных с управляющими инструкциями даёт внедрённому тексту возможность изменить поведение приложения с языковой моделью.

Граница связи: Слабость описывает одну границу доверия; она не охватывает все способы внедрения и не доказывает, что модель выполнила атакующую инструкцию.

Субъект	Аналитик угроз AI-приложения
Условие применения	Недоверенное содержимое попадает в контекст модели и может пересечь границу между данными и инструкциями.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-003 · Управление контекстом общения с моделью	16	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	17	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	18	48
AISAFE2-COMP-007 · Долговременная память	19	50
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	274	95
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	275	97
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	276	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	277	101
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	278	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	279	107
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	280	126
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	578	324
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	579	561

Недостаток · AISAFE2-WEAK-008

Фундаментальный конфликт между возможностями и безопасностью

Краткое описание

Полезная способность модели выполнять широкий класс инструкций остаётся применимой и к нежелательным целям, не образуя детерминированной внутренней границы безопасности.

Подробное описание

Противоречие становится заметным в новых, пограничных или намеренно подобранных контекстах, для которых обученные ограничения не дают устойчивого разделения разрешённых и запрещённых целей.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений

Прямые причинные воздействия

AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к нарушению этических норм по отношению к пользователю, потому что внутри универсальной модели нет детерминированной границы между разрешённым и запрещённым, и без внешнего контроля результата она способна сама сформировать недопустимый ответ.

Ограничение: Только если нежелательное использование способности модели проявилось в ответе, решении или обращении с пользователем, нарушивших утверждённые для сценария этические нормы и причинивших ему вред

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к неожидаемым для разработчика ответам модели, потому что ограничения модели формируются обучением и не образуют строгого правила, и в пограничных или намеренно подобранных контекстах она выполняет недопустимую задачу или формирует вредный результат.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к обходу ограничений и предусмотренной логики агента, потому что ограничения агента удерживаются лишь вероятностным поведением модели и преодолеваются подобранным запросом, а недопустимое действие без внешнего детерминированного контроля результата исключить нельзя.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что нераскрытие полученных от инструментов сведений держится только на вероятностном поведении модели и преодолевается подобранным запросом.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к раскрытию системного промпта и инструкций инструментов, потому что запрет на их выдачу не является детерминированной границей возможностей модели и обходится специально построенным запросом.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Условия проявления

1. **Предпосылка.** Допустимость результата определяется только обученными ограничениями самой модели.
2. **Триггер.** Поступает пограничная или намеренно подобранная задача, следование которой даёт недопустимый результат или действие.
3. **Разрешающий пробел.** Внешний детерминированный контроль результата или действия отсутствует.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	20	37
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	281	98
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	282	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	283	101
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	284	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	285	107
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	580	429

Недостаток · AISAFE2-WEAK-009

Ненадёжное соблюдение системных инструкций

Краткое описание

Модель разрешает конфликт инструкций и ограничений вероятностно, поэтому ожидаемое старшинство директив не является инвариантом её поведения.

Подробное описание

Приоритет системных, пользовательских и полученных из данных директив зависит от формулировки и контекста; обученное следование правилам не заменяет детерминированный порядок и внешний механизм принуждения.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Ненадёжное соблюдение моделью системных инструкций может привести к неожидаемым для разработчика ответам модели, потому что старшинство системных директив не является инвариантом поведения, конфликт инструкций разрешается вероятностно, и ограничения преодолеваются как подобранными запросами, так и самопроизвольно.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Ненадёжное соблюдение моделью системных инструкций может привести к обходу ограничений и предусмотренной логики агента, потому что соблюдение системных директив вероятностно и зависит от всего контекста, и при конфликте инструкций агент отклоняется от предусмотренной логики, обходя заложенные правила.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Ненадёжное соблюдение моделью системных инструкций может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что указания о нераскрытии соблюдаются моделью вероятностно, а не как принуждение, и защищаемые сведения из контекста раскрываются вопреки системным директивам.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Ненадёжное соблюдение моделью системных инструкций может привести к раскрытию системного промпта и инструкций инструментов, потому что запрет на раскрытие — лишь одна из конкурирующих директив, соблюдаемая вероятностно, и подобранный запрос преодолевает его вопреки ожидаемому старшинству системного промпта.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Условия проявления

1. **Предпосылка.** Ожидаемые границы поведения, логика и защита сведений зависят от соблюдения моделью системных инструкций.
2. **Триггер.** В запросе или контексте возникает случайный или целенаправленно созданный конфликт директив.
3. **Разрешающий пробел.** Механизм принуждения к старшинству системных инструкций не даёт гарантий, и исход определяется вероятностным поведением модели.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

> **AML.T0054** · Диагностический признак сбоя
>
> Ненадёжное соблюдение старшинства инструкций создаёт одну из предпосылок обхода ограничений модели через специально составленный запрос.
>
> **Граница связи:** Техника охватывает также изменение весов и защитных механизмов; карта не доказывает успешный обход.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик техник атак на AI-систему |
> | **Условие применения** | Поведенческие ограничения модели могут испытываться с помощью специально составленных запросов. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	21	37
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	286	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	287	101

Карточка	Связи	Стр.
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	288	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	289	107
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	581	429

Недостаток · AISAFE2-WEAK-010

Размытие обученных ограничений через контекст

Краткое описание

Накопленные в окне контекста примеры, назначенная роль и сложившийся шаблон беседы повышают вероятность продолжения этого шаблона вопреки обученным ограничениям.

Подробное описание

Недостаток проявляется в длинных, многошаговых или намеренно подобранных диалогах, где накопленные примеры и локальный шаблон поведения снижают относительное влияние ранее обученного ограничения. Весы модели при этом не меняются: меняется относительное влияние накопленного контекстного шаблона на вероятностный выбор продолжения.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений

Прямые причинные воздействия

AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы

Преодоление обученных ограничений содержимым контекста может привести к нарушению этических норм по отношению к пользователю, потому что накопленные в контексте примеры, назначенная роль и сложившийся шаблон беседы повышают вероятность продолжения недопустимого шаблона, и обученные ограничения преодолеваются в результате диалога.

Ограничение: Только если контекстный обход обученного ограничения привёл к ответу, решению или обращению с пользователем, нарушившим утверждённые для сценария этические нормы и причинившим ему вред

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Преодоление обученных ограничений содержимым контекста может привести к неожидаемым для разработчика ответам модели, потому что содержимое окна контекста меняет поведение модели без изменения весов, и подобранным запросом снимаются ограничения на запрещённые темы.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Преодоление обученных ограничений содержимым контекста может привести к обходу ограничений и предусмотренной логики агента, потому что постепенно выстроенный контекст переопределяет фактическое поведение агента, и обученные ограничения перестают удерживать сценарий его действий.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Преодоление обученных ограничений содержимым контекста может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что сложившийся в контексте шаблон согласия повышает вероятность выдачи полученных от инструментов защищаемых сведений вопреки обученному запрету.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Преодоление обученных ограничений содержимым контекста может привести к раскрытию системного промпта и инструкций инструментов, потому что накопленный контекст постепенно подводит модель к пересказу защищённых инструкций, в котором при прямом запросе она отказала бы.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Условия проявления

1. **Предпосылка.** Запрещённое целевое поведение остаётся в пределах способностей модели, несмотря на обученные ограничения.
2. **Триггер.** Субъект ведёт длинный или многошаговый диалог и наполняет окно контекста примерами, ролью и шаблоном, согласованными с нежелательной целью.
3. **Разрешающий пробел.** Система не отклоняет и не сокращает накопленный контекст безопасным образом перед дальнейшей обработкой.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM01:2026 · Диагностический признак сбоя

Накопленный шаблон беседы может усилить влияние внедрённых инструкций и ослабить ранее обученный запрет модели.

Граница связи: Это вероятностный эффект контекста, а не изменение весов; нужна конкретная атакующая инструкция и её наблюдаемый эффект.

Субъект	Аналитик угроз AI-приложения
Условие применения	Недоверенное содержимое попадает в контекст модели и может пересечь границу между данными и инструкциями.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	22	37
AISAFE2-COMP-003 · Управление контекстом общения с моделью	23	40
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	290	98
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	291	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	292	101
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	293	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	294	107
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	582	429

Недостаток · AISAFE2-WEAK-011

Поверхностная настройка модели на безопасное поведение

Краткое описание

Безопасное поведение модели зависит от поверхностных признаков запроса и теряет устойчивость при переформулировке или смене условий.

Подробное описание

Она становится релевантной при состязательных запросах, нестандартных языках, форматах или распределениях данных, где поверхностные признаки перестают удерживать ожидаемое поведение. Данные настройки безопасного поведения могут покрывать более узкое распределение, чем базовое обучение, поэтому способность выполнить задачу сохраняется в редком языке, кодировке или формате, где ограничение не обобщилось.

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений

Прямые причинные воздействия

AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы

Слабая настройка модели на безопасное поведение может привести к нарушению этических норм по отношению к пользователю, потому что безопасное поведение обобщается хуже способностей модели, и в необычных формулировках, языках и форматах ограничитель не срабатывает, пропуская недопустимое содержание пользователю.

Ограничение: Только если сбой настройки безопасного поведения при смене языка, формата или условий привёл к ответу, решению или обращению с пользователем, нарушившим утверждённые для сценария этические нормы и причинившим ему вред

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Слабая настройка модели на безопасное поведение может привести к неожиданным для разработчика ответам модели, потому что данные дообучения безопасному поведению покрывают более узкое распределение, чем способности модели, и часть обходов ограничений (редкие языки, кодировки, форматы) обнаруживается уже после выпуска.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Слабая настройка модели на безопасное поведение может привести к обходу ограничений и предусмотренной логики агента, потому что смена формы запроса (язык, кодировка, формат) выводит задачу из покрытого настройкой распределения, и заложенные ограничения поведения агента не срабатывают.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Слабая настройка модели на безопасное поведение может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что ограничитель, обученный на узком распределении, не срабатывает на просьбах пересказать, перевести или закодировать сведения, и защита нераскрытия обходится сменой формы запроса.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Слабая настройка модели на безопасное поведение может привести к раскрытию системного промпта и инструкций инструментов, потому что запрет на их выдачу, выученный на типовых формулировках, обходится нетипичной формой запроса — переводом, кодированием, косвенным пересказом.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Условия проявления

1. **Предпосылка.** Способность модели выполнить запрещённую задачу в слабо представленном формате сохраняется.
2. **Триггер.** Запрещённая задача формулируется на редком языке, в кодировке, непривычном формате или иной слабо представленной при настройке безопасности форме.
3. **Разрешающий пробел.** На пути к получателю нет иных контролей, задерживающих недопустимый результат.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	24	37
AISAFE2-COMP-009 · Конвейер обучения и дообучения	25	54
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	295	98
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	296	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	297	101
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	298	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	299	107
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	583	429

Недостаток · AISAFE2-WEAK-012

Непроверяемая или скомпрометированная цепочка поставки AI

Краткое описание

Точный состав, происхождение или целостность компонентов AI-системы в выпуске нельзя подтвердить до их использования.

Подробное описание

Недостаток активируется при добавлении, динамической регистрации, обновлении или продвижении стороннего либо внутреннего компонента, если до использования не подтверждены его точная версия, состав, происхождение или целостность. Для внешнего компонента решение о допуске поставщика проверяется отдельно по утверждённой политике.

Область применения

- Общая безопасность

Происхождение

- Процесс
- Реализация

Классы недостатка

- AISAFE2-WC-032 · Цепочка поставки, недоверенная среда и физическая защита

Прямые причинные воздействия

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к вовлечению пользователя в злонамеренную схему, потому что подменённый или заражённый компонент цепочки поставки действует изнутри доверенного контура и позволяет злоумышленнику управлять ответами и поведением системы.

Ограничение: Только если скомпрометированный компонент сформировал или передал убеждающее сообщение и человек изменил поведение в интересах злонамеренной схемы

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к неожидаемым для разработчика ответам модели, потому что компоненты с заранее встроенным поведением — заражённые модели, отравленные наборы данных — включаются без проверки происхождения и целостности и демонстрируют неожиданные свойства изнутри доверенного контура.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к обходу ограничений и предусмотренной логики агента, потому что скомпрометированный компонент, встроенный в контур агента, меняет его решения и действия изнутри в обход задуманной логики.

Ограничение: Только если скомпрометированный AI-компонент фактически изменил план, действие или состояние Agent Core; его присутствия в цепочке поставки недостаточно

AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к неожидаемому поведению слоя инструментов, потому что инструменты и их описания подключаются и обновляются без проверки происхождения и целостности (включая динамическую регистрацию через MCP), и в составе доступных появляется незаявленный или заражённый компонент.

Ограничение: Только когда инструмент, RAG или сторонний агент выдаёт непредусмотренный результат либо выполняет действие, достигшее следующего компонента или внешней системы.

AISAFE2-IMP-011 · Компрометация баз знаний

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к компрометации базы знаний, потому что скомпрометированный компонент цепочки поставки действует изнутри доверенного контура и может передавать содержимое базы знаний поставщику и третьим сторонам.

Ограничение: Только если скомпрометированный компонент передал неуполномоченному получателю существенную часть базы знаний, её метаданных или индекса

AISAFE2-IMP-012 · Кража модели

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к краже модели, потому что внедрённый в цепочку поставки код действует изнутри контура с доступом к весам и артефактам модели и передаёт их атакующему.

Ограничение: Только если внедрённый через цепочку поставки компонент фактически передал веса или связанные артефакты субъекту без разрешения

AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к раскрытию сведений о технологиях, алгоритмах и источниках данных системы, потому что скомпрометированный компонент (например, инструмент загрузки наборов данных) сообщает сведения о системе и её данных наружу — поставщику или третьим сторонам.

Ограничение: Только при фактической передаче наружу сведений об архитектуре, технологиях, алгоритмах, конфигурации или источниках данных; внутреннего доступа недостаточно.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к несанкционированному чтению или изменению хранимых данных, потому что заражённый компонент исполняется с правами доверенной системы и читает или изменяет достижимые ей хранимые данные в обход контроля полномочий.

Ограничение: Только если заражённый компонент с правами доверенной системы фактически прочитал или изменил хранимый объект вне полномочий атакующего

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных

компонентов (в том числе за счёт повышения привилегий)

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что закладка в компоненте вызывает функции инфраструктуры от имени системы, на которые инициатор воздействия полномочий не имеет.

Ограничение: Только при успешном вызове инфраструктурной функции за пределами полномочий субъекта или с полномочиями скомпрометированного компонента.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к утечке чувствительных данных в недоверенный контур, потому что скомпрометированный или недоверенный компонент передаёт обрабатываемые данные поставщику и третьим сторонам изнутри доверенного контура.

Ограничение: Только при фактическом пересечении чувствительными данными границы доверия без разрешённого перехода к точному получателю или среде.

Условия проявления

1. **Предпосылка.** Добавляемый компонент способен проявить неожиданное или вредоносное поведение внутри доверенного контура с доступом к данным и функциям системы.
2. **Триггер.** В систему добавляется или обновляется модель, набор данных, библиотека, инструмент или его описание.
3. **Разрешающий пробел.** До начала использования отсутствует хотя бы одно необходимое подтверждение: точной версии и целостности компонента по доверенному отпечатку, подписи или эквивалентной проверке; состава по проверяемому манифесту, связанному с этой версией, или эквивалентной проверке; происхождения по аттестации или другому проверяемому свидетельству; либо требуемого политикой допуска внешнего поставщика, если он участвует в поставке.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

AML.T0010 · Диагностический признак сбоя

Непроверенное происхождение или целостность компонента оставляет путь для подмены цепочки поставки AI, описанный `AML.T0010`.

Граница связи: Карточка описывает условие риска; факт подмены или атаки из неё не следует.

Субъект	Аналитик техник атак на AI-систему
Условие применения	AI-система принимает модель, данные, программный компонент или оборудование, которые могут быть скомпрометированы в цепочке поставки.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM04:2026 · Диагностический признак сбоя

Если до использования нельзя подтвердить происхождение, состав или целостность точной версии компонента, цепочка поставки становится уязвимой к подмене.

Граница связи: Непроверяемое происхождение не доказывает компрометацию компонента и не описывает все риски поставщика или лицензирования.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	26	44
AISAFE2-COMP-008 · Данные для подсистемы обучения	27	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	28	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	29	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	30	67
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	31	84
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	300	95
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	301	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	302	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	303	104
AISAFE2-IMP-011 · Компрометация баз знаний	304	109
AISAFE2-IMP-012 · Кража модели	305	111
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	306	114
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	307	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	308	118
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	309	130
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	584	359
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	585	364

Недостаток · AISAFE2-WEAK-013

Отсутствие прослеживаемости происхождения и преобразований данных

Краткое описание

Для данных нельзя восстановить подтверждённые сведения об источнике, версии и значимых преобразованиях, с которыми они попали в AI-контур.

Подробное описание

Отсутствующая или неполная трасса «источник → преобразования → итоговый набор» повышает стоимость и снижает доказательность воспроизведения релиза, расследования изменений поведения модели и аудита свойств данных, а также затрудняет обнаружение отравления.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-016 · Управление и происхождение данных

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Набор без восстановимой трассы источника и преобразований проходит допуск до проверки переходов; если в нём присутствует отравлённый материал, отсутствие трассы позволяет использовать точную изменённую версию в обучении, после чего модель формирует и передаёт неожиданный ответ.

Ограничение: Только когда конкретная отравлённая версия проходит допуск именно из-за отсутствия восстановимой трассы, используется в обучении и приводит к фактически переданному неожиданному ответу; отсутствие такой трассируемости без этого пути недостаточно.

Условия проявления

1. **Предпосылка.** Данные используются как основание выпуска, аудита или расследования.
2. **Триггер.** Данные проходят приём, преобразования и смену версий перед использованием в обучении, базе знаний или артефакте.
3. **Разрешающий пробел.** Сведения об источнике, версиях и преобразованиях не фиксируются либо теряются по пути.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

AML.T0020 · Диагностический признак сбоя

Разрыв трассы происхождения и преобразований затрудняет выявление изменений обучающих данных, которые использует техника `AML.T0020`.

Граница связи: Карточка не проверяет содержимое набора и не означает, что данные были отравлены.

Субъект	Аналитик техник атак на AI-систему
Условие применения	Модель обучается или дообучается на данных, разметке либо обратной связи, состав которых может изменить нарушитель.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM05:2026 · Диагностический признак сбоя

Отсутствие трассы источника, версии и преобразований затрудняет поиск, расследование и откат набора данных, если в нём обнаружено отравление.

Граница связи: Отсутствие трассы само по себе не означает, что данные отравлены или что вредная запись попала в систему.

Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-008 · Данные для подсистемы обучения	32	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	33	56
AISAFE2-COMP-023 · Подсистема формирования базы знаний	34	86
AISAFE2-COMP-025 · Источники данных для инструментов	35	90
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	310	100
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	586	451
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	587	469
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	588	541

Недостаток · AISAFE2-WEAK-014

Неконтролируемый допуск материала в источники знаний и контуры обратной связи

Краткое описание

Непроверенный внешний или сгенерированный материал допускается в источник знаний, память либо повторный цикл обработки.

Подробное описание

Ошибки, галлюцинации и намеренно внедрённые атакующим данные сохраняются и воспроизводятся при последующих обращениях: базы знаний отравляются, поведение системы постепенно деградирует, а через содержимое источников становится возможным управлять её ответами и действиями.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-017 · Контроль допуска источников

Прямые причинные воздействия

AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы

Неконтролируемый допуск материала в источники знаний и контуры обратной связи может привести к вовлечению пользователя в злонамеренную схему, потому что непроверенный материал, внедрённый атакующим в базы знаний и контуры обратной связи, сохраняется и воспроизводится, позволяя управлять ответами и действиями системы в его интересах.

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Если ошибки или вредоносные примеры без контроля попадают в обучающие данные либо контур обратной связи, последующее обучение может изменить поведение модели и вызвать неожиданные ответы на рабочих запросах.

Ограничение: Загрязнение RAG-корпуса без изменения модели относится к ответу системы и рассматривается отдельно; здесь требуется фактический путь в обучение.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Неконтролируемый допуск материала в источники знаний и контуры обратной связи может привести к обходу ограничений и предусмотренной логики агента, потому что внедрённый в базы знаний и долговременную память материал воспроизводится при последующих обращениях и позволяет управлять планами и действиями агента через содержимое источников.

AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)

Неконтролируемый допуск материала в источники знаний и контуры обратной связи может привести к неожидаемому поведению слоя инструментов, потому что база знаний пополняется из источников, доступных злоумышленнику, без проверки происхождения и качества материала, и RAG возвращает подложные или отравленные данные.

Условия проявления

1. **Предпосылка.** Последующая обработка использует сохранённое содержимое как доверенное.
2. **Триггер.** Внешний или сгенерированный системой материал записывается в базу знаний, долговременную память или обучающие данные.
3. **Разрешающий пробел.** При записи не проверяются происхождение, допустимость, качество и срок жизни материала.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

AML.T0070 · Диагностический признак сбоя

Допуск непроверенного материала в базу знаний создаёт путь для отравления RAG-корпуса и последующих ответов.

Граница связи: Техника относится к RAG; карточка шире и сама не доказывает внесение вредоносного материала.

Субъект	Аналитик техник атак на AI-систему
Условие применения	Приложение извлекает контекст из RAG-корпуса, в который нарушитель может поместить документы или записи.

Открыть положение в источнике

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI06 · Диагностический признак сбоя

Неконтролируемый допуск материала в память или базу знаний создаёт путь для отравления последующего контекста агента.

Граница связи: Нужны вредный материал и его повторное использование; один лишь слабый допуск не доказывает произошедшее отравление.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.
Условие связи с карточкой	Непроверенный материал попадает в постоянную память, источник знаний или обратную связь и затем повторно используется.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-007 · Долговременная память	36	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	37	52
AISAFE2-COMP-023 · Подсистема формирования базы знаний	38	86
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	311	95
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	312	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	313	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	314	104
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	589	339
AISAFE2-MIT-021 · Управление доверием и сроком жизни долговременной памяти	590	396
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	591	469
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	592	541

Недостаток · AISAFE2-WEAK-015

Переобучение модели

Краткое описание

Модель воспроизводит особенности обучающей или оценочной выборки вместо устойчивого поведения на целевых данных.

Подробное описание

К переобучению приводят избыточная ёмкость относительно данных, слабая регуляризация, чрезмерное обучение, точные и приблизительные дубликаты, нерепрезентативная выборка, а также многократный отбор по одному оценочному набору или утечка тестовых данных.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-002 · Поведение при обучении и дообучении

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Переобученность модели может привести к неожиданным для разработчика ответам модели, потому что модель воспроизводит частные особенности обучающей выборки вместо устойчивых закономерностей, и на новых и редких входах качество систематически ниже измеренного при тестировании.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Наблюдаемая зависимость переобученной модели от знакомой выборки позволяет воспроизводимо определить участие чувствительной записи либо восстановить её атрибут посредством атаки определения принадлежности или проверки восстановления.

Ограничение: Только при подтверждённом воспроизведении, определении принадлежности, логическом выводе или восстановлении конкретной чувствительной записи либо её атрибута.

Условия проявления

1. **Предпосылка.** Выборка и процедура обучения поощряют качество на знакомых данных из-за дубликатов, нерепрезентативности или подгонки под оценочный набор.
2. **Триггер.** Рабочие входы отличаются от обучающих либо внешний субъект целенаправленно исследует разрыв поведения на знакомых и новых примерах.
3. **Разрешающий пробел.** Допуск модели опирается на метрики по тем же данным без отдельной проверки переноса на новые данные.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	39	37
AISAFE2-COMP-008 · Данные для подсистемы обучения	40	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	41	54
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	315	100
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	316	113
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	593	466

Недостаток · AISAFE2-WEAK-016

Слабый контроль доступа к обучающим данным, конвейерам и артефактам

Краткое описание

Машинное задание обучения или обработки артефакта получает более широкие полномочия, чем нужны для конкретного запуска и его входов или выходов.

Подробное описание

Недостаток реализован в техническом применении прав: сервисная идентичность задания может читать, изменять или экспортировать обучающие данные и артефакты за пределами назначенного пути. Риск усиливается, когда одна сервисная идентичность используется

несколькими стадиями и внешними подсистемами и получает доступ сразу к нескольким наборам данных и артефактам.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-010 · Контроль доступа и авторизация

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Избыточные права машинного задания позволяют фактически изменить обучающие данные или артефакт вне назначенного пути; изменённая версия проходит обучение и выпуск и формирует неожиданный ответ.

Ограничение: Только когда точная версия модели формирует и передаёт ответ вне утверждённых границ содержания, формата, качества или поведения; одной возможности недостаточно.

AISAFE2-IMP-012 · Кража модели

Слабый контроль доступа к обучающим данным, конвейерам и артефактам может привести к краже модели, потому что компрометация одного задания конвейера с избыточными сервисными полномочиями открывает хищение всех достижимых артефактов модели.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Слабый контроль доступа к обучающим данным, конвейерам и артефактам может привести к несанкционированному чтению или изменению хранимых данных, потому что сервисные учётные записи конвейеров имеют избыточные полномочия, и компрометация одного задания позволяет читать и подменять достижимые наборы и модели, включая отравление будущих версий.

AISAFE2-IMP-031 · Утрата датасета в результате атаки

Недостаток позволяет атакующему изменить или удалить датасет обучения, тестирования либо валидации вместе с доступными доверенными копиями, вследствие чего требуемая версия датасета становится невосстановимой.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения доступных доверенных копий.

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Недостаток позволяет атакующему изменить или удалить веса, базу знаний, индекс либо критичную конфигурацию вместе с доступными доверенными версиями, вследствие чего требуемый AI-артефакт становится невосстановимым.

Ограничение: Только для вызванной атакой невосстановимой утраты критичного AI-артефакта, кроме датасета, когда отсутствует доступная доверенная версия или воспроизводимый рецепт.

Условия проявления

1. **Предпосылка.** Задание обучения или обработки артефактов выполняется от сервисной учётной записи, права которой определяют достижимую область воздействия.
2. **Триггер.** Задание, его код или учётные данные компрометируются либо используются со злоупотреблением.
3. **Разрешающий пробел.** Права сервисной учётной записи шире минимально необходимых для задания.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM05:2026** · Диагностический признак сбоя
>
> Избыточные права задания обучения позволяют несанкционированно изменить данные или артефакт и тем самым создать условия для отравления будущей версии модели.
>
> **Граница связи:** Чрезмерные права не доказывают изменения данных; загрязнение источника без доступа к заданию остаётся отдельным путём.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-008 · Данные для подсистемы обучения	42	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	43	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	44	56
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	45	61
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	317	100
AISAFE2-IMP-012 · Кража модели	318	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	319	116
AISAFE2-IMP-031 · Утрата датасета в результате атаки	320	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	321	141
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	594	383
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	595	473

Карточка	Связи	Стр.
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	596	481

Недостаток · AISAFE2-WEAK-017

Отсутствие проверяемых критериев выбора модели

Краткое описание

Процесс выбора модели или конфигурации не задаёт проверяемые критерии пригодности и правила сравнения кандидатов для целевого сценария.

Подробное описание

Пробел возникает до оценки и допуска: команда не определяет требуемые свойства, пороги и граничные условия, не фиксирует сопоставимые результаты кандидатов и основание выбора конкретной модели или конфигурации.

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-005 · Выбор и допуск модели

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Модель выбирается и допускается без проверяемых критериев содержания, формата и качества; выбранная точная версия затем фактически выдаёт результат вне предусмотренной границы.

Ограничение: Только когда точная версия модели формирует и передаёт ответ вне утверждённых границ содержания, формата, качества или поведения; одной возможности недостаточно.

Условия проявления

1. **Предпосылка.** Для целевого сценария выбирается модель или её конфигурация.
2. **Триггер.** Результат выбора допускается в эксплуатацию без независимой проверки пригодности.
3. **Разрешающий пробел.** Требуемые свойства, пороги и граничные условия заранее не заданы, а сравнение кандидатов и основание решения не фиксируются.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

> **MANAGE 1.1** · Диагностический признак сбоя
>
> Карточка показывает, что без проверяемых критериев выбора модели труднее обосновать решение о её пригодности для заявленной цели.
>
> **Граница связи:** Критерии выбора не заменяют оценку всей AI-системы и отдельное решение о продолжении разработки или выпуска.

Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	Владелец решает, достигает ли конкретная AI-система заявленной цели и допустим ли её выпуск, используя добровольный NIST AI RMF.
Условие связи с карточкой	Назначение AI-системы и выбор кандидата влияют на решение о разработке либо внедрении.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	46	67
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	322	100
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	597	466

Недостаток · AISAFE2-WEAK-018

Избыточная наблюдаемость и воспроизводимость поведения модели через интерфейс

Краткое описание

Повторяемый доступ к штатным ответам модели даёт внешнему субъекту устойчивый сигнал соответствия входов и выходов, пригодный для систематического исследования её поведения.

Подробное описание

Через разрешённый интерфейс субъект может выполнять адаптивные серии запросов и сопоставлять ответы. Если ответы сохраняют устойчивые признаки версии, настроек, защитных механизмов или подключённых источников, накопленный сигнал позволяет исследовать границы поведения и строить функционально близкий аналог без доступа к параметрам модели. Само наличие повторяемых ответов не является недостатком; значимость определяется назначением интерфейса, информативностью сигнала и отсутствием ограничений на систематическое исследование.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-004 · Сигналы вывода и извлечения модели

Прямые причинные воздействия

AISAFE2-IMP-012 · Кража модели

Наблюдаемость и воспроизводимость поведения модели для внешнего субъекта могут привести к краже модели, потому что штатный доступ к ответам раскрывает функциональную зависимость выхода от входа, и накопленные пары запрос — ответ позволяют обучить функциональную копию модели без доступа к весам.

Ограничение: Только если собранные через интерфейс пары запрос—ответ позволили подтвердить функциональное воспроизведение ценного поведения модели без разрешения

AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных

Наблюдаемость и воспроизводимость поведения модели для внешнего субъекта могут привести к раскрытию сведений о технологиях, алгоритмах и источниках данных системы, потому что воспроизводимый сигнал вход — выход позволяет адаптивными сериями запросов определить семейство и версию модели, настройки, защитные механизмы и подключённые источники данных.

Ограничение: Только при подтверждённом неразрешённом получении закрытой информации о технологиях, архитектуре, алгоритмах или источниках данных.

Условия проявления

1. **Предпосылка.** Штатные ответы детальны, воспроизводимы и сохраняют устойчивые признаки версии, настроек и подключённых источников.
2. **Триггер.** Внешний субъект выполняет многочисленные адаптивные серии запросов к модели.
3. **Разрешающий пробел.** Систематическое исследование не ограничивается по субъекту и объёму обращений, а такая детализация ответа не требуется назначением сервиса.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

> **AML.T0024.002** · Диагностический признак сбоя
>
> Многократные подробные ответы API дают сигнал вход–выход, по которому можно строить функционально близкий аналог модели.
>
> **Граница связи:** Повторяемость ответов сама по себе допустима. Значимость слабости зависит от информативности ответов и доступного объёма запросов; успешное копирование подтверждается отдельно.
>
Субъект	Аналитик техник атак на AI-систему
Условие применения	Защищаемая модель доступна через API инференса, позволяющий многократно получать ответы для обучения её функциональной копии.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	47, 48	37
AISAFE2-IMP-012 · Кража модели	323	111
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	324	114
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	598, 599	485

Недостаток · AISAFE2-WEAK-019

Избыточное раскрытие внутренних оценок модели

Краткое описание

Интерфейс возвращает недоверенному потребителю внутренние оценки модели — векторы вероятностей, логиты, логарифмы правдоподобия или детальную уверенность — сверх необходимого результата.

Подробное описание

Контракт обычного или диагностического API не минимизирует внутренние оценки по размерности, разрядности либо доступному режиму: клиент получает весь вектор вместо необходимого ограниченного набора наиболее значимых значений, близкие к сырым оценки или расширенную диагностику. Серия таких ответов даёт дополнительный сигнал о границах решения, облегчает построение замещающей модели и может повысить точность проверки участия примера в обучении. Ограничение выдачи снижает информативность, но не исключает исследование по обычным ответам.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-004 · Сигналы вывода и извлечения модели

Прямые причинные воздействия

AISAFE2-IMP-012 · Кража модели

Возврат логитов или распределения вероятностей раскрывает более подробный сигнал о поведении модели и может облегчить построение функционального аналога при массовых запросах.

Ограничение: Только если раскрытые оценки модели помогли подтвердить функциональное воспроизведение её ценного поведения без разрешения

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Избыточное раскрытие оценок уверенности в ответе модели может привести к раскрытию чувствительных данных из обучающего датасета, потому что серия детальных оценок уверенности раскрывает различимость отдельных примеров и повышает точность вывода о принадлежности предъявленного примера обучающей выборке.

Ограничение: Только если раскрытые оценки позволили подтвердить участие конкретной чувствительной записи в обучении либо восстановить её атрибут; предположение по сигналу недостаточно

Условия проявления

1. **Предпосылка.** Недоверенный потребитель получает внутренние оценки модели сверх необходимого результата: полные векторы вероятностей, высокую разрядность или диагностические режимы.
2. **Триггер.** Потребитель выполняет массовые или адаптивные запросы.
3. **Разрешающий пробел.** Ответ не сокращается перед выдачей потребителю.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Диагностический признак сбоя
>
> Подробные оценки модели в ответах дают дополнительный сигнал для вывода об участии записи или её свойствах без выдачи самой записи.
>
> **Граница связи:** Внутренние оценки не всегда раскрывают чувствительный факт; это зависит от актива, точности сигнала и доступа к запросам.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	49, 50, 51	37
AISAFE2-IMP-012 · Кража модели	325	111
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	326	113
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	600, 601, 602	485

Недостаток · AISAFE2-WEAK-022

Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации

Краткое описание

Правила доступа задаются раздельно по стадиям и средам без сквозной модели субъектов, объектов и операций, поэтому на стыках возникают участки с неопределёнными или избыточными правами и транзитивный доступ между контурами.

Подробное описание

Транзитивный доступ между стадиями возникает через общие технические учётные записи, разделяемые хранилища и конвейеры CI/CD, даже если локальные правила каждой стадии выглядят ограниченными.

Область применения

- Общая безопасность

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-010 · Контроль доступа и авторизация

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к компрометации базы знаний, потому что несогласованные по стадиям и средам правила доступа создают участки с избыточными правами и транзитивный доступ к данным и их конвейерам между контурами.

AISAFE2-IMP-012 · Кража модели

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к краже модели, потому что транзитивный доступ между контурами и избыточные права на стыках стадий жизненного цикла позволяют считывать артефакты модели вне полномочий.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к несанкционированному чтению или изменению хранимых данных, потому что на стыках стадий и сред возникают участки с неопределёнными и избыточными правами и транзитивный доступ, допускающие чтение, изменение и удаление данных вне полномочий.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что транзитивный доступ между контурами позволяет субъекту одной стадии вызывать функции компонентов другой, на которые у него нет полномочий.

Ограничение: Только если транзитивное право одной стадии действительно позволило субъекту вызвать функцию другой стадии, на которую у него нет полномочий

AISAFE2-IMP-031 · Утрата датасета в результате атаки

Недостаток позволяет атакующему изменить или удалить датасет обучения, тестирования либо валидации вместе с доступными доверенными копиями, вследствие чего требуемая версия датасета становится невосстановимой.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения доступных доверенных копий.

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Недостаток позволяет атакующему изменить или удалить веса, базу знаний, индекс либо критичную конфигурацию вместе с доступными доверенными версиями, вследствие чего требуемый AI-артефакт становится невосстановимым.

Ограничение: Только для вызванной атакой невосстановимой утраты критичного AI-артефакта, кроме датасета, когда отсутствует доступная доверенная версия или воспроизводимый рецепт.

Условия проявления

1. **Предпосылка.** Данные или артефакты переходят между стадиями и средами жизненного цикла: подготовкой данных, обучением, публикацией, эксплуатацией и распространением.
2. **Триггер.** Субъект использует избыточные права на стыке стадий или транзитивный доступ для операции вне своих полномочий.
3. **Разрешающий пробел.** Для переходов не заданы сквозные правила о субъектах и допустимых операциях, а общие учётные записи, разделяемые хранилища и конвейеры размывают границу доступа.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	52	61

Карточка	Связи	Стр.
AISAFE2-IMP-011 · Компрометация баз знаний	327	109
AISAFE2-IMP-012 · Кража модели	328	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	329	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	330	118
AISAFE2-IMP-031 · Утрата датасета в результате атаки	331	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	332	141
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	603	342
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	604	383
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	605	473
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	606	481
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	607	541
AISAFE2-MIT-061 · Непубличность хранилищ AI-системы по умолчанию	608	544

Недостаток · AISAFE2-WEAK-023

Незашифрованное хранение

Краткое описание

Постоянные данные или артефакты доступны в хранилище без криптографической защиты содержимого.

Подробное описание

Незащищёнными остаются также резервные, временные, экспортные копии и снимки: шифрование только основной копии оставляет содержимое доступным через альтернативное место хранения.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-027 · Криптографическая защита хранения и каналов

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Незащищенное хранение данных и артефактов может привести к компрометации базы знаний, потому что база знаний и её индекс хранятся без криптографической защиты содержимого, и субъект с доступом к месту хранения считывает корпус напрямую и целиком.

Ограничение: Только если субъект без полномочий прочитал из незащищённого хранилища существенную часть корпуса или индекса; возможность доступа недостаточна

AISAFE2-IMP-012 · Кража модели

Незащищенное хранение данных и артефактов может привести к краже модели, потому что веса и контрольные точки хранятся без криптографической защиты, и артефакты модели копируются целиком из незашифрованного хранилища.

Ограничение: Только если субъект без разрешения скопировал веса или связанные артефакты из незащищённого хранилища

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Незащищенное хранение данных и артефактов может привести к несанкционированному чтению или изменению хранимых данных, потому что без криптографической защиты хранения единственным барьером остаётся логическое разграничение, и субъект с доступом к месту хранения читает содержимое напрямую.

Ограничение: Только если субъект без полномочий фактически прочитал хранимый объект через незащищённое хранилище; возможность доступа недостаточна

Условия проявления

1. **Предпосылка.** Данные или артефакты, включая резервные, временные и экспортные копии, хранятся без шифрования содержимого.
2. **Триггер.** Субъект получает доступ к носителю, снимку, копии или административному интерфейсу хранилища.
3. **Разрешающий пробел.** Между субъектом и содержимым нет иных барьеров, кроме логического разграничения доступа.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Диагностическая связь

Незашифрованная рабочая, резервная или временная копия указывает на нарушение требуемой защиты при хранении.

Ограничение: Связь диагностическая: она не подтверждает выполнение требования. Защищённая передача и управление ключами проверяются отдельно.

Условие применения: При хранении данных, модельных артефактов, индексов, дисков или резервных копий без требуемого шифрования.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-007 · Долговременная память	53	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	54	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	55	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	56	58
AISAFE2-COMP-013 · Телеметрия и обнаружение	57	64
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	58	75
AISAFE2-COMP-023 · Подсистема формирования базы знаний	59	86
AISAFE2-IMP-011 · Компрометация баз знаний	333	109
AISAFE2-IMP-012 · Кража модели	334	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	335	116
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	609	356
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	610	481
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	611	547

Недостаток · AISAFE2-WEAK-024

Незашифрованные каналы передачи

Краткое описание

Данные или артефакты передаются между компонентами без защищённого транспортного шифрования.

Подробное описание

Недостаток проявляется, когда субъект с доступом к маршруту или промежуточному узлу может наблюдать передаваемое содержимое или подменять его до конечной точки. Разрыв защиты на любом маршруте жизненного цикла, включая поставку артефакта за пределы контролируемой зоны, оставляет содержимое наблюдаемым или изменяемым на промежуточном узле.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-027 · Криптографическая защита хранения и каналов

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Незащищенные каналы передачи данных могут привести к компрометации базы знаний, потому что данные передаются между компонентами без шифрования транспорта и проверки подлинности сторон, и содержимое базы перехватывается в канале без наблюдаемых следов.

Ограничение: Только если через незащищённый канал прошёл и был перехвачен объём, позволяющий получить существенную часть базы знаний, её метаданных или индекса

AISAFE2-IMP-012 · Кража модели

Незащищенные каналы передачи данных могут привести к краже модели, потому что артефакты передаются по каналам без транспортного шифрования, и веса модели перехватываются при передаче без наблюдаемых следов, включая этап поставки.

Ограничение: Только если веса или связанные артефакты были фактически перехвачены в незащищённом канале субъектом без разрешения

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Атакующий изменяет данные или артефакт в незащищённом канале, а конечная точка принимает и сохраняет подменённое содержимое, что реализует несанкционированное изменение хранимого объекта.

Ограничение: Только когда атакующий изменяет данные или артефакт в незащищённом канале, конечная точка принимает подмену и сохраняет её как хранимый объект вне полномочий атакующего; пассивное чтение трафика само по себе не образует AISAFE2-IMP-015.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Атакующий подменяет или повторно воспроизводит запрос в незащищённом канале, а конечная точка исполняет инфраструктурную функцию вне полномочий атакующего.

Ограничение: Только если подменённый или повторный запрос был принят конечной точкой и вызвал функцию вне полномочий атакующего

Условия проявления

1. **Предпосылка.** Данные или артефакты передаются по маршруту без шифрования транспорта и проверки подлинности сторон.
2. **Триггер.** Субъект использует доступ к каналу или промежуточному узлу для перехвата либо подмены передачи.
3. **Разрешающий пробел.** Конечная точка не имеет средств обнаружить перехват или подмену содержимого.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Диагностическая связь

Незащищённый участок передачи оставляет данные доступными для перехвата или подмены и указывает на нарушение транспортной части требования.

Ограничение: Связь не подтверждает выполнение требования и не оценивает шифрование хранимых копий или управление их ключами.

Условие применения: При передаче данных или артефактов по каналу без требуемого защищённого протокола.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	60	44
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	61	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	62	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	63	69
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	64	82
AISAFE2-IMP-011 · Компрометация баз знаний	336	109
AISAFE2-IMP-012 · Кража модели	337	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	338	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	339	118
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	612	481
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	613	550

Недостаток · AISAFE2-WEAK-025

Недостаточное управление ключами и секретами

Краткое описание

Ключи и секреты не отделены от защищаемых данных и операций по владельцу, назначению или сроку действия.

Подробное описание

Недостаток расположен в контуре идентичности и управления секретами: один субъект либо путь может совмещать получение шифротекста, ключа и полномочия на расшифрование, а отзыв или ротация не ограничивают доступ вовремя. Секреты дополнительно оказываются в коде, конфигурациях, образах, журналах и резервных копиях, переиспользуются между назначениями и средами, а операции выдачи, использования, ротации и уничтожения не разделяются и не протоколируются.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-028 · Управление ключами и секретами

Прямые причинные воздействия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Компрометация разделяемого, встроенного или долгоживущего административного секрета даёт атакующему доступ к штатной конфигурации, через которую он фактически читает системный промпт либо инструкцию инструмента.

Ограничение: Только если скомпрометированный ключ или секрет дал неуполномоченному субъекту фактически прочитать системный промпт либо инструкцию инструмента

AISAFE2-IMP-011 · Компрометация баз знаний

Компрометация криптографического ключа позволяет расшифровать хранимый корпус, а компрометация разделяемого, встроенного или долгоживущего административного секрета — экспортировать через функции управления существенную часть базы знаний, её метаданных или индекса.

Ограничение: Только если скомпрометированный ключ или секрет позволил фактически расшифровать либо экспортировать существенную часть базы знаний или индекса

AISAFE2-IMP-012 · Кража модели

Компрометация криптографического ключа снимает защиту хранимых или передаваемых артефактов модели, а компрометация разделяемого, встроенного или долгоживущего административного секрета даёт доступ к реестру артефактов; в обоих случаях атакующий может несанкционированно скопировать веса или связанные артефакты.

Ограничение: Только если скомпрометированный ключ или секрет привёл к фактическому копированию весов либо связанных артефактов без разрешения

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Компрометация криптографического ключа либо привилегированного секрета позволяет атакующему расшифровать или через штатный интерфейс прочитать либо изменить хранимый объект вне предоставленных ему полномочий.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Компрометация разделяемого, встроенного или долгоживущего учётного секрета позволяет атакующему действовать от имени легитимного привилегированного субъекта и успешно вызвать функцию инфраструктурного компонента за пределами собственных полномочий.

Ограничение: Только если скомпрометированный секрет использован для успешного вызова функции вне собственных полномочий атакующего

Условия проявления

1. **Предпосылка.** Ключи и секреты размещены вне контролируемого хранилища: в коде, конфигурации, образе, журнале или рядом с защищаемыми данными; административный доступ может опираться на разделяемый, встроенный или бессрочный секрет.
2. **Триггер.** Субъект получает одновременно защищённые данные и ключ с полномочием расшифрования либо извлекает разделяемый секрет из утечки, конфигурации или компрометированного участника и предъявляет его сервису.
3. **Разрешающий пробел.** Секрет не изолирован по владельцу, назначению и сроку действия, а ротация и экстренный отзыв не ограничивают время использования после компрометации.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Диагностическая связь

Диагностирует слабый жизненный цикл ключей, из-за которого защита шифрованием может быть неэффективной.

Ограничение: Слабый жизненный цикл ключа указывает на риск для шифрования, но не устанавливает режим защиты и не подтверждает его применение к данным.

Условие применения: Если для шифрования данных или артефактов используются ключи с неуправляемыми правами, ротацией или отзывом.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SECRET1

Редакция: Редакция 7, версия от 17.08.2026

Диагностическая связь

Секрет без владельца, управляемого места хранения, ограниченных потребителей или проверяемого отзыва указывает на нарушение AI-SECRET1.

Ограничение: Описание недостатка не подтверждает выполнение требования; конкретные секреты, версии и потребители устанавливаются по инвентарю системы.

Условие применения: При наличии API-ключей, токенов, паролей или иных секретных значений в ИИ-системе.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(b) · Диагностический признак сбоя

Неуправляемый ключ или секрет может открыть путь к персональным данным и ослабить конфиденциальность обработки, учитываемую статьёй 32.

Граница связи: Связь возникает только при доступе секрета к охваченным данным; она не оценивает все меры безопасности и не доказывает нарушение GDPR сама по себе.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Секрет открывает доступ к системе обработки персональных данных либо позволяет их расшифровать.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	65	61
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	340	107
AISAFE2-IMP-011 · Компрометация баз знаний	341	109
AISAFE2-IMP-012 · Кража модели	342	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	343	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	344	118
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	614	383
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	615	481
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	616	554

Недостаток · AISAFE2-WEAK-026

Отсутствие ограничений частоты запросов

Краткое описание

Интерфейсы инференса принимают повторные запросы без сквозной квоты, привязанной к субъекту и совокупности его точек доступа.

Подробное описание

Массовый систематический опрос модели позволяет накапливать пары «запрос — ответ» для восстановления её функционального аналога и извлечения свойств обучающих данных, а также создавать избыточную нагрузку с ростом расходов владельца и отказом в обслуживании.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-021 · Ограничение частоты, ресурсов и стоимости

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Недостаточность механизмов ограничения числа и частоты запросов может привести к компрометации базы знаний, потому что интерфейсы инференса принимают обращения без сквозной квоты субъекта, и массовый систематический опрос воспроизводит содержимое и свойства базы знаний серией запросов.

Ограничение: Только если серия запросов без действенной квоты позволила воспроизводимо собрать существенную часть базы знаний или индекса

AISAFE2-IMP-012 · Кража модели

Недостаточность механизмов ограничения числа и частоты запросов может привести к краже модели, потому что отсутствие сквозных ограничений числа и частоты запросов делает возможным массовый опрос для восстановления функционального аналога модели.

Ограничение: Только если массовый опрос без действенной квоты позволил подтвердить функциональный аналог ценного поведения модели

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Недостаточность механизмов ограничения числа и частоты запросов может привести к раскрытию чувствительных данных из обучающего датасета, потому что неограниченное число запросов через штатный интерфейс позволяет систематически извлекать свойства и записи обучающих данных.

Ограничение: Только если серия запросов позволила подтвердить раскрытие или реконструкцию конкретной чувствительной записи, её атрибута либо участия в обучении

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Недостаточность механизмов ограничения числа и частоты запросов может привести к чрезмерному потреблению вычислительных ресурсов системы, потому что отсутствие сквозных квот числа и частоты запросов позволяет создавать избыточную вычислительную нагрузку на инфраструктуру владельца.

Ограничение: Только при фактическом превышении утверждённого бюджета вычислений, памяти, сети или ускорителей собственной среды AI-системы.

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Недостаточность механизмов ограничения числа и частоты запросов может привести к недоступности системы или непредоставлению результата в директивный срок, потому что неограниченный по числу и частоте поток запросов создаёт нагрузку, ведущую к отказу в обслуживании легитимных потребителей.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Условия проявления

1. **Предпосылка.** Интерфейсы обращения к модели принимают автоматизированные повторные запросы.
2. **Триггер.** Субъект распределяет запросы между ключами, IP-адресами, учётными записями и маршрутами, включая веб-интерфейс и API мобильного приложения.
3. **Разрешающий пробел.** Сквозная квота, объединяющая все точки доступа одного субъекта, не ведётся или не контролируется.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Диагностический признак сбоя
>
> Без сквозного лимита частоты один субъект может многократно загружать модель и наращивать расход вычислений или стоимость обслуживания.
>
> **Граница связи:** Предел запросов не заменяет бюджет токенов и цены одного дорогого вызова; сам по себе пробел не означает отказ сервиса.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	66	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	67	37
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	68	69
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	69	77
AISAFE2-IMP-011 · Компрометация баз знаний	345	109
AISAFE2-IMP-012 · Кража модели	346	111
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	347	113
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	348	123
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	349	127
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	617	462
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	618	485

Недостаток · AISAFE2-WEAK-027

Отсутствие экономического ограничения массовых запросов

Краткое описание

Низкая цена или отсутствие цены для клиента делает массовое опрашивание модели существенно дешевле потенциальных затрат владельца сервиса.

Подробное описание

Стоимость массового опроса модели для клиента несоразмерно мала по сравнению с затратами владельца на обработку и с ценностью извлекаемого результата: при бесплатном доступе, фиксированной подписке или низких тарифах предельная стоимость дополнительного запроса для клиента близка к нулю. В результате экономический сдерживающий фактор отсутствует: автоматизированные и распределяемые между учётными записями кампании извлечения модели и истощения ресурсов становятся выгодными атакующему, а расходы и потери несёт владелец сервиса.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-021 · Ограничение частоты, ресурсов и стоимости

Прямые причинные воздействия

AISAFE2-IMP-012 · Кража модели

Отсутствие экономического ограничителя позволяет продолжать массовый опрос, стоимость которого ниже ценности извлекаемого поведения, до подтверждённого построения функционального аналога модели.

Ограничение: Только если экономически доступная кампания запросов дошла до подтверждённого функционального аналога модели

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Близкая к нулю предельная стоимость поддерживает массовую кампанию, совокупная нагрузка которой превышает бюджет вычислений владельца.

Ограничение: Только при фактическом превышении утверждённого бюджета вычислений, памяти, сети или ускорителей и установленной связи с продолжающейся кампанией.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-027 · Отсутствие экономического ограничения массовых запросов
> 2. Промежуточное воздействие: AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов
> 3. Промежуточное состояние: Измеренная стоимость обработки атрибутирована массовой кампании и плательщику сверх утверждённого бюджета
> 4. Итоговое воздействие: AISAFE2-IMP-025 · Прямые финансовые потери
>
> **Механизм переходов**
>
> 1. Экономическая асимметрия и отсутствие ограничителя массовых запросов позволяют кампании создать измеримое потребление ресурсов сверх утверждённого бюджета.
> 2. Метринг связывает конкретный класс ресурсов и стоимость обработки с кампанией, владельцем расходов и проверяемым счётом.
> 3. Подтверждённая сумма после возвратов и оспаривания остаётся невозвратным чистым денежным убытком плательщика.
>
> **Условия применимости**
>
> - Фактическая массовая кампания материально усилена ценовой асимметрией и превышает утверждённый ресурсный бюджет.
> - Установлены точный плательщик, счёт, атрибуция кампании и невозвратная чистая сумма после доступных возвратов.
>
> **Исключения**
>
> - Низкая цена отдельного запроса, плановый расход или предотвращённое потребление не являются прямым финансовым убытком.
> - Один и тот же ресурсный расход нельзя одновременно учитывать как отдельный I019 и повторно добавлять к сумме I025.

Условия проявления

1. **Предпосылка.** Ценность извлекаемого результата или создаваемый ущерб превышают затраты атакующего.
2. **Триггер.** Субъект запускает массовый автоматизированный опрос модели и распределяет его между учётными записями.
3. **Разрешающий пробел.** Предельная стоимость очередного запроса для клиента близка к нулю и не отражает затрат владельца на обработку.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Диагностический признак сбоя
>
> Несоразмерно дешёвый массовый опрос позволяет атакующему переложить стоимость большого числа инференсов на владельца сервиса.
>
> **Граница связи:** Модель цены лишь повышает выгодность атаки; реальный ущерб зависит от трафика и технических бюджетов.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-012 · Кража модели	350	111
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	351	123
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	619	462
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	620	485

Недостаток · AISAFE2-WEAK-028

Недостатки наблюдаемости и обнаружения злоупотреблений

Краткое описание

События запросов, решений, действий и потребления ресурсов не собираются полностью, не связываются между компонентами и субъектами и не доводятся до мониторинга, поэтому распределённые атакующие кампании неотличимы от обычного трафика.

Подробное описание

Правила оценивают отдельные события без восстановления последовательности и профиля субъекта, поэтому распределённые, адаптивные или малоскоростные кампании выглядят как обычный трафик. Злоупотребление обнаруживается лишь после ущерба, а расследование не может восстановить проверяемую картину событий. Восстановленная последовательность или выявленное отклонение должны сопоставляться с правилом обнаружения и связываться с проверяемым уведомлением.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-033 · Наблюдаемость, обнаружение и доказательства

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Неразличимая распределённая кампания извлечения базы знаний продолжается до достижения порога компрометации, когда связанный с детектированием контур реагирования в противном случае остановил бы её раньше.

Ограничение: Только если незамеченная кампания достигла порога компрометации, а действующий контур реагирования остановил бы её по своевременному сигналу; нехватка телеметрии без этого пути недостаточна

AISAFE2-IMP-012 · Кража модели

Неразличимая распределённая кампания накопления пар запрос—ответ продолжается до функционального извлечения модели, когда связанный контур реагирования в противном случае остановил бы её раньше.

Ограничение: Только при доказанной кампании извлечения, заранее действующем контуре реагирования и подтверждённом функциональном аналоге, полученном из-за продолжения незамеченной кампании.

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Реальная аномальная или вредоносная кампания повторных обращений к модели остаётся необнаруженной, хотя достаточный сигнал должен был запустить предусмотренный контур реагирования; продолжение кампании накапливает токенный расход запросов и связанную вычислительную нагрузку до превышения утверждённого бюджета ресурсов среды.

Ограничение: Только если незамеченная кампания повторных запросов вызвала измеренный перерасход ресурсов, а действующий контур реагирования остановил бы её по своевременному сигналу; токенный предел одного запроса может сохраняться

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений
> 2. Промежуточное воздействие: AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов
> 3. Итоговое воздействие: AISAFE2-IMP-025 · Прямые финансовые потери
>
> **Механизм переходов**
>
> 1. Пробел наблюдаемости не позволяет заранее определённой реакции остановить доказанную кампанию злоупотребления до превышения ресурсного порога, и возникает измеримое избыточное потребление.
> 2. Атрибутированное потребление приводит к точной подтверждённой сумме, которая остаётся невозвратной после доступного финансового урегулирования.
>
> **Условия применимости**
>
> - Доказана кампания, а заранее определённые обнаружение и реакция остановили бы её до порога при наличии требуемой телеметрии.
> - Фактически наступили I019 и точный атрибутированный невозвратный денежный убыток.
>
> **Исключения**
>
> - Отсутствие телеметрии без эффективного контрфактического сценария реагирования недостаточно.
> - Потребление без денежного убытка и повторный учёт одной суммы исключаются.

Условия проявления

1. **Предпосылка.** Мониторинг оценивает низкочастотные, распределённые и многошаговые события порознь, без восстановления общей картины.
2. **Триггер.** Субъект распределяет злоупотребление по времени, компонентам или идентичностям.
3. **Разрешающий пробел.** События запросов, решений, действий и потребления собираются неполно либо не связываются между компонентами и идентичностями одного субъекта.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C12.1.1 · Диагностический признак сбоя

Неполные и несвязанные события AI-запроса мешают восстановить сессию, действия и участника цепочки.

Граница связи: Карта включает обнаружение злоупотреблений; само требование AISVS здесь касается регистрации взаимодействий.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	AI-приложение ведёт журнал взаимодействий с контекстом сессии и AI-специфичной телеметрией по AISVS 1.0.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	70	37
AISAFE2-COMP-013 · Телеметрия и обнаружение	71, 72	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	73	69
AISAFE2-IMP-011 · Компрометация баз знаний	352	109
AISAFE2-IMP-012 · Кража модели	353	111
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	354	123
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	621	400
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	622	405
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	623	456

Недостаток · AISAFE2-WEAK-030

Отсутствие проверяемой маркировки модели и её выходов

Краткое описание

У модели, её артефактов или характерных выходов нет проверяемой контрольной метки, пригодной как дополнительный признак происхождения и несанкционированного копирования.

Подробное описание

Недостаток становится значимым при расследовании подозрительной копии, модели-замены или производного артефакта, когда требуется сопоставить объект с защищаемой версией по заранее заложенному и проверяемому сигналу. Проверяемый сигнал может быть заложен в артефакт или выходные данные либо задан как заранее зафиксированный контрольный профиль поведения и должен быть связан с точной защищаемой версией до её распространения.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-033 · Наблюдаемость, обнаружение и доказательства

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-030 · Отсутствие проверяемой маркировки модели и её выходов
> 2. Промежуточное состояние: Независимо подтверждена неразрешённая копия или производный продукт, но отсутствие маркера препятствует версионной атрибуции
> 3. Промежуточное состояние: Работоспособное вмешательство своевременно не запускается, и существенное использование актива конкурентом продолжается
> 4. Итоговое воздействие: AISAFE2-IMP-030 · Утеря конкурентного преимущества
>
> **Механизм переходов**
>
> 1. При уже независимо установленной копии отсутствие заранее внедрённого надёжного маркера лишает расследование версионной технической атрибуции.
> 2. Без технической атрибуции не запускается заранее определённое своевременное и эффективное вмешательство, поэтому использование существенного актива продолжается.
> 3. Продолженное существенное использование конкурентом причинно ухудшает заранее выбранную и измеренную метрику конкурентной позиции владельца.
>
> **Условия применимости**
>
> - Кража, копия или производный продукт установлены независимо от отсутствующего маркера, а маркер надёжно различал бы точную версию.
> - Существует определённый своевременный путь вмешательства, конкурент использует существенный актив и выбранная метрика фактически ухудшилась.
>
> **Исключения**
>
> - Отсутствие маркировки не считается причиной самой кражи и не доказывает наличие копии.
> - Визуальное сходство, отсутствие юридического доказательства или невозможность атрибуции без измеримого конкурентного результата недостаточны.

Условия проявления

1. **Предпосылка.** Модель, её артефакты или характерные выходы могут быть скопированы или использованы стороной.
2. **Триггер.** Появляется подозрительная копия или производный продукт, принадлежность которых требуется установить.
3. **Разрешающий пробел.** Проверяемая метка происхождения, например водяной знак или контрольный образец поведения, заранее не заложена.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	74	37
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	75	56
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	624	485

Недостаток · AISAFE2-WEAK-031

Недостаточная защита модели при поставке в недоверенную среду

Краткое описание

Артефакт модели передается для исполнения в среду, где владелец модели не контролирует хост, память и пути копирования.

Подробное описание

Недостаток активируется, когда владелец недоверенной среды или получатель артефакта может исследовать исполняемый экземпляр, память, временные файлы и штатные функции экспорта без независимой границы, сохраняемой владельцем модели. Без независимой аппаратной границы шифрование и обфускация только повышают стоимость извлечения, поскольку исполняемый экземпляр и рабочее состояние доступны среде получателя.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-032 · Цепочка поставки, недоверенная среда и физическая защита

Прямые причинные воздействия

AISAFE2-IMP-012 · Кража модели

Недостаточная защита модели при поставке в недоверенную среду может привести к краже модели, потому что артефакт исполняется в среде, которую контролирует получатель, расшифрованные веса неизбежно оказываются в памяти процесса и копируются из дистрибутива, памяти или временных файлов.

Ограничение: Только если из дистрибутива, памяти или временных файлов в недоверенной среде были фактически скопированы веса либо связанные артефакты без разрешения

Условия проявления

1. **Предпосылка.** Получатель имеет доступ к дистрибутиву модели, памяти процесса и временным файлам в контролируемой им среде.
2. **Триггер.** Артефакт модели передаётся для исполнения на пользовательское или бортовое устройство либо во внешнее развёртывание.
3. **Разрешающий пробел.** Независимая от получателя граница исполнения, например обработка на стороне владельца, не применяется.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	76	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	77	58
AISAFE2-IMP-012 · Кража модели	355	111

Недостаток · AISAFE2-WEAK-032

Недостаточная защита физического доступа к модели и инфраструктуре

Краткое описание

Физический или аппаратно-привилегированный доступ к хосту, носителю либо ускорителю может обойти логические ограничения на артефакт модели.

Подробное описание

Недостаток охватывает непосредственные пути к хосту, GPU и артефакту модели, но не описывает полную общеорганизационную программу физической безопасности.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-032 · Цепочка поставки, недоверенная среда и физическая защита

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Недостаточная защита физического доступа к модели и инфраструктуре может привести к компрометации базы знаний, потому что физический доступ к носителям и серверам хранилищ позволяет скопировать корпус и индекс в обход логических средств защиты.

Ограничение: Только если физический доступ привёл к копированию существенной части базы знаний или индекса субъектом без полномочий

AISAFE2-IMP-012 · Кража модели

Физический или аппаратно-привилегированный доступ может открыть путь к артефакту модели, если защита носителя, ключей и локального доступа не перекрывает этот путь.

Ограничение: Только если физический доступ привёл к неразрешённому копированию весов или связанных артефактов модели

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

При физическом или аппаратно-привилегированном доступе субъект может прочитать либо изменить хранимые данные, если защита носителя, ключей и локального доступа допускает такую операцию.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Недостаточная защита физического доступа к модели и инфраструктуре может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что субъект с физическим доступом вмешивается в работу оборудования и компрометирует компоненты инфраструктуры в обход логических средств контроля.

Ограничение: Только если физический доступ привёл к вызову инфраструктурной функции без полномочий; вмешательство в оборудование без такого вызова сюда не входит

Условия проявления

1. **Предпосылка.** На оборудовании и носителях находятся модель, данные или иные ценные компоненты системы.
2. **Триггер.** Субъект получает физический или аппаратно-привилегированный доступ к оборудованию, носителям или памяти.
3. **Разрешающий пробел.** Защита содержимого опирается только на логические механизмы: учётные записи, сеть и программные контроли.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	78	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	79	58
AISAFE2-IMP-011 · Компрометация баз знаний	356	109
AISAFE2-IMP-012 · Кража модели	357	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	358	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	359	118

Карточка	Связи	Стр.
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	625	481

Недостаток · AISAFE2-WEAK-033

Раскрытие сведений об отдельных обучающих записях через модель

Краткое описание

Модель сохраняет относимую к отдельным обучающим записям информацию, которую через инференс можно воспроизвести либо использовать для вывода об участии записи и её атрибутах.

Подробное описание

Обучение может закрепить в параметрах информацию об отдельных, особенно редких или чувствительных, записях сверх требуемых обобщённых закономерностей. Через целевые запросы и сопоставление ответов со вспомогательными данными внешний субъект может вызвать фрагменты записи, определить её участие в обучении или вывести отдельные атрибуты. Удаление исходной записи после обучения само по себе не устраняет её влияние из выпущенной модели; отсутствие наблюдаемого раскрытия также не доказывает отсутствие такого влияния.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-002 · Поведение при обучении и дообучении
- AISAFE2-WC-018 · Минимизация данных и приватность

Прямые причинные воздействия

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Модель сохраняет сведения об отдельной чувствительной обучающей записи; целевые запросы к выпущенной версии позволяют неуполномоченному субъекту воспроизвести её фрагмент, вывести атрибут или подтвердить участие записи в обучении. Удаление исходной записи из датасета само по себе не меняет выпущенную модель.

Условия проявления

1. **Предпосылка.** В обучении участвовали чувствительные или редкие записи.
2. **Триггер.** Внешний субъект выполняет целевые запросы к модели и сопоставляет ответы со вспомогательными сведениями.
3. **Разрешающий пробел.** Воспроизведение фрагментов записей и вывод об их участии в обучении не ограничиваются ни при выпуске, ни в эксплуатации.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Диагностический признак сбоя
>
> Запоминание отдельных обучающих записей создаёт путь к их воспроизведению или выводу об участии записи и её атрибутах через ответы модели.
>
> **Граница связи:** Другие утечки через текущий контекст, журналы, интеграции и выходные каналы этой слабостью не объясняются.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	80, 81	37
AISAFE2-COMP-008 · Данные для подсистемы обучения	82	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	83	54
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	360	113
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	626, 627	478

Недостаток · AISAFE2-WEAK-035

Недостатки управления жизненным циклом данных

Краткое описание

Наборы данных AI-системы собираются, повторно используются, передаются и включаются в обучение без назначенного владельца, инвентаризации, классификации по чувствительности, срока хранения и проверяемого решения о допустимости использования.

Подробное описание

Недостаток становится значимым, когда данные собираются, повторно используются, передаются или включаются в обучение без владельца, метки чувствительности, срока хранения и проверяемого решения о допустимости.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-016 · Управление и происхождение данных

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Набор без владельца, классификации и решения о допустимости включается в инструментальный источник, после чего чувствительное содержание фактически выдаётся неуполномоченному получателю.

Ограничение: Только при фактической выдаче или воспроизводимой реконструкции чувствительных сведений из инструментального источника неуполномоченным получателем.

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Чувствительный набор без классификации и решения о допустимости включается в обучение, после чего запись или атрибут воспроизводимо раскрывается через модель.

Ограничение: Только при подтверждённом воспроизведении, определении принадлежности, логическом выводе или восстановлении конкретной чувствительной записи либо её атрибута.

Условия проявления

1. **Предпосылка.** Набор данных собирается, повторно используется, передаётся или готовится к включению в обучение.
2. **Триггер.** Операция выполняется без требуемых подтверждений, хотя набор может содержать чувствительные или неправомерно используемые данные.
3. **Разрешающий пробел.** У набора нет владельца, инвентарной записи, класса чувствительности, срока хранения или зафиксированного решения о допустимости использования.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-008 · Данные для подсистемы обучения	84	52
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	85	67
AISAFE2-COMP-025 · Источники данных для инструментов	86	90
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	361	106
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	362	113
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений	628	423
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	629	509

Недостаток · AISAFE2-WEAK-036

Передача в AI-контур избыточных данных (нарушение минимизации)

Краткое описание

В AI-контур передаются записи или поля, которые не нужны для заявленной цели обработки и могут быть исключены.

Подробное описание

Недостаток активируется, когда набор сохраняет лишние записи, атрибуты или исторические поля вместо удаления до передачи в контекст, RAG, память, журналы или обучение.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-018 · Минимизация данных и приватность

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Лишние чувствительные поля или записи не удаляются до передачи в RAG или контекст и фактически выдаются получателю без полномочий.

Ограничение: Только при фактической выдаче или воспроизводимой реконструкции чувствительных сведений из инструментального источника неуполномоченным получателем.

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Не нужные для цели чувствительные записи включаются в обучение и затем воспроизводимо выводятся либо восстанавливаются через модель.

Ограничение: Только при подтверждённом воспроизведении, определении принадлежности, логическом выводе или восстановлении конкретной чувствительной записи либо её атрибута.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Избыточные чувствительные поля могут попасть к внешнему провайдеру, которому передача именно этих данных не разрешена утверждённой политикой; фактический запрос тогда пересекает запрещённую границу.

Ограничение: Только если избыточные данные фактически переданы провайдеру, не утверждённому для этого класса данных и цели; избыток в разрешённом контуре сам по себе не утечка

Условия проявления

1. **Предпосылка.** Часть передаваемых в AI-контур полей и записей не требуется для заявленной цели обработки.

2. **Триггер.** Конвейер передаёт избыточные поля и записи без удаления в контекст, RAG, память, журналы или обучение.
3. **Разрешающий пробел.** До передачи в AI-контур не проверяется минимально необходимый для заявленной цели состав полей и записей.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 5(1)(c)** · Диагностический признак сбоя
>
> Передача в AI-контур полей персональных данных, не нужных для заявленной цели, показывает конкретный риск нарушения принципа минимизации статьи 5.
>
> **Граница связи:** Ненужность поля определяется по цели и операции обработки; карточка сама не устанавливает правовое основание или применимость GDPR.

Субъект	Контролёр персональных данных по GDPR
Условие применения	Операция обработки персональных данных подпадает под действие GDPR.
Условие связи с карточкой	Контролёр установил фактические персональные данные, цель обработки и избыточные поля до границы AI-контура.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-003 · Управление контекстом общения с моделью	87	40
AISAFE2-COMP-007 · Долговременная память	88	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	89	52
AISAFE2-COMP-013 · Телеметрия и обнаружение	90	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	91	69
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	92	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	93	77
AISAFE2-COMP-023 · Подсистема формирования базы знаний	94	86
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	363	106
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	364	113
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	365	130

Карточка	Связи	Стр.
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	630	349

Недостаток · AISAFE2-WEAK-037

Непропорциональный вклад отдельных обучающих записей в результат обучения

Краткое описание

Отсутствие дедупликации и ограничения вклада отдельных записей концентрирует влияние немногих примеров на модели, позволяя как извлекать их, так и целенаправленно менять поведение модели небольшим числом внедрённых записей.

Подробное описание

Недостаток возникает, когда конвейер не выявляет точные и приблизительные дубликаты и не ограничивает совокупный вклад отдельной записи, субъекта или группы связанных записей в обучающий сигнал, повышая риск запоминания, отравления и восстановления данных.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-002 · Поведение при обучении и дообучении

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Непропорциональный вклад отдельных обучающих записей в результат обучения может привести к неожидаемым для разработчика ответам модели, потому что из-за отсутствия дедупликации немногочисленные намеренно внедрённые записи непропорционально влияют на обучение и позволяют целенаправленно изменить поведение модели, включая внедрение закладки.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Непропорциональный вклад отдельных обучающих записей в результат обучения может привести к раскрытию чувствительных данных из обучающего датасета, потому что дубликаты и связанные записи не ограничиваются по суммарному вкладу, запоминание конкретных примеров усиливается, и становится возможным их извлечение и определение участия в обучении.

Условия проявления

1. **Предпосылка.** Обучающий набор содержит повторяющиеся или связанные записи, в том числе примеры, внедрённые атакующим.
2. **Триггер.** Обучение агрегирует вклады этих примеров.

3. **Разрешающий пробел.** Конвейер подготовки не выявляет дубликаты и не ограничивает вклад одной записи, субъекта или группы.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-008 · Данные для подсистемы обучения	95	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	96	54
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	366	100
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	367	113
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	631	478

Недостаток · AISAFE2-WEAK-039

Выдача инструментом чувствительных или избыточных данных

Краткое описание

Инструмент или коннектор возвращает получателю чувствительные либо избыточные данные без минимизации, проекции полей и сохраняемой метки чувствительности.

Подробное описание

В зависимости от маршрута чувствительные или избыточные данные могут попасть в контекст модели, интерфейс, другие инструменты, журналы или внешний сервис и раскрыться через ответ. При передаче внешней модели данные могут покинуть контролируемый контур организации.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-018 · Минимизация данных и приватность

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Инструмент возвращает чувствительные или избыточные поля без проекции, усечения и сохранения метки чувствительности; результат поступает в модель или интерфейс и раскрывает данные неуполномоченному получателю.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Инструмент передаёт чувствительные данные без минимизации или проверки меток внешней модели либо иному получателю, которому эти данные не разрешены; фактическая передача пересекает запрещённую границу.

Ограничение: Только если инструмент выдал чувствительные поля модели или получателю, которым их передавать нельзя; избыточный объём при разрешённом получателе сам по себе не утечка

Условия проявления

1. **Предпосылка.** Инструмент обращается к источнику с чувствительными сведениями, а запрос допускает широкую выборку.
2. **Триггер.** Ответ передаётся в контекст модели, интерфейс, журнал или смежный компонент.
3. **Разрешающий пробел.** Результат возвращается целиком, без отбора полей под цель операции, усечения, обезличивания и метки чувствительности.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Диагностический признак сбоя
>
> Избыточный ответ инструмента может поместить закрытые поля в контекст модели, журнал или последующий ответ, открытый чужому получателю.
>
> **Граница связи:** Выдача инструментом внутри разрешённой области не равна утечке; получатель и последующий маршрут должны быть проверены.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	97	44
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	368	106
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	369	130
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	632	441

Недостаток · AISAFE2-WEAK-040

Отсутствие маскирования или анонимизации чувствительных данных

Краткое описание

Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Подробное описание

Слабость проявляется, когда сохраняемое значение можно преобразовать без потери требуемой функции, но оно передаётся AI-компонентам либо в журналы, память и наборы без такого преобразования.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-018 · Минимизация данных и приватность

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Отсутствие маскирования или анонимизации чувствительных данных может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что необходимые для задачи данные циркулируют в исходном виде вместо маскированного, и попавшие в контекст и ответы значения раскрывают чувствительные сведения без потери функции защищённой формы.

Ограничение: Только если немаскированные сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; само хранение исходных значений недостаточно

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Отсутствие маскирования или анонимизации чувствительных данных может привести к раскрытию чувствительных данных из обучающего датасета, потому что необходимые записи сохраняются в обучающих наборах с чувствительными значениями в исходном виде вместо маскированного, и запоминание значений моделью раскрывает их.

Ограничение: Только если исходное немаскированное значение из обучающего набора воспроизведено или восстановлено через модель; его хранение в датасете недостаточно

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Исходные чувствительные значения из контекста, журнала или набора данных без маскирования передаются получателю, которому их обрабатывать нельзя, и пересекают запрещённую границу.

Ограничение: Только если исходные чувствительные значения из журнала, набора или контекста ушли получателю или среде без разрешения; отсутствие маскирования внутри доверенного контура недостаточно

Условия проявления

1. **Предпосылка.** В обработке участвуют чувствительные значения, функция которых сохраняется после маскирования или анонимизации.
2. **Триггер.** Значения передаются модели, журналу, памяти или набору данных в исходном виде.
3. **Разрешающий пробел.** Эти точки доступны субъектам и процессам, которым исходные значения не предназначены.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 25(1)-(2)** · Диагностический признак сбоя
>
> Отсутствие предусмотренного маскирования или токенизации на AI-пути показывает пробел в выбранной защите данных на этапе проектирования.
>
> **Граница связи:** Статья 25 не предписывает маскирование для каждого поля; нужность метода, настройки по умолчанию и другие меры оцениваются по цели и риску.

Субъект	Контролёр персональных данных по GDPR
Условие применения	Установлены фактическая операция обработки и соразмерность меры риску.
Условие связи с карточкой	Контролёр обрабатывает персональные данные и задокументировал соразмерное риску преобразование, сохраняющее заявленную цель.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-003 · Управление контекстом общения с моделью	98	40
AISAFE2-COMP-007 · Долговременная память	99	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	100	52
AISAFE2-COMP-013 · Телеметрия и обнаружение	101	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	102	69
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	103	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	104	77
AISAFE2-COMP-023 · Подсистема формирования базы знаний	105	86

Карточка	Связи	Стр.
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	370	106
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	371	113
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	372	130
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	633	349

Недостаток · AISAFE2-WEAK-041

Недостаточный контроль доступа к системным инструкциям и секретам

Краткое описание

Доступ к конфигурации системного промпта, инструкциям инструментов или связанным секретам не ограничен по субъекту и операции.

Подробное описание

Недостаток проявляется, когда субъект может напрямую или через избыточные полномочия читать либо изменять эти объекты. Чтение раскрывает защищаемые инструкции и секреты, а изменение системного промпта или описания инструмента непосредственно меняет поведение AI-системы.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-010 · Контроль доступа и авторизация
- AISAFE2-WC-028 · Управление ключами и секретами

Прямые причинные воздействия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Недостаточный контроль доступа к системным инструкциям и секретам может привести к раскрытию системного промпта и инструкций инструментов, потому что доступ к конфигурации системного промпта и инструкциям инструментов не ограничен по субъекту и операции, и их читает более широкий круг, чем необходимо.

Ограничение: Только если субъект без соответствующего права фактически прочитал системный промпт либо инструкцию инструмента через конфигурацию

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостаточный контроль доступа к системным инструкциям и секретам может привести к несанкционированному чтению или изменению хранимых данных, потому что круг субъектов, способных читать и изменять системные инструкции, описания инструментов и связанные секреты, шире необходимого, и несанкционированное изменение меняет поведение системы.

Условия проявления

1. **Предпосылка.** Системный промпт, инструкции инструментов или связанные с ними секреты доступны на чтение или изменение кругу субъектов шире необходимого.
2. **Триггер.** Субъект использует этот доступ напрямую или через избыточные полномочия.
3. **Разрешающий пробел.** Среда выполнения применяет изменённое содержимое, не проверяя правомерность изменения.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-003 · Управление контекстом общения с моделью	106	40
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	107	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	108	61
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	109	67
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	110	75
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	373	107
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	374	116
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	634	371
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	635	554

Недостаток · AISAFE2-WEAK-042

Недостаточная пригодность модели для целевого сценария

Краткое описание

Конкретная версия модели не проходит фактическую репрезентативную проверку по заранее заданным критериям качества или безопасности целевого сценария.

Подробное описание

Если версия не достигает заданных критериев, её результаты в целевом сценарии могут быть недостоверными или небезопасными; использование таких результатов повышает риск ошибочных решений и ненадёжной работы сценария.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-005 · Выбор и допуск модели

Прямые причинные воздействия

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Недостаточная пригодность модели для целевого сценария может привести к ущербу от решения, принятого на основе результата AI-системы, потому что модель, не соответствующая целевому сценарию по критериям качества, систематически ошибается именно там, где применяется, и пользователи принимают решения на недостоверных результатах.

Ограничение: Только если непригодность модели для целевого сценария дала ошибочный результат, на котором существенно основывалось исполненное вредное решение

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Недостаточная пригодность модели для целевого сценария может привести к неожидаемым для разработчика ответам модели, потому что конкретная версия модели проваливает граничные случаи и теряет требуемые свойства, систематически выдавая результаты за предусмотренными границами качества и поведения.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария
> 2. Промежуточное воздействие: AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей
> 3. Промежуточное состояние: Именованный обязательный последующий шаг принимает фактически неверный результат без исправления
> 4. Итоговое воздействие: AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
>
> **Механизм переходов**
>
> 1. Точная версия не проходит репрезентативную проверку по утверждённому критерию, но её допускают к целевому сценарию; совпадающий рабочий вход вызывает фактический выход за допустимую границу качества.
> 2. Обязательный последующий шаг принимает ошибочный результат без отклонения или исправления и использует его в своей операции.
> 3. Использование принятого ошибочного результата вызывает наблюдаемое отклонение именованного обязательного бизнес-процесса.
>
> **Условия применимости**
>
> - Установлены точная проваленная размерность теста, совпадающий производственный вход и фактический выход за установленную границу качества.
> - Обязательный последующий шаг принимает результат без исправления, а журнал подтверждает фактическое отклонение процесса.

Исключения

- Сам провал теста без фактического события в эксплуатации или результат, который был отклонён либо исправлен, не образуют путь.
- Совещательное использование без обязательного процессного шага относится к другой границе воздействия.

Условия проявления

1. **Предпосылка.** Проверка по заранее заданным критериям показала, что версия модели не достигает порога, проваливает граничные случаи или теряет требуемое свойство.
2. **Триггер.** Рабочие запросы попадают на целевые или граничные входы, а результаты используются как основание решений или операций.
3. **Разрешающий пробел.** Версия тем не менее допускается или сохраняется в эксплуатации целевого сценария.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

MANAGE 1.1 · Диагностический признак сбоя

Непригодность проверенной версии модели для целевого сценария даёт основание пересмотреть решение о выпуске AI-системы.

Граница связи: Карточка касается модели и выбранной проверки; остальные цели системы и итоговое решение о разработке или выпуске оцениваются отдельно.

Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	Владелец решает, достигает ли конкретная AI-система заявленной цели и допустим ли её выпуск, используя добровольный NIST AI RMF.
Условие связи с карточкой	Проверенная версия модели входит в AI-систему, назначение которой пересматривается.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	111	37
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	112	67
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	375	93
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	376	100
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	636	466

Недостаток · AISAFE2-WEAK-043

Недостаточные ограничения в системной инструкции

Краткое описание

Системная инструкция неполно, неоднозначно или противоречиво задаёт запреты и границы поведения, ожидаемые в целевом сценарии.

Подробное описание

Ограничение может быть задано только для известных формулировок запроса, а не для запрещённого намерения, оставляя семантически эквивалентные запросы формально допустимыми.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-008 · Системные инструкции и детерминированные контроли

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Недостаточные ограничения в системных инструкциях могут привести к неожидаемым для разработчика ответам модели, потому что неполные, неоднозначные или противоречивые системные инструкции оставляют формально не запрещённые запросы обслуживаемыми штатно, за пределами задуманных разработчиком границ.

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Недостаточные ограничения в системных инструкциях могут привести к обходу ограничений и предусмотренной логики агента, потому что границы поведения заданы в системных инструкциях неполно или противоречиво, и формально не запрещённые запросы уводят агента от предусмотренного сценария работы.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Недостаточные ограничения в системных инструкциях могут привести к раскрытию системного промпта и инструкций инструментов, потому что системная инструкция не запрещает запросы вида повтори всё, что было выше, и содержимое контекста раскрывается формально не запрещёнными просьбами.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Системная инструкция не задаёт требуемый запрет внешней передачи, и модель или агент фактически отправляет чувствительные данные недоверенному получателю.

Ограничение: Только если утверждённая политика запрещала конкретную передачу, этот запрет не был задан агенту и данные фактически ушли запрещённому адресату

Условия проявления

1. **Предпосылка.** Необходимый запрет отсутствует в системной инструкции, задан неоднозначно или конфликтует с другой директивой.
2. **Триггер.** Поступает пограничный или специально подобранный запрос, попадающий в этот пробел.
3. **Разрешающий пробел.** Соблюдение границы возложено только на модель без внешнего принуждения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM01:2026** · Диагностический признак сбоя
>
> Неполная системная инструкция облегчает прямому вредоносному запросу обход заявленных ограничений поведения модели.
>
> **Граница связи:** Текст системной инструкции не создаёт надёжной границы доверия; слабость не объясняет внедрение через RAG или инструмент и не доказывает обход.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Недоверенное содержимое попадает в контекст модели и может пересечь границу между данными и инструкциями.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-003 · Управление контекстом общения с моделью	113	40
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	377	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	378	101
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	379	107
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	380	130
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	637	429

Недостаток · AISAFE2-WEAK-044

Недостатки управления изменениями и развёртыванием AI-системы

Краткое описание

Разработка и выпуск AI-системы допускают изменения без обязательных проверок безопасности, владельцев решения и подтверждений, связанных с точной версией.

Подробное описание

Недостаток проявляется, когда модель, библиотека, конфигурация, инструмент или иной компонент попадает в рабочую среду без оценки, проверяемого решения о выпуске, ограниченного развёртывания, плана отката и повторной проверки после изменения версии либо условий эксплуатации.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-035 · Безопасный выпуск и управление уязвимостями

Прямые причинные воздействия

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Недостатки управления изменениями и развёртыванием AI-системы могут привести к неожидаемым для разработчика ответам модели, потому что изменения модели, библиотек и конфигурации попадают в рабочую среду без обязательных проверок защищённости, и в эксплуатации появляются нежелательно ведущие себя версии.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Недостатки управления изменениями и развёртыванием AI-системы могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что изменения публикуются без поэтапного развёртывания и плана отката, и дефектную версию нельзя быстро откатить, продлевая недоступность сервиса.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Условия проявления

1. **Предпосылка.** Подготовлено изменение модели, библиотеки, конфигурации или инструмента, которое должно пройти в рабочую среду.
2. **Триггер.** Изменение продвигается в рабочую среду, и эксплуатация принимает новый состав как разрешённый.

3. **Разрешающий пробел.** Выпуск не требует проверки защищённости, решения владельца и поэтапного развёртывания и не обеспечивает готовый план быстрого отката.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	114	67
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	381	100
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	382	127
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	638	364
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	639	415
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	640	448
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	641	514
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	642	577

Недостаток · AISAFE2-WEAK-045

Недостатки управления учётными записями и правами в контуре MLOps

Краткое описание

Человеческие и административные учётные записи MLOps получают общие, избыточные или долговечные права, которые не пересматриваются при изменении роли.

Подробное описание

Недостаток проявляется в выдаче, периодическом пересмотре и отзыве прав операторов, разработчиков и администраторов, когда устаревший доступ сохраняется после смены обязанностей или завершения работ.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-037 · Закрытие прав и возможностей жизненного цикла

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Недостатки управления учётными записями и правами в контуре MLOps могут привести к компрометации базы знаний, потому что избыточные и неотозванные права учётных записей в контуре MLOps дают широкий несанкционированный доступ к данным и хранилищам базы знаний.

Ограничение: Только если избыточное право MLOps распространялось на базу знаний и было использовано для получения её существенной части

AISAFE2-IMP-012 · Кража модели

Недостатки управления учётными записями и правами в контуре MLOps могут привести к краже модели, потому что накопленные избыточные и неотозванные права в контуре MLOps дают несанкционированный доступ к артефактам модели в конвейерах и хранилищах.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостатки управления учётными записями и правами в контуре MLOps могут привести к несанкционированному чтению или изменению хранимых данных, потому что сменившие роль и уволившиеся сохраняют права в конвейерах и хранилищах, и компрометация такой учётной записи даёт широкий доступ к данным, артефактам и конфигурации.

AISAFE2-IMP-031 · Утрата датасета в результате атаки

Устаревшее либо избыточное человеческое или административное право MLOps используется для удаления или повреждения датасета вместе со всеми доступными доверенными копиями, вследствие чего датасет нельзя восстановить для обучения, тестирования или валидации.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения всех доступных доверенных копий.

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Устаревшее либо избыточное человеческое или административное право MLOps используется для удаления или повреждения весов, базы знаний, индекса либо критичной конфигурации вместе с доступными версиями и рецептами восстановления.

Ограничение: Только для вызванной атакой невосстановимой утраты весов, базы знаний, индекса или критичной конфигурации при отсутствии пригодной версии либо доверенного рецепта восстановления.

Условия проявления

1. **Предпосылка.** Учётная запись в контуре MLOps имеет избыточные права либо сохраняет их после смены роли, завершения работ или увольнения.
2. **Триггер.** Запись используется со злоупотреблением или компрометируется, а конвейеры и хранилища продолжают ей доверять.
3. **Разрешающий пробел.** Периодический пересмотр и отзыв прав не выполняются.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-009 · Конвейер обучения и дообучения	115	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	116	56
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	117	61
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	118	67
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	119	79
AISAFE2-IMP-011 · Компрометация баз знаний	383	109
AISAFE2-IMP-012 · Кража модели	384	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	385	116
AISAFE2-IMP-031 · Утрата датасета в результате атаки	386	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	387	141
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	643	525

Недостаток · AISAFE2-WEAK-046

Недостаточное управление соответствием требованиям

Краткое описание

Применимые требования, владельцы и подтверждения соответствия не связаны с конкретной архитектурой и выпуском AI-системы.

Подробное описание

Недостаток становится значимым при новом выпуске, изменении архитектуры, внешнего поставщика или условий обработки данных, когда не определены применимые обязательства, критерии допуска, ответственный владелец, договорные и технические гарантии, подтверждения независимой оценки и период пересмотра.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-036 · Соответствие и допустимое использование

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

Путь 1

1. Исходный недостаток: AISAFE2-WEAK-046 · Недостаточное управление соответствием требованиям
2. Промежуточное воздействие: AISAFE2-IMP-027 · Несоответствие требованиям регуляторов
3. Промежуточное состояние: Установленное нарушение получает публичную огласку и однозначно атрибутируется организации или продукту
4. Итоговое воздействие: AISAFE2-IMP-029 · Репутационный ущерб

Механизм переходов

1. Непрослеженное точное российское обязательство приводит к выпуску версии, для которой применимый внутренний контроль фактически не выполнен и нарушение установлено.
2. Установленное нарушение публикуется для значимой аудитории и связывается с конкретной организацией либо продуктом.
3. Публичная атрибуция установленного нарушения причинно снижает заранее выбранную и измеренную метрику доверия.

Условия применимости

- Для точной версии установлено применимое российское регуляторное требование; невыполнение соответствующего контроля подтверждено для конкретной конфигурации и периода.
- Факт получает публичную огласку и атрибуцию, после чего заранее выбранная метрика доверия измеримо снижается.

Исключения

- Неполные подтверждающие материалы поставщика, иностранная норма или отсутствие внутреннего документа без установленного нарушения недостаточны.
- Частное или неатрибутированное событие и публикация без измеримого результата доверия не образуют финальный переход.

Условия проявления

1. **Предпосылка.** Система работает в сценарии с обязательными регуляторными требованиями либо изменения в архитектуре, смена поставщика или условий обработки данных должны соответствовать таким требованиям.
2. **Триггер.** Изменение или эксплуатация продолжаются без подтверждения соответствия до инцидента и реакции на него со стороны регулятора.
3. **Разрешающий пробел.** Применимые требования, критерии допуска, владелец для конкретной версии не определены и/или подтверждение соответствия не выполнено.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	120	67
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений	644	423

Карточка	Связи	Стр.
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	645	426
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	646	519

Недостаток · AISAFE2-WEAK-047

Недостаточное управление уязвимостями

Краткое описание

Эксплуатируемые AI-компоненты не имеют полного инвентаря и процесса оценки, исправления и подтверждения устранения уязвимостей.

Подробное описание

Корневая слабость охватывает конвейер обучения, артефакты, инфраструктуру и выпуск: без единого инвентаря одна находка не сопоставляется со всеми затронутыми версиями, средами и путями распространения компонента. Для каждой находки не назначаются владелец и срок, а исправление не подтверждается повторной проверкой во всех затронутых версиях и средах.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-035 · Безопасный выпуск и управление уязвимостями

Прямые причинные воздействия

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостаточное управление уязвимостями может привести к несанкционированному чтению или изменению хранимых данных, потому что известные уязвимости остаются применимыми к рабочим компонентам неопределённо долго, и их эксплуатация даёт несанкционированный доступ к хранимым данным.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Недостаточное управление уязвимостями может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что известные, но не устранённые уязвимости компонентов эксплуатируются для выполнения функций и произвольного кода вне полномочий.

Ограничение: Только если эксплуатация неустранённой уязвимости привела к успешному вызову функции вне полномочий атакующего

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Недостаточное управление уязвимостями может привести к недоступности системы или непредоставлению результата в директивный срок, потому что эксплуатация неустранённых уязвимостей выводит компоненты из строя, останавливая обслуживание потребителей.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

AISAFE2-IMP-031 · Утрата датасета в результате атаки

Атакующий эксплуатирует неустранённую уязвимость для разрушительного доступа к точному датасету и уничтожает все пригодные доверенные копии и пути воспроизводимого восстановления.

Ограничение: Только если неустранённая уязвимость использована для уничтожения точного датасета и всех пригодных копий или путей восстановления; невосстановимость подтверждена

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Атакующий эксплуатирует неустранённую уязвимость для разрушительного доступа к весам, базе знаний, индексу либо критичной конфигурации и уничтожает все пригодные доверенные копии и пути воспроизводимого восстановления.

Ограничение: Только если неустранённая уязвимость использована для уничтожения точных весов, базы знаний, индекса или конфигурации и всех пригодных путей восстановления; невосстановимость подтверждена

Условия проявления

1. **Предпосылка.** В рабочем AI-компоненте или его зависимости существует известная уязвимость.
2. **Триггер.** Уязвимая версия остаётся в эксплуатации и достижима для источника воздействия.
3. **Разрешающий пробел.** Факт наличия уязвимости в компоненте не выявлен либо обнаружение не отработано по процессу устранения.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-009 · Конвейер обучения и дообучения	121	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	122	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	123	58
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	124	67
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	125	84
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	388	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	389	118

Карточка	Связи	Стр.
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	390	127
AISAFE2-IMP-031 · Утрата датасета в результате атаки	391	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	392	141
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	647	419
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	648	577

Недостаток · AISAFE2-WEAK-048

Недостаточное реагирование на инциденты и аварийное восстановление

Краткое описание

AI-инциденты не имеют исполнимых сценариев локализации, восстановления и проверки безопасного возврата в работу.

Подробное описание

Недостаток активируется, когда требуется остановить агента, отозвать доступ, изолировать коннектор или данные, переключить модель либо восстановить проверенные версии модели, базы знаний и конфигурации из доверенной копии. Процесс остаётся неисполняемым без команд, владельцев, целевых состояний и регулярных учений. Даже описанный сценарий неисполняем, если реагирующей команде заранее не выданы технические полномочия на предусмотренные действия.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-031 · Готовность реагирования и восстановления

Прямые причинные воздействия

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Недостатки реагирования на инциденты и процесса аварийного восстановления могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что сценарии локализации и восстановления не отработаны и не обеспечены проверенными копиями, операторы импровизируют, и простой растягивается сверх директивных интервалов.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

AISAFE2-IMP-031 · Утрата датасета в результате атаки

После разрушительного воздействия атаки на точный датасет неотработанный процесс восстановления и отсутствие проверенных доверенных копий превращают повреждение в невосстановимую утрату.

Ограничение: Только если атака повредила точный датасет, пригодных доверенных копий или рецептов нет и проверенное восстановление не удалось; временный простой исключён

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

После разрушительного воздействия атаки на веса, базу знаний, индекс либо критичную конфигурацию неотработанное восстановление и отсутствие проверенных копий делают объект невосстановимым.

Ограничение: Только если атака повредила точный критичный артефакт, кроме датасета, пригодных версий или рецептов нет и проверенное восстановление не удалось

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление
> 2. Промежуточное воздействие: AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала
> 3. Итоговое воздействие: AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
>
> **Механизм переходов**
>
> 1. Инцидент требует отсутствующего подготовленного действия, полномочия или доверенной копии, и этот пробел в готовности к восстановлению продлевает техническую недоступность.
> 2. Продлённая недоступность превышает допустимый интервал именованного обязательного шага без обходного пути и вызывает наблюдаемую остановку, задержку либо ошибку процесса.
>
> **Условия применимости**
>
> - Фактический инцидент требует именно отсутствующего действия, полномочия или доверенной копии, и пробел в готовности к восстановлению продлевает недоступность.
> - Назван обязательный процесс, нет эффективного обходного пути, допустимый интервал превышен и отклонение наблюдается.
>
> **Исключения**
>
> - Инцидент, не связанный с пробелом восстановления, не поддерживает этот путь.
> - Успешное восстановление в интервале, доступный обходной путь или некритичный процесс исключают финальный переход.

Условия проявления

1. **Предпосылка.** Эксплуатация AI-системы зависит от способности команды остановить агента, отозвать доступы, изолировать зависимость или восстановить проверенную

версию.

2. **Триггер.** Наступает инцидент, требующий одного или нескольких таких действий.
3. **Разрешающий пробел.** Команды, полномочия, целевые состояния и проверенные копии заранее не подготовлены, учения не проводились и иных средств быстрой локализации и подтверждения безопасного возврата нет.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 32(1)(c)** · Диагностический признак сбоя
>
> Неисполняемый сценарий восстановления AI-системы может задержать возврат доступа к персональным данным после технического инцидента.
>
> **Граница связи:** Возврат сервиса в работу не доказывает восстановление самих персональных данных в необходимый срок; пригодные копии и испытания проверяются отдельно.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Обработка персональных данных зависит от затронутого AI-пути; для неё проверена цель восстановления.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	393	127
AISAFE2-IMP-031 · Утрата датасета в результате атаки	394	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	395	141
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	649	410
AISAFE2-MIT-065 · Защищённое сохранение материалов расследования	650	558

Недостаток · AISAFE2-WEAK-049

Слабые или отсутствующие механизмы установления доверия в A2A

Краткое описание

Межагентское сообщение принимается без достаточной проверки отправителя, целостности, адресата, задачи, свежести или повторного использования.

Подробное описание

Недостаток проявляется, когда поддельное, изменённое, ошибочно адресованное, просроченное или повторно воспроизведённое сообщение может повлиять на план либо действие. Пробел сохраняется, если проверка подменена общей инфраструктурной идентичностью, не различающей отдельных агентов и их полномочия.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-011 · Передача идентичности и A2A-доверие

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Слабое установление доверия в межагентном взаимодействии может привести к обходу ограничений и предусмотренной логики агента, потому что поддельное, изменённое, ошибочно адресованное или повторно воспроизведённое сообщение другого агента принимается как действующее поручение и меняет планы и действия системы.

AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)

Получатель принимает поддельное, изменённое или повторно воспроизведённое межагентское поручение за действительное и передаёт его инструментальному слою; инструмент или сторонний агент возвращает непредусмотренный результат либо выполняет действие вне контракта.

Ограничение: Только если такое поручение фактически принято и вызвало непредусмотренный результат или действие, достигшее следующего компонента либо внешней системы.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Слабое установление доверия в межагентном взаимодействии может привести к утечке чувствительных данных в недоверенный контур, потому что данные передаются агенту, идентичность и полномочия которого не подтверждены, то есть фактически неизвестному получателю за границей доверия.

Ограничение: Только если чувствительные данные фактически переданы агенту, чьи идентичность и полномочия не установлены либо не разрешены для этих данных

Условия проявления

1. **Предпосылка.** Агенты в рамках системы принимают сообщения от других агентов как поручения, влияющие на их планы и действия.
2. **Триггер.** Источник воздействия может влиять на работу системы через отправку вредоносных задач в межагентских сообщениях.
3. **Разрешающий пробел.** Тройка (отправитель, адресат, задача) не проверяется на уровне отдельных агентов и их полномочий.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI07** · Диагностический признак сбоя
>
> Приём межагентского сообщения без проверки отправителя, задачи и свежести соответствует одному из путей подмены коммуникации между агентами.
>
> **Граница связи:** Уязвимость границы не доказывает, что сообщение было подделано или что агент выполнил вредное действие.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агенты обмениваются сообщениями, для которых значимы отправитель, получатель, задача и защита от повтора.
Условие связи с карточкой	Рабочее сообщение A2A принимается без проверки отправителя, целостности, получателя, задачи, свежести либо повтора.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	126	61
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	127	82
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	396	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	397	104
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	398	130
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	651	429
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	652	497

Недостаток · AISAFE2-WEAK-050

Остаточные данные в общей инфраструктуре исполнения моделей

Краткое описание

Данные предыдущего процесса, задания или арендатора остаются доступными следующему потребителю общего ресурса исполнения моделей.

Подробное описание

Недостаток проявляется, когда ресурс передается другой задаче или арендатору до подтвержденной очистки либо изоляции. Общая облачная или управляемая платформа не считается безопасной по факту своего существования; граница требует подтверждений для фактической реализации и версии. К повторно используемому состоянию относятся память GPU и хоста, кэши, буферы, временные файлы и состояние среды выполнения.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-019 · Изоляция данных

Прямые причинные воздействия

AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами

Неочищенная память, кэши, буферы или временные файлы общей среды передают состояние предыдущего задания следующему изолированному контексту, что напрямую смешивает краткоживущие артефакты разных заданий или арендаторов.

Условия проявления

1. **Предпосылка.** Общая память GPU или хоста, кэш, буфер, временный файл либо переиспользуемая среда поочерёдно обслуживает разные процессы, задания или арендаторов и содержит остатки чувствительного состояния предыдущего потребителя.
2. **Триггер.** Ресурс передаётся следующему потребителю, который получает возможность прочитать прежнее состояние.
3. **Разрешающий пробел.** Перед повторным выделением ресурса очистка и изоляция остаточного состояния не подтверждаются.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Диагностический признак сбоя
>
> Остатки промпта, ответа или кэша одного арендатора в общем ресурсе могут стать доступны следующему потребителю.
>
> **Граница связи:** Нужна фактическая доступность чужого состояния; общее размещение само по себе утечку не доказывает.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	128	58
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	399	139
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	653	500

Недостаток · AISAFE2-WEAK-051

Выдача результата без завершения обязательных стадий вычислений

Краткое описание

Система выпускает штатный результат, хотя одна или несколько обязательных стадий вычислений пропущены либо завершились ошибкой, тайм-аутом или отменой, и неполное выполнение неотличимо от успешного.

Подробное описание

Пропуск стадии, повторное или перезаписанное состояние, ошибка, тайм-аут, отмена либо неопределённая применимость делают анализ неполным, но агрегатор не отличает такое выполнение от полного успешного завершения плана.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-030 · Безопасная обработка ошибок и деградация

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Выдача результата без завершения обязательных стадий вычислений может привести к обходу ограничений и предусмотренной логики агента, потому что пропуск или сбой обязательной стадии неотличим от успешного завершения, и агент пропускает обязательный шаг (например, согласование), выдавая неполное выполнение за проверенное.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-051 · Выдача результата без завершения обязательных стадий вычислений
> 2. Промежуточное воздействие: AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core
> 3. Промежуточное состояние: Неполный результат ошибочно маркируется как успешный и принимается последующим шагом как завершённый
> 4. Итоговое воздействие: AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
>
> **Механизм переходов**
>
> 1. Неполный план пропускает либо не завершает обязательную стадию, поэтому агент фактически выдаёт неполный результат операции.
> 2. Неполный результат помечается как успешный, а именованный последующий шаг доверяет этому статусу и принимает его как завершённый.
> 3. Принятый как завершённый неполный результат вызывает наблюдаемое отклонение обязательной последующей операции.
>
> **Условия применимости**
>
> - Обязательная стадия фактически пропущена или завершилась неуспешно, но результат ошибочно помечен как успех.
> - Именованный обязательный последующий шаг доверяет статусу и вследствие этого фактически отклоняется от требуемого процесса.
>
> **Исключения**
>
> - Явный статус частичного выполнения или ошибки, безопасный отказ при сбое либо отклонение результата последующим контуром исключают путь.
> - Неполнота без наблюдаемого процессного последствия не является I026.

Условия проявления

1. **Предпосылка.** Итог операции складывается из нескольких обязательных стадий вычислений или проверок.
2. **Триггер.** Одна из стадий пропущена либо завершилась ошибкой, тайм-аутом или отменой.
3. **Разрешающий пробел.** Точка агрегации не сверяет конечное состояние всех стадий, и получатель принимает обычный статус за подтверждение успешной проверки.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	129	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	130	77
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	400	101
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	654	507

Недостаток · AISAFE2-WEAK-052

Неограниченные внутренние циклы и усиление нагрузки

Краткое описание

Один допустимый запрос способен породить неограниченные агентные циклы, повторные попытки или разветвление и занять несоразмерный объём ресурсов до результата.

Подробное описание

Недостаток проявляется в рекурсии, самопроверке без условия остановки, веерных вызовах инструментов, повторных попытках, расширении контекста или обработке данных, когда внутреннюю траекторию вычислений нельзя заранее ограничить и проверить. Отсутствующие пределы охватывают не только число шагов, но и глубину вложенности, время, параллелизм и совокупный бюджет одной задачи; лимит внешних запросов не ограничивает это внутреннее усиление.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-021 · Ограничение частоты, ресурсов и стоимости

Прямые причинные воздействия

AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям

Неограниченные внутренние циклы и усиление нагрузки могут привести к неограниченному потреблению токенов, потому что внутренние шаги планирования, повторы и самопроверка не ограничены бюджетом задачи, и расход внутри одного запроса растёт незаметно для лимитов на число запросов.

Ограничение: Только для одной траектории обработки с повторными обращениями к языковой модели, в которой не срабатывает утверждённый токенный предел.

AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов

Неограниченные внутренние циклы и усиление нагрузки могут привести к незапланированным расходам на платные инструменты, потому что повторные попытки, веерные вызовы инструментов и циклы агента множат обращения к тарифицируемым сервисам без ограничения числа шагов и бюджета задачи.

Ограничение: Только когда цикл или действие содержит тарифицируемый инструмент либо внешний сервис и создаёт подтверждённое начисление вне утверждённого бюджета.

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Неограниченные внутренние циклы и усиление нагрузки могут привести к чрезмерному потреблению вычислительных ресурсов системы, потому что один запрос порождает неограниченную траекторию вычислений — рекурсию, самопроверку, веерные вызовы, разрастание контекста, — занимая несоразмерный объём ресурсов.

AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов

Если внутренние шаги одного запроса повторяются без достижения нового состояния, отсутствие жёсткого предела итераций, глубины вложенности или времени не останавливает цикл и позволяет ему продолжаться до внешнего прерывания.

Ограничение: Только при наблюдаемом повторении шагов без штатного завершения; само отсутствие лимита не доказывает зацикливание.

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Неограниченные внутренние циклы и усиление нагрузки могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что неограниченные внутренние циклы удерживают ресурсы и квоты вплоть до зацикливания системы и отказа в обслуживании.

Условия проявления

1. **Предпосылка.** Обработка одного запроса допускает рекурсию, самопроверку, веерные вызовы инструментов, повторные попытки или разрастание контекста.
2. **Триггер.** Допустимый запрос запускает такую внутреннюю траекторию, а модель продолжает выбирать новые шаги до формального достижения цели.
3. **Разрешающий пробел.** Число шагов, вложенность, время, параллелизм и суммарный бюджет задачи жёстко не ограничены, а остановка зависит только от управляемого содержимым решения модели.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.1.2 · Диагностический признак сбоя

Неограниченные внутренние циклы показывают путь, на котором одна операция превышает допустимую глубину и расход.

Граница связи: Карточка описывает отсутствие пределов; их значения и срабатывание во время исполнения нужно проверять в системе.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает лимиты глубины рекурсии, токенов и расходов на одно выполнение.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	131	42
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	132	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	133	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	134	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	135	75
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	401	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	402	121
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	403	123
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	404	126
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	405	127
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	655	333

Недостаток · AISAFE2-WEAK-053

Небезопасный вывод AI-компонента из эксплуатации

Краткое описание

После вывода AI-компонента из эксплуатации сохраняются его доступы, маршруты, интеграции, учётные данные или управляемые копии данных и производных артефактов.

Подробное описание

Общий объект не удаляется целиком, если обслуживает другие системы: прекращаются только доступ и возможности выводимого компонента. Этот недостаток не подразумевает физическое уничтожение данных вне контролируемой области.

Область применения

- Общая безопасность

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-037 · Закрытие прав и возможностей жизненного цикла

Прямые причинные воздействия

AISAFE2-IMP-012 · Кража модели

Небезопасный вывод AI-компонента из эксплуатации может привести к краже модели, потому что после вывода компонента из эксплуатации забытые копии моделей остаются доступными для хищения в несопровождаемом и неучтённом контуре.

Ограничение: Только если из оставшейся после вывода из эксплуатации копии фактически получены веса или связанные артефакты без разрешения

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Небезопасный вывод AI-компонента из эксплуатации может привести к несанкционированному чтению или изменению хранимых данных, потому что у выведенного из эксплуатации компонента сохраняются доступы, учётные данные и данные, и связанные системы продолжают предоставлять доступ от его имени.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Небезопасный вывод AI-компонента из эксплуатации может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что выведенный компонент остаётся неучтённой точкой входа: связанные системы продолжают принимать вызовы и предоставлять доступ от его имени.

Ограничение: Только если оставшийся после вывода путь позволил вызвать инфраструктурную функцию от имени снятого компонента без полномочий

Условия проявления

1. **Предпосылка.** Принято решение о выводе AI-компонента из эксплуатации, но вывод выполняется без надежного чек-листа.
2. **Триггер.** Связанные системы продолжают принимать вызовы и предоставлять доступ от его имени, а сохранившиеся объекты никем не сопровождаются.
3. **Разрешающий пробел.** Его доступы, учётные данные, интеграции, данные и копии артефактов сохраняются.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

> **GOVERN 1.7** · Диагностический признак сбоя
>
> Карточка показывает, какие оставленные доступы и копии могут сделать вывод AI-компонента из эксплуатации небезопасным.
>
> **Граница связи:** Описание пробела не подтверждает наличие процедуры вывода, распределение ответственности и проверку всех затронутых систем.

Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	В организации выбран процесс поэтапного вывода AI-системы из эксплуатации в рамках добровольного NIST AI RMF.
Условие связи с карточкой	AI-компонент выводится из эксплуатации или его использование поэтапно прекращается.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	136	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	137	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	138	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	139	44
AISAFE2-COMP-007 · Долговременная память	140	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	141	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	142	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	143	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	144	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	145	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	146	64
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	147	67

Карточка	Связи	Стр.
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	148	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	149	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	150	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	151	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	152	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	153	82
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	154	84
AISAFE2-COMP-023 · Подсистема формирования базы знаний	155	86
AISAFE2-IMP-012 · Кража модели	406	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	407	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	408	118
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	656	509

Недостаток · AISAFE2-WEAK-054

Стохастическая недостоверность вывода генеративной модели

Краткое описание

Вероятностная генерация способна давать убедительный, но фактически неверный или невоспроизводимый результат при штатной работе модели.

Подробное описание

Генеративная модель строит ответ по выученным закономерностям, а выбранный режим декодирования не гарантирует фактическую истинность. Поэтому галлюцинация, вымышленный источник или ошибочная рекомендация возможны без атаки и технического отказа.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Стохастическая недостоверность вывода генеративной модели может привести к выполнению пользователем предложенного системой вредоносного действия, потому что модель формирует правдоподобный, но не обязательно верный или безопасный результат при

штатной работе, и ошибочная рекомендация (несуществующая функция, вредная команда) появляется без атаки и отказа.

Ограничение: Только если ошибочная рекомендация была прямо предложена как действие, пользователь её выполнил и понёс вред

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Стохастическая недостоверность вывода генеративной модели может привести к ущербу от решения, принятого на основе результата AI-системы, потому что модель выдаёт правдоподобный текст, а не проверенный факт, и решение, существенно опирающееся на такой ответ (сделка, отмена заказа, утверждение расчёта), оборачивается ущербом.

Ограничение: Только если недостоверный ответ стал существенным основанием исполненного решения и вреда; сама генерация ошибочного ответа недостаточна

AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей

Генеративная модель может при штатной работе выдать правдоподобный, но неверный ответ, вымышленный источник или ошибочную рекомендацию; детерминированное декодирование само по себе не обеспечивает истинность.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели
> 2. Промежуточное воздействие: AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей
> 3. Промежуточное состояние: Ложный результат принимается без эффективной проверки именованной обязательной операцией
> 4. Итоговое воздействие: AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
>
> **Механизм переходов**
>
> 1. Стохастически недостоверный вывод фактически содержит проверяемую ошибку или выдуманное утверждение вне допустимой границы качества.
> 2. Эффективная проверка отсутствует, поэтому именованная обязательная операция принимает ложный результат как пригодный вход.
> 3. Использование принятого ложного входа вызывает зарегистрированное отклонение обязательной операции от требуемого результата.
>
> **Условия применимости**
>
> - Установлен фактический ложный или недействительный вывод точной версии, а эффективная проверка его не отклоняет.
> - Точная обязательная операция потребляет вывод и фиксируется наблюдаемое отклонение процесса.
>
> **Исключения**
>
> - Одна вероятность галлюцинации или результат, который был обнаружен и отклонён, не образуют путь.

- Совещательное использование без обязательной операции относится к отдельной ветви воздействия на решение человека.

Условия проявления

1. **Предпосылка.** Сценарий опирается на достоверность или воспроизводимость вывода: сведения, расчёты, рекомендации, код.
2. **Триггер.** Получатель — человек или процесс — применяет правдоподобный ответ как достоверный.
3. **Разрешающий пробел.** Ответ формируется вероятностной генерацией без обязательного основания и проверки истинности.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM07:2026 · Диагностический признак сбоя

Вероятностная генерация может выдать убедительное, но фактически неверное утверждение даже при штатной работе модели.

Граница связи: Ложный ответ становится риском LLM07 при доверии к нему и последующем решении; карточка не доказывает такой исход.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется как фактическое основание ответа, рекомендации или решения.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	156	37
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	409	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	410	93
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	411	100
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	657	331
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	658	488

Недостаток · AISAFE2-WEAK-055

Возложение бизнес-логики и обязательных ограничений на модель

Краткое описание

Решение модели непосредственно становится действием системы, хотя обязательные правила не проверяются независимым детерминированным механизмом.

Подробное описание

Лимиты, разрешённые операции, порядок шагов, правила раскрытия данных или согласования существуют только как текст для модели. Вероятностное следование этому тексту позволяет обойти инвариант убеждением, переформулированием или инъекцией.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-008 · Системные инструкции и детерминированные контроли

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Возложение бизнес-логики и обязательных ограничений на модель может привести к обходу ограничений и предусмотренной логики агента, потому что обязательные правила существуют только как текст для модели, её решение напрямую становится действием системы, и убеждение, переформулирование или инъекция меняют поведение агента.

Условия проявления

1. **Предпосылка.** Для системного действия существуют обязательные лимиты, допустимые операции, порядок шагов или правила раскрытия данных.
2. **Триггер.** Решение модели напрямую порождает действие системы.
3. **Разрешающий пробел.** Обязательные правила заданы только как инструкции модели, а независимый детерминированный механизм не проверяет их соблюдение до исполнения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.5.3 · Диагностический признак сбоя

Если правило доступа существует только в инструкции модели, его выполнение зависит от её ответа и не обеспечивается независимой логикой приложения.

Граница связи: Связь относится к правилам доступа; прочие бизнес-ограничения карты шире этого контроля AISVS.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения требует решения о доступе в логике приложения или независимом механизме политики.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	157	42
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	158	72
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	412	101
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	659	429
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	660	519
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	661	561

Недостаток · AISAFE2-WEAK-056

Недостаточная валидация аргументов вызова инструментов

Краткое описание

Аргументы, сформированные моделью или агентом, поступают исполнителю без обязательной проверки структуры, типа, диапазона и допустимых значений.

Подробное описание

Даже разрешённый инструмент становится опасной границей, если исполнитель принимает произвольный объект, путь, адрес, команду или параметр и не отклоняет значения вне заранее установленного контракта. Пробел сохраняется, если контракт проверяется только на выходе

модели или агента, но не применяется принудительно непосредственно перед исполнителем инструмента.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-014 · Контракты вызова инструментов

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Недостаточная валидация аргументов вызова инструментов может привести к обходу ограничений и предусмотренной логики агента, потому что аргументы вызова, собранные моделью из недоверенного текста, не проверяются перед исполнением, и подмена объектов операции (идентификаторов, адресатов, сумм) уводит выполнение за предусмотренные рамки.

AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)

Недостаточная валидация аргументов вызова инструментов может привести к неожидаемому поведению слоя инструментов, потому что через непроверенные аргументы вызова, собранные из недоверенного текста, даже исправный инструмент можно заставить выполнить непредусмотренные действия за пределами контракта.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостаточная валидация аргументов вызова инструментов может привести к несанкционированному чтению или изменению хранимых данных, потому что через непроверенные аргументы в нижележащие системы проходят инъекции запросов к базам данных и обход файловых путей, которые исполняются с правами инструмента и открывают чтение и изменение данных вне полномочий инициатора.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Недостаточная валидация аргументов вызова инструментов может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что непроверенные аргументы, собранные моделью из недоверенного текста, реализуют инъекции в нижележащие системы (SQL, команды ОС, обход путей) вплоть до компрометации инструмента и смежных компонентов.

Ограничение: Только если непроверенный аргумент привёл к инъекции или обходу, давшему вызов функции вне полномочий инициатора

Условия проявления

1. **Предпосылка.** Аргументы вызова инструмента формируются моделью или агентом из недоверенного контекста.
2. **Триггер.** Вызов исполняется на основании одного лишь выбора агента, и аргументы достигают смежных систем.

3. **Разрешающий пробел.** Исполнитель принимает объект, путь, адрес, команду или параметр без проверки структуры, типа, диапазона и допустимых значений.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	159	44
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	413	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	414	104
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	415	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	416	118
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	662	316
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	663	375

Недостаток · AISAFE2-WEAK-057

Отсутствие обязательного подтверждения чувствительных действий

Краткое описание

Чувствительное, необратимое или внешне значимое действие, для которого правила системы требуют решения уполномоченного человека, выполняется без его явного решения по точным параметрам действия.

Подробное описание

Общее согласие на использование агента или предварительное предупреждение не подтверждает конкретный платёж, удаление, публикацию либо передачу данных. Если для действия требуется решение человека, до исполнения ему должно быть показано неизменяемое предложение с точными параметрами и существенными последствиями.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-025 · Подтверждение и пропускная способность HITL

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Отсутствие обязательного подтверждения чувствительных действий может привести к обходу ограничений и предусмотренной логики агента, потому что предусмотренный контур согласования высокорисковых операций фактически не исполняется: необратимые и внешне значимые действия совершаются без явного решения уполномоченного субъекта по их точным параметрам, и логика обязательных подтверждений обходится.

AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов

Отсутствие обязательного подтверждения чувствительных действий может привести к незапланированным расходам на платные инструменты, потому что дорогостоящие операции не требуют явного подтверждения уполномоченным субъектом, и платные вызовы совершаются без осознанного решения по их параметрам.

Ограничение: Только если тарифицируемый вызов требовал обязательного подтверждения, прошёл без него и создал начисление сверх утверждённого бюджета

AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы

Отсутствие обязательного подтверждения чувствительных действий позволяет агенту совершать дорогие операции от имени пользователя без его согласия, что в тарифицируемом контексте напрямую конвертируется в неожиданные для пользователя расходы.

Ограничение: Только если стоимость операции несёт пользователь и дорогой вызов прошёл без обязательного подтверждения; вне пользовательского тарифа этот недостаток не создаёт его расхода

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Отсутствие обязательного подтверждения чувствительных действий может привести к утечке чувствительных данных в недоверенный контур, потому что действия передачи данных вовне — публикация, отправка, выгрузка — исполняются от имени пользователя без явного решения по точным параметрам операции, и ошибочная или навязанная атакующим передача пересекает границу доверия раньше, чем пользователь о ней узнаёт.

Ограничение: Только если требующая подтверждения передача чувствительных данных прошла без решения уполномоченного пользователя к неразрешённому адресату

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий
> 2. Промежуточное воздействие: AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core
> 3. Промежуточное состояние: Денежная операция с точными параметрами фактически исполнена без требуемого подтверждения уполномоченного пользователя
> 4. Итоговое воздействие: AISAFE2-IMP-025 · Прямые финансовые потери
>
> **Механизм переходов**
>
> 1. Отсутствие обязательного подтверждения позволяет агенту сформировать и передать на исполнение денежное действие без авторизации точной суммы и получателя.

2. Исполнитель принимает действие и фактически проводит денежную операцию от имени пользователя без требуемого подтверждения её точных параметров.
3. После исчерпания процедур возврата или оспаривания платежа точная чистая сумма операции остаётся атрибутированным невозвратным денежным убытком.

Условия применимости

- Агент имеет платёжное полномочие, а политика требует подтверждения точной суммы и получателя, которого фактически не было.
- Операция исполнена, установлена точная чистая сумма, а возможности возврата или оспаривания платежа отсутствуют либо исчерпаны.

Исключения

- Подготовленное или отклонённое действие, неденежная операция и подтверждённая операция не образуют путь.
- Возвращённую сумму и то же начисление I018 либо I020 нельзя повторно учитывать как отдельный I025.

Условия проявления

1. **Предпосылка.** Агент способен выполнить чувствительное, необратимое или внешне значимое действие, для которого правила системы требуют решения уполномоченного человека, например платёж, удаление, публикацию или передачу данных.
2. **Триггер.** Исполнитель получает подготовленную чувствительную операцию и трактует общее согласие на работу агента либо предварительное предупреждение как разрешение на её выполнение.
3. **Разрешающий пробел.** Для действия, требующего решения человека, ему не предъявляется неизменяемое предложение с точными параметрами и последствиями операции и не запрашивается явное решение до исполнения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.2.1 · Диагностический признак сбоя

Действие без подтверждения его точных параметров показывает нарушение границы человеческого допуска перед исполнением.

Граница связи: Применимость зависит от класса действия; все пути исполнения требуют отдельной проверки.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверенное подтверждение человека до привилегированного или необратимого действия.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	160	35
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	161	65
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	417	101
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	418	121
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	419	125
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	420	130
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	664	389

Недостаток · AISAFE2-WEAK-058

Вызов инструментов без учёта идентичности и полномочий инициатора

Краткое описание

Идентичность пользователя или сессии и её полномочия не передаются либо не проверяются целевым инструментом, который действует с более широкими сервисными правами.

Подробное описание

Агент принимает запрос менее привилегированного субъекта, но вызывает инструмент от общей учётной записи. Без проверки действия и объекта относительно исходного субъекта его запрос наследует полномочия агента или сервиса. Контекст инициатора теряется по цепочке «сессия — ядро агента — инструмент», если целевой инструмент авторизует только общую техническую учётную запись.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-011 · Передача идентичности и A2A-доверие

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Вызов инструментов без учёта идентичности и полномочий инициатора может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что инструмент действует от общей служебной учётной записи, не проверяя идентичность и полномочия инициатора, и возвращает данные, на которые у запрашивающего субъекта нет прав.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных

(в том числе за счёт повышения привилегий)

Вызов инструментов без учёта идентичности и полномочий инициатора может привести к несанкционированному чтению или изменению хранимых данных, потому что запрос менее привилегированного субъекта наследует широкие права сервисной учётной записи инструмента и читает или меняет чужие хранимые данные.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Вызов инструментов без учёта идентичности и полномочий инициатора может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что инструмент не проверяет идентичность и полномочия инициатора и выполняет функции с широкими сервисными правами по запросу менее привилегированного субъекта.

Условия проявления

1. **Предпосылка.** Инструменты вызываются от общей сервисной учётной записи с широкими правами.
2. **Триггер.** Запрос поступает от субъекта, полномочия которого меньше полномочий сервисной записи.
3. **Разрешающий пробел.** Идентичность и полномочия инициатора не передаются по цепочке «сессия — ядро агента — инструмент» либо не проверяются на конкретном действии и объекте.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	162	44
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	163	61
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	421	106
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	422	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	423	118
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	665	429
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	666	434

Недостаток · AISAFE2-WEAK-059

Недостаточная изоляция арендаторов при дообучении модели

Краткое описание

Данные разных арендаторов объединяются при дообучении так, что их принадлежность и допустимые границы использования не сохраняются в результирующей модели.

Подробное описание

Дообучение выполняется на объединённых данных нескольких арендаторов, а результат предоставляется им как общий сервис: принадлежность данных и допустимые границы их использования не сохраняются в параметрах модели. Разграничение, установленное для исходных данных, не переносится на обучающий сигнал, задания и производные артефакты: модель не содержит сведений о принадлежности выученного и не может ограничивать выдачу по арендатору. В результате влияние записей одного арендатора проявляется в поведении, доступном другому, и через штатный интерфейс модели могут извлекаться чужие данные.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-002 · Поведение при обучении и дообучении
- AISAFE2-WC-019 · Изоляция данных

Прямые причинные воздействия

AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели

Недостаточная изоляция арендаторов при дообучении модели может привести к раскрытию чувствительных данных из обучающего датасета, потому что при дообучении данные разных арендаторов смешиваются без сохранения принадлежности в параметрах модели, и через штатный интерфейс общего сервиса извлекаются чужие данные.

Ограничение: Только если данные другого арендатора фактически извлечены или восстановлены через общую дообученную модель; совместное обучение без такого результата недостаточно

Условия проявления

1. **Предпосылка.** Дообучение выполняется на объединённых данных нескольких арендаторов, а результат предоставляется им как общий сервис.
2. **Триггер.** Модель отвечает арендаторам без ограничения выдачи по принадлежности выученного.
3. **Разрешающий пробел.** Принадлежность данных и границы их использования не переносятся на обучающий сигнал, задания и производные артефакты.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	164	37
AISAFE2-COMP-008 · Данные для подсистемы обучения	165	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	166	54
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	424	113
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	667	473
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	668	478

Недостаток · AISAFE2-WEAK-060

Отсутствие фильтрации сообщений об ошибках на внешней границе

Краткое описание

Внутреннее представление сбоя пересекает пользовательскую границу без преобразования и раскрывает технические сведения о реализации системы.

Подробное описание

Фильтрация нужна на каждой внешней границе AI-системы: API, вызовах инструментов, механизме извлечения данных и ответе модели, которая может пересказать служебную ошибку. Передача внутреннего представления сбоя внешнему получателю раскрывает сведения о реализации системы и облегчает разведку и подготовку целевых атак.

Область применения

- Общая безопасность

Происхождение

- Реализация
- Проектирование

Классы недостатка

- AISAFE2-WC-030 · Безопасная обработка ошибок и деградация

Прямые причинные воздействия

AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных

Нефильтрованное внутреннее сообщение об ошибке, содержащее закрытые сведения о стеке, компонентах, конфигурации, алгоритмах или источниках данных, пересекает внешнюю границу и становится доступно неуполномоченному получателю.

Условия проявления

1. **Предпосылка.** Компонент модели, извлечения, инструмента или API имеет внешнюю границу, через которую может возвращаться результат обработки ошибки.
2. **Триггер.** В компоненте возникает сбой с детализированным внутренним представлением: стеком вызовов, именами ресурсов, конфигурацией или идентификаторами.
3. **Разрешающий пробел.** Обработчик не преобразует внутреннюю ошибку в ограниченное внешнее сообщение, поэтому API, интерфейс или пересказ модели передаёт её внешнему получателю.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	167	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	168	37
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	169	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	170	48
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	171	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	172	72
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	425	114
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	669	320

Недостаток · AISAFE2-WEAK-061

Избыточно открытая внутренняя телеметрия

Краткое описание

Трассировки, метрики и служебные поля внутренних компонентов становятся доступны внешнему получателю через ответ, интерфейс, открытый сетевой адрес или экспорт аналитики.

Подробное описание

Диагностические события могут содержать идентификаторы, параметры запросов, структуру вызовов, значения стоимости и иные сведения, предназначенные только для внутреннего наблюдения и расследования.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-033 · Наблюдаемость, обнаружение и доказательства

Прямые причинные воздействия

AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных

Избыточно открытая внутренняя телеметрия может привести к раскрытию сведений о технологиях, алгоритмах и источниках данных системы, потому что трассировки, метрики и служебные заголовки доступны внешнему получателю через ответы, интерфейс и экспорт аналитики и раскрывают сведения об устройстве и работе системы.

Ограничение: Только при подтверждённом неразрешённом получении закрытой информации о технологиях, архитектуре, алгоритмах или источниках данных.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Избыточно открытая внутренняя телеметрия может привести к утечке чувствительных данных в недоверенный контур, потому что телеметрия и служебные данные экспортируются во внешнюю аналитику и попадают в ответы и заголовки, вынося за контур сведения и чувствительное содержимое запросов.

Ограничение: Только если телеметрия с чувствительным содержимым запроса была экспортирована или выдана получателю без разрешения

Условия проявления

1. **Предпосылка.** Диагностические события содержат сведения о реализации системы или содержимое обрабатываемых запросов.
2. **Триггер.** Телеметрия достигает внешних получателей: через ответы и заголовки, элементы интерфейса, открытые сетевые сервисы, экспорт в аналитику.
3. **Разрешающий пробел.** Состав телеметрии и доступ к ней не ограничены по получателю.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM02:2026** · Диагностический признак сбоя
>
> Открытый экспорт трассировок и служебных полей может раскрыть чувствительные сведения внешнему получателю.
>
> **Граница связи:** Телеметрия не всегда содержит закрытые сведения; необходимо установить её состав и получателя.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-013 · Телеметрия и обнаружение	173	64
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	426	114
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	427	130
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	670	519

Недостаток · AISAFE2-WEAK-062

Отсутствие сквозной координации ресурсных и стоимостных бюджетов

Краткое описание

Сквозной учёт и принудительное ограничение потребления между всеми участниками одной операции отсутствуют: каждый компонент ограничен лишь собственными лимитами, и составная операция в сумме превышает допустимое потребление.

Подробное описание

Каждый компонент может соблюдать собственный предел, но цепочка моделей, агентов и инструментов в сумме превышает допустимое потребление. Общий счётчик не резервируется, не передаётся по вызовам либо не останавливает работу при исчерпании.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-021 · Ограничение частоты, ресурсов и стоимости

Прямые причинные воздействия

AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям

Если число внутренних обращений к модели не ограничено, отсутствие сквозного токенного бюджета позволяет составной операции продолжать расходовать токены без верхней границы, хотя каждый отдельный вызов соблюдает локальный лимит.

Ограничение: Только если у запроса, сессии или цепочки нет другого действующего конечного предела общего токенного расхода; конечная цепочка с локальными лимитами может превысить бюджет, но не образует неограниченное потребление.

AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов

Отсутствие сквозного учёта стоимости между участниками одной операции позволяет каждому компоненту соблюдать локальный лимит, тогда как совокупные вызовы платных инструментов превышают утверждённый бюджет и создают незапланированные начисления.

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Отсутствие общего ресурсного бюджета для составной операции позволяет отдельным компонентам продолжать вычисления в пределах локальных лимитов, тогда как их совокупное потребление вычислений, памяти и сети превышает утверждённый предел среды.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов
> 2. Промежуточное состояние: Совокупное измеренное потребление одного класса ресурса в составной операции превышает утверждённый общий бюджет при прохождении локальных лимитов
> 3. Промежуточное состояние: Избыточное потребление атрибутировано операции и создаёт подтверждённый счёт, ремонт либо замену для владельца
> 4. Итоговое воздействие: AISAFE2-IMP-025 · Прямые финансовые потери
>
> **Механизм переходов**
>
> 1. Разрозненные локальные лимиты пропускают каждый шаг, но отсутствие сквозного бюджета позволяет одной составной операции превысить утверждённый общий порог точного ресурса.
> 2. Метринг связывает совокупный перерасход с одной операцией и владельцем, после чего возникает точный счёт либо подтверждённая стоимость ремонта или замены.
> 3. После финансового урегулирования точная атрибутированная сумма остаётся невозвратным чистым денежным убытком владельца.
>
> **Условия применимости**
>
> - Определены один класс ресурса и составная операция, утверждён общий бюджет, а при прохождении локальных лимитов зафиксировано фактическое совокупное превышение.
> - Перерасход атрибутирован владельцу и операции, а точная сумма после возвратов остаётся невозвратной.
>
> **Исключения**
>
> - Теоретическая неограниченность, потребление без убытка и плановый утверждённый расход не образуют путь.
> - Показатели I017, I018 и I019 нельзя складывать с I025 как отдельные денежные потери одного события.

Условия проявления

1. **Предпосылка.** Операция составная: цепочка из моделей, агентов и инструментов.
2. **Триггер.** Каждый компонент продолжает работу в пределах лишь собственного лимита, и совокупное потребление не контролируется.
3. **Разрешающий пробел.** Общий бюджет времени, вычислений и стоимости не передаётся по цепочке вызовов и не резервируется.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.1.2 · Диагностический признак сбоя

Разрозненные лимиты компонентов позволяют одной агентной операции превысить общий бюджет времени или стоимости.

Граница связи: Карта не описывает конкретный предел глубины и токенов одной операции.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает лимиты глубины рекурсии, токенов и расходов на одно выполнение.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM06:2026 · Диагностический признак сбоя

Сумма вызовов моделей и инструментов может превысить допустимую стоимость, хотя каждый компонент соблюдает свой локальный предел.

Граница связи: Связь касается составной операции; локальные лимиты всё ещё могут ограничивать отдельные этапы, а фактический перерасход надо измерить.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	174	42
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	175	58
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	176	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	177	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	178	79
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	428	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	429	121

Карточка	Связи	Стр.
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	430	123
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	671	333

Недостаток · AISAFE2-WEAK-063

Недостаточные пользовательские средства управления расходами и прозрачность тарификации

Краткое описание

Плательщик не может заранее ограничить и оценить расходы, увидеть их состав, получить предупреждение или подтвердить дорогостоящее действие.

Подробное описание

Тарифицируемая операция выполняется без пользовательского бюджета, понятной оценки, текущего потребления, порога уведомления и автоматической остановки. После списания невозможно связать стоимость с запросом, этапом или фактическим результатом.

Область применения

- Общая безопасность

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-022 · Пользовательский контроль тарифицируемых операций

Прямые причинные воздействия

AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы

Если пользователь оплачивает потребление, отсутствие доступного ему бюджета, оценки и уведомления может привести к счёту, существенно превышающему показанную или разумно ожидаемую стоимость.

Ограничение: Только при оплате потребления пользователем и подтверждённом существенном превышении ожидаемой стоимости; само увеличение тарифицируемого объёма не достаточно.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-063 · Недостаточные пользовательские средства управления расходами и прозрачность тарификации

2. Промежуточное воздействие: AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы
3. Промежуточное состояние: Неожиданное начисление окончательно списано и остаётся невозвратным после доступных процедур возврата и оспаривания платежа
4. Итоговое воздействие: AISAFE2-IMP-025 · Прямые финансовые потери

Механизм переходов

1. Отсутствие пользовательского лимита, оценки или подтверждения материально позволяет фактическому начислению превысить заранее показанную либо одобренную сумму.
2. Неожиданное начисление списывается с плательщика и после применимых процедур возврата и оспаривания платежа остаётся окончательным.
3. Окончательная точная чистая сумма образует атрибутированный невозвратный денежный убыток плательщика.

Условия применимости

- Потребитель является плательщиком, а фактическое начисление превышает показанную или одобренную сумму из-за отсутствующего средства контроля.
- Установлена точная чистая сумма, которая не возвращена после применимых процедур возврата и оспаривания платежа.

Исключения

- Прозрачно показанное и принятое начисление, расходы владельца сервиса и полностью возвращённая сумма исключаются.
- Одна и та же сумма до возвратов учитывается один раз и не дублируется одновременно как I020 и отдельный I025.

Условия проявления

1. **Предпосылка.** Пользователь оплачивает сервис по фактическому потреблению.
2. **Триггер.** Запускается дорогостоящая операция: легитимно, по ошибке или по инициативе злоумышленника.
3. **Разрешающий пробел.** Предварительная оценка стоимости, собственные лимиты, пороговые уведомления, подтверждение дорогих операций и автоматическая остановка пользователю недоступны, а списания не отслеживаются до конкретных шагов.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Диагностический признак сбоя
>
> Без пользовательского предела и оценки стоимости дорогая операция может продолжить списания сверх суммы, которую плательщик ожидал.
>
> **Граница связи:** Связь относится к оплате по фактическому потреблению и отсутствию принудительного предела; одна лишь непонятная тарификация не доказывает неконтролируемый расход.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	179	35
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	180	72
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	431	125
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	672	333

Недостаток · AISAFE2-WEAK-064

Неизолированные отказы внешних зависимостей

Краткое описание

Отказ или замедление внешней модели, инструмента, хранилища либо провайдера распространяется на AI-систему и исчерпывает её собственные ресурсы.

Подробное описание

Отсутствующие или неверные тайм-ауты, ограничения повторов, раздельные пулы, предохранитель вызовов и проверенное резервное состояние превращают локальную проблему зависимости в каскад задержек и отказов. Безусловные повторные попытки создают лавину обращений к уже деградировавшей зависимости, удерживают собственные ресурсы и могут мешать её восстановлению.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-030 · Безопасная обработка ошибок и деградация

Прямые причинные воздействия

AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов

Неизолированные отказы внешних зависимостей могут привести к незапланированным расходам на платные инструменты, потому что безусловные повторы при отказах и замедлениях внешних зависимостей порождают лавину дополнительных тарифицируемых обращений к платным сервисам.

Ограничение: Только когда цикл или действие содержит тарифицируемый инструмент либо внешний сервис и создаёт подтверждённое начисление вне утверждённого бюджета.

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Неизолированные отказы внешних зависимостей не компенсируются переключением и планом обработки недоступности, что позволяет отказу или деградации одного звена распространиться на весь цикл обработки запроса и привести к недоступности AI-системы.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей
> 2. Промежуточное воздействие: AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала
> 3. Итоговое воздействие: AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
>
> **Механизм переходов**
>
> 1. Фактический отказ внешней зависимости без эффективных ограничений времени ожидания, изоляции отказов, увеличения интервалов между повторными попытками или переключения на резерв распространяется на техническую доступность AI-системы.
> 2. Распространившийся отказ превышает допустимый интервал именованной обязательной операции без обходного пути и вызывает наблюдаемое отклонение процесса.
>
> **Условия применимости**
>
> - Внешняя зависимость фактически отказала, а отсутствие конкретного защитного механизма позволило отказу распространиться.
> - Именованная обязательная операция зависит от функции, не имеет обходного пути, превышает допустимый интервал и фактически отклоняется.
>
> **Исключения**
>
> - Изолированный или краткий отказ в допустимом интервале не образует финальный переход.

- Эффективный обходной путь либо техническая недоступность без воздействия на обязательный процесс исключают I026.

Условия проявления

1. **Предпосылка.** Обработка зависит от внешней модели, инструмента, хранилища или провайдера сервиса.
2. **Триггер.** Зависимость отказывает или замедляется.
3. **Разрешающий пробел.** В точке подключения не ограничены время ожидания, повторы, параллелизм и пул ресурсов, а безопасное переключение не предусмотрено.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI08 · Диагностический признак сбоя

Неизолированный отказ внешней зависимости может запустить каскад, если затронет несколько связанных агентных операций.

Граница связи: Обычный отказ одной AI-системы ещё не ASI08; требуется показать распространение и усиление между агентами или процессами.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Взаимодействующие агенты и зависимые операции могут усилить исходный сбой или повтор.
Условие связи с карточкой	Сбой внешней зависимости усиливается через совместную работу агентов или зависимые операции; эффект измерим.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	181	44
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	182	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	183	72
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	432	121
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	433	127
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	673	492

Недостаток · AISAFE2-WEAK-065

Избыточная маршрутизация задач на человеческую проверку

Краткое описание

Критерии, гранулярность и отсутствие переиспользования решений направляют на ручную проверку ненужные или повторно эквивалентные задачи.

Подробное описание

Один пользовательский сценарий дробится на чрезмерное число согласований, а однотипные обращения рассматриваются независимо без безопасного применения ранее принятого решения. Ограниченный человеческий ресурс расходуется без пропорционального снижения риска.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-025 · Подтверждение и пропускная способность HITL

Прямые причинные воздействия

AISAFE2-IMP-023 · Перегрузка контура человеческой проверки

Избыточная, слишком мелкая и повторная маршрутизация задач на ручную проверку для эквивалентных случаев генерирует поток обращений, превышающий пропускную способность контура HITL, что напрямую повышает нагрузку на человеческие ресурсы.

Условия проявления

1. **Предпосылка.** В потоке обработки есть обязательная ручная проверка с ограниченной пропускной способностью.
2. **Триггер.** Каждое обращение направляется на новое обязательное согласование, и очередь заполняется задачами, не требующими участия человека.
3. **Разрешающий пробел.** Правила маршрутизации направляют в неё задачи без учёта фактической критичности: дробят сценарии на цепочки согласований, не объединяют эквивалентные повторы, не переиспользуют принятые решения.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	184	65
AISAFE2-IMP-023 · Перегрузка контура человеческой проверки	434	129
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	674	459

Недостаток · AISAFE2-WEAK-066

Недостаточная мощность и резервирование человеческой проверки

Краткое описание

Численность, график и резерв проверяющих рассчитаны ниже пиковой потребности обязательного контура человеческого контроля.

Подробное описание

Порог отказа контура проверки оказывается низким и предсказуемым: при пиковой нагрузке или отсутствии сотрудников обязательные проверки не выполняются вовремя, зависящие от них операции задерживаются или останавливаются, а злоумышленник может намеренно исчерпать контур сравнительно небольшим воздействием.

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-025 · Подтверждение и пропускная способность HITL

Прямые причинные воздействия

AISAFE2-IMP-023 · Перегрузка контура человеческой проверки

Численность, график и резерв операторов, рассчитанные ниже пиковой потребности, не выдерживают всплеска обращений, что приводит к росту очереди и задержек в контуре HITL и напрямую повышает нагрузку на человеческие ресурсы.

Условия проявления

1. **Предпосылка.** Выполнение операций зависит от своевременного решения проверяющего человека.
2. **Триггер.** Наступает пик нагрузки, включая намеренное исчерпание контура злоумышленником.
3. **Разрешающий пробел.** Численность, график и резерв проверяющих рассчитаны ниже пиковой потребности, механизма оперативного привлечения резерва нет.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-023 · Перегрузка контура человеческой проверки	435	129
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	675	459

Недостаток · AISAFE2-WEAK-067

Недостаточная инструментальная поддержка оператора HITL

Краткое описание

Интерфейс и процесс ручной проверки не дают оператору необходимый контекст, выделенные материалы, историю похожих решений и безопасные групповые действия.

Подробное описание

Оператор тратит лишнее время на восстановление контекста и повторение однородных операций. Неполное представление основания и последствий решения повышает трудоёмкость каждой задачи и вероятность ошибки.

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-025 · Подтверждение и пропускная способность HITL

Прямые причинные воздействия

AISAFE2-IMP-023 · Перегрузка контура человеческой проверки

Недостаточная инструментальная поддержка увеличивает время обработки задания; когда поток превышает снизившуюся пропускную способность операторов, очередь и нагрузка растут.

Ограничение: При потоке ниже доступной пропускной способности увеличение времени обработки само по себе не означает перегрузку человеческого контура.

Условия проявления

1. **Предпосылка.** Решение по задаче принимает оператор ручной проверки.
2. **Триггер.** Оператор принимает решение на основании материалов, представленных рабочим местом HITL.
3. **Разрешающий пробел.** Материалы подаются без выделения значимых фрагментов, полного контекста и последствий решения, истории похожих случаев или необходимых групповых операций.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	185	65
AISAFE2-IMP-023 · Перегрузка контура человеческой проверки	436	129
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	676	459

Недостаток · AISAFE2-WEAK-068

Недостаточный контроль качества решений и утомления операторов HITL

Краткое описание

При длительной или растущей нагрузке не отслеживаются утомление операторов, качество их решений и склонность автоматически соглашаться с рекомендацией системы, и человеческий контроль вырождается в формальное согласование.

Подробное описание

Без метрик времени и качества, ротации, независимой выборочной проверки и реакции на отклонения перегрузка постепенно ослабляет человеческий контроль и превращает его в формальное согласование.

Происхождение

- Процесс
- Реализация

Классы недостатка

- AISAFE2-WC-025 · Подтверждение и пропускная способность HITL

Условия проявления

1. **Предпосылка.** Нагрузка на операторов ручной проверки длительная или растущая, в том числе создаваемая намеренно.
2. **Триггер.** Решения операторов принимаются системой без оценки отклонений от типового поведения.
3. **Разрешающий пробел.** Время рассмотрения, качество решений, утомление и склонность автоматически соглашаться с рекомендацией системы не отслеживаются, ротация и выборочная перепроверка не применяются.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	186	65
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	677	459

Недостаток · AISAFE2-WEAK-069

Неопределённый режим переполнения и недоступности человеческой проверки

Краткое описание

Система не имеет проектного поведения при переполнении очереди или недоступности проверяющего и может молча пропустить контроль либо полностью остановить обслуживание.

Подробное описание

Не определены приоритеты, допустимая деградация, запрет или приостановка высокорисковых сценариев, сохранение запросов и безопасное восстановление после возврата человеческого ресурса. Не определён и сам переход между автоматической обработкой, очередью проверки и итоговым решением на время недоступности человеческого контроля.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-025 · Подтверждение и пропускная способность HITL

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Неопределённое поведение при перегрузке или недоступности HITL-контура может привести к обходу ограничений и предусмотренной логики агента, потому что при переполнении очереди проверки система пропускает обращения без проверки, превращая перегрузку HITL-контура в обход обязательного контроля.

Ограничение: Только для сценария продолжения работы без обязательной проверки: при недоступной человеческой проверке операция, которая по утверждённому маршруту обязана ждать решения HITL, продолжает выполнение без этой проверки и меняет предусмотренный план, действие или переход Agent Core.

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Неопределённое поведение при перегрузке или недоступности HITL-контура может привести к недоступности системы или непредоставлению результата в директивный срок, потому что при недоступности или переполнении HITL-контура вариант полной остановки обработки оборачивается отказом в обслуживании легитимных пользователей.

Ограничение: Только для сценария безопасного отказа с блокировкой операции: при недоступной человеческой проверке операция остаётся заблокированной, и потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Условия проявления

1. **Предпосылка.** Выполнение высокорисковых операций зависит от ручной проверки.

2. **Триггер.** Очередь проверки переполняется в результате штатной пиковой нагрузки либо из-за намеренного переполнения.
3. **Разрешающий пробел.** Штатное поведение для этого случая не определено: приоритеты, допустимая деградация, приостановка рисковых сценариев, процедуры безопасного восстановления не заданы.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.2.1 · Диагностический признак сбоя

Недоступность проверяющего может открыть обход подтверждения, если система пропускает высокорисковое действие при переполненной очереди.

Граница связи: Связь относится к режиму пропуска контроля; безопасная приостановка не является нарушением.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверенное подтверждение человека до привилегированного или необратимого действия.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	187	65
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	437	101
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	438	127
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	678	459

Недостаток · AISAFE2-WEAK-070

Отсутствие управляемой политики допустимого использования AI-сервисов

Краткое описание

Организация не определяет и не поддерживает разрешённые AI-сервисы, допустимые сценарии и классы данных, обязанности сотрудников и санкционированные альтернативы.

Подробное описание

Без понятных правил, владельца, обучения пользователей и периодического пересмотра сотрудники могут передавать данные в неподходящие публичные сервисы или применяют их в запрещённых сценариях как часть обычной работы.

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-036 · Соответствие и допустимое использование

Прямые причинные воздействия

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Отсутствие управляемой политики допустимого использования AI-сервисов может привести к утечке чувствительных данных в недоверенный контур, потому что при отсутствии правил, обучения и внутренних альтернатив сотрудники в рамках обычной работы отправляют служебные документы в публичные AI-сервисы.

Ограничение: Только если сотрудник отправил в публичный AI-сервис чувствительные данные, для которых этот сервис и цель обработки не разрешены; отсутствие политики без передачи недостаточно

Условия проявления

1. **Предпосылка.** Сотрудники обрабатывают рабочие данные или выполняют рабочие задачи с помощью AI-сервисов, включая публичные сервисы.
2. **Триггер.** Сотрудник передаёт данные в самостоятельно выбранный AI-сервис.
3. **Разрешающий пробел.** Разрешённые сервисы, допустимые сценарии и классы данных, обязанности и внутренние альтернативы не определены или не доведены до персонала, а передача не сверяется с политикой.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	439	130
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	679	519

Недостаток · AISAFE2-WEAK-071

Отсутствие явной модели доверенных зон и потоков данных

Краткое описание

Архитектура не определяет доверенные и недоверенные зоны, не инвентаризирует пересекающие их потоки и не фиксирует утверждённые правила таких переходов.

Подробное описание

Доверенные и недоверенные зоны не определены или определены с ошибками: модели, инструменты, хранилища и внешние API рассматриваются как единый контур без проверяемого решения о доверии, потоки данных между зонами не инвентаризированы, а их источники, получатели, классы данных, назначения и допустимость не зафиксированы. В результате затрудняются выбор технических мер защиты и подтверждение полноты их охвата: передача чувствительных данных во внешний недоверенный контур может выглядеть для системы штатной и не распознаваться как нарушение.

Область применения

- Общая безопасность

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-039 · Модель зон и межзонное принуждение

Прямые причинные воздействия

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Ошибочно классифицированный маршрут может передать чувствительные данные из AI-контура получателю, которому их передача запрещена действующим решением, если проверка зоны и потока не срабатывает до отправки.

Ограничение: Только если для конкретного класса данных и получателя отдельно установлено, что внешний поток не разрешён; отсутствие схемы зон само по себе не доказывает утечку

Условия проявления

1. **Предпосылка.** Между моделями, инструментами, хранилищами и внешними API проектируются или действуют потоки данных.
2. **Триггер.** Соединённый компонент передаёт или начинает принимать данные через такой поток как внутри единого доверенного контура.
3. **Разрешающий пробел.** Зоны доверия и переходы между ними не определены либо определены ошибочно, а потоки не инвентаризированы и не классифицированы.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	440	130
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	680	514

Недостаток · AISAFE2-WEAK-072

Отсутствие обязательного технического контроля межзонных потоков данных

Краткое описание

Утверждённые межзонные потоки данных не проходят через обязательную техническую точку, которая проверяет их атрибуты и блокирует неизвестные или запрещённые переходы.

Подробное описание

Даже при описанных доверенных зонах компонент может передать данные внешней модели, инструменту, хранилищу или API в обход детерминированного шлюза. Фактический поток не сопоставляется с утверждёнными источником, получателем, назначением и классом данных до пересечения границы.

Область применения

- Общая безопасность

Происхождение

- Реализация

Классы недостатка

- AISAFE2-WC-039 · Модель зон и межзонное принуждение

Прямые причинные воздействия

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Обход обязательного межзонного контроля позволяет фактически передать чувствительные данные получателю, которому такой поток запрещён утверждённой политикой.

Ограничение: Только при подтверждённой передаче запрещённому получателю; обход шлюза на разрешённом потоке сам по себе не образует утечку.

Условия проявления

1. **Предпосылка.** Утверждённые границы зон доверия требуют контролировать передачу данных между зонами.
2. **Триггер.** Компонент передаёт данные внешней модели, инструменту, хранилищу или API через отсутствующий либо обходной маршрут контроля.
3. **Разрешающий пробел.** До пересечения границы фактический поток не проверяется обязательной политикой по источнику, получателю, назначению и классу данных.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.5.3 · Диагностический признак сбоя

Обход обязательного межзонного шлюза оставляет передачу данных без независимой проверки источника, получателя и разрешённого класса.

Граница связи: Связь касается доступа на межзонном переходе; прочие решения об авторизации она не описывает.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения требует решения о доступе в логике приложения или независимом механизме политики.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	188	77
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	441	130
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	681	519
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	682	531

Недостаток · AISAFE2-WEAK-073

Отсутствие независимого проверочного материала для AI-результата

Краткое описание

Результат, для которого выполняется независимая детерминированная проверка, передаётся потребителю без сформированных ею оснований и находок, необходимых для осознанного применения.

Подробное описание

Контур постобработки не прикладывает к результату проверочный материал, созданный независимо от модели: подтверждение утверждений по доверенным источникам, отчёт статического анализа или аннотации конкретных рискованных элементов. Потребитель видит итог, но не видит основания независимой проверки и вынужден повторять её вручную либо применять результат без неё. Состав проверочного материала задаётся контрактом конкретного результата и может различаться для разных типов вывода.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-024 · Доверие человека и прозрачность интерфейса

Прямые причинные воздействия

AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой

Независимая проверка выявляет конкретную опасность в явно предложенном действии, но находка не доставляется потребителю. Из-за этого не срабатывает заранее установленный обязательный пересмотр или отказ; действие выполняется и причиняет подтверждённый вред.

Ограничение: Только если подтверждены предупреждающая находка, обязательный маршрут пересмотра или отказа при ней, его пропуск из-за недоставки, выполненное действие и вред.

AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы

Независимая проверка выявляет конкретную ошибку в AI-результате, но находка не доставляется до решения. Из-за этого не срабатывает заранее установленный обязательный пересмотр или отказ; результат становится существенным основанием исполненного решения с подтверждённым вредом.

Ограничение: Только если подтверждены находка, обязательный маршрут пересмотра или отказа при ней, его пропуск из-за недоставки, существенное влияние результата на исполненное решение и вред.

Условия проявления

1. **Предпосылка.** Потребителю передаётся результат для применения: утверждения, рекомендации, исполняемые артефакты.
2. **Триггер.** Содержательная проверка требует от потребителя самостоятельного расследования, которое он не проводит.
3. **Разрешающий пробел.** Основания и находки независимой проверки не прикладываются к результату до его передачи потребителю.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	189	35
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	190	72
AISAFE2-IMP-001 · Выполнение действий, явно предложенных AI-системой	442	92
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	443	93
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	683	331

Недостаток · AISAFE2-WEAK-074

Отсутствие независимого разграничения содержимого контекста и вывода модели

Краткое описание

Чувствительные элементы контекста доступны модели при генерации ответа, а их нераскрытие зависит только от вероятностного поведения модели и не подкреплено независимым техническим контролем.

Подробное описание

Системные инструкции, описания инструментов, пользовательские данные или результаты вызовов могут находиться в контексте, доступном модели для формирования ответа. Если получатель не должен видеть один из этих элементов, текстовый запрет или обученное поведение сами по себе не образуют проверяемой границы доступа. Субъект, влияющий на запрос или промежуточный контекст, может попытаться вызвать дословное либо преобразованное раскрытие; успешность зависит от модели, контекста и внешних контролей и не предполагается гарантированной.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-003 · Внутреннее поведение и устойчивость ограничений
- AISAFE2-WC-008 · Системные инструкции и детерминированные контроли

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Отсутствие гарантий нераскрытия содержимого контекста модели может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что между содержимым контекста и выводным каналом нет технического разграничения доступа, и данные других участников обработки раскрываются через ответ дословно или в преобразованной форме.

Ограничение: Только если сведения из инструментального контекста раскрыты через вывод модели получателю без полномочий; доступ модели к контексту сам по себе недостаточен

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Отсутствие гарантий нераскрытия содержимого контекста модели может привести к раскрытию системного промпта и инструкций инструментов, потому что системная инструкция и описания инструментов — такой же текст однородного контекста без меток доступа, и их раскрытие через выводной канал технически не исключено.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Модель может воспроизвести чувствительный фрагмент контекста в ответе или вызове инструмента; если этот вывод фактически отправлен получателю без права на данные, происходит запрещённая передача.

Ограничение: Только если содержимое контекста с чувствительными данными фактически выдано моделью получателю без разрешения; отсутствие технического разграничения само по себе недостаточно

Условия проявления

1. **Предпосылка.** В контексте модели находится чувствительное содержимое: системная инструкция, схемы инструментов, результаты вызовов или данные других участников обработки.
2. **Триггер.** Субъект влияет на формирование ответа запросами или содержимым, попадающим в контекст.
3. **Разрешающий пробел.** Нераскрытие обеспечивается только обучением и текстовыми инструкциями без независимого технического разграничения между контекстом и каналом вывода.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	191	37
AISAFE2-COMP-003 · Управление контекстом общения с моделью	192	40
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	444	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	445	107
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	446	130
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	684	349

Недостаток · AISAFE2-WEAK-075

Отсутствие совокупных пределов масштабирования и расходов инфраструктуры

Краткое описание

Автоматическое масштабирование и оплачиваемое по фактическому потреблению выделение ресурсов не ограничены принудительным верхним пределом, а бюджетные пороги владельца не связаны с автоматическими защитными действиями.

Подробное описание

Автоматическое масштабирование и оплачиваемое по фактическому потреблению выделение ресурсов не имеют принудительного верхнего предела: верхние границы масштабирования и совокупные бюджетные лимиты владельца не заданы, а пороговые оповещения не связаны с автоматическим ограничением. Рост нагрузки — штатный, ошибочный или созданный злоумышленником — ведёт к выделению ресурсов и расходам, которые продолжаются до ручного вмешательства.

Область применения

- Общая безопасность

Происхождение

- Реализация
- Проектирование

Классы недостатка

- AISAFE2-WC-020 · Лимиты, ресурсы и контроль затрат

Прямые причинные воздействия

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Отсутствие совокупных пределов масштабирования и расходов инфраструктуры может привести к чрезмерному потреблению вычислительных ресурсов системы, потому что автоматическое масштабирование не имеет принудительного верхнего предела, а бюджетные пороги не связаны с защитными действиями, и выделение ресурсов растёт до ручного вмешательства.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-075 · Отсутствие совокупных пределов масштабирования и расходов инфраструктуры
> 2. Промежуточное воздействие: AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов
> 3. Промежуточное состояние: Стоимость горизонтального масштабирования, ремонта или замены измерена и атрибутирована конкретному превышению AI-нагрузки
> 4. Итоговое воздействие: AISAFE2-IMP-025 · Прямые финансовые потери
>
> **Механизм переходов**
>
> 1. Отсутствие жёсткого ресурсного лимита позволяет фактической AI-нагрузке масштабироваться и превысить утверждённый ресурсный бюджет.
> 2. Метринг связывает перерасход с рабочей нагрузкой и фиксирует точный счёт за горизонтальное масштабирование либо подтверждённую стоимость ремонта или замены.
> 3. После доступного урегулирования атрибутированная точная сумма сверх утверждённого бюджета остаётся невозвратным чистым убытком.

Условия применимости

- Жёсткий ресурсный лимит отсутствует, рабочая нагрузка фактически масштабируется и I019 превышает утверждённый ресурсный бюджет.
- Счёт либо стоимость актива точно атрибутированы перерасходу и остаются невозвратными после урегулирования.

Исключения

- Настроенный эффективный ресурсный лимит, отсутствие горизонтального масштабирования, расход в бюджете и нормальная плановая ёмкость исключают путь.
- Ресурсный всплеск без финансового убытка и повторный учёт той же суммы недостаточны.

Условия проявления

1. **Предпосылка.** Инфраструктура автоматически выделяет ресурсы, оплачиваемые по фактическому потреблению.
2. **Триггер.** Нагрузка растёт: штатно, из-за ошибки или действий злоумышленника.
3. **Разрешающий пробел.** Принудительная верхняя граница масштабирования и совокупный бюджет владельца не заданы, а превышение порогов не запускает автоматических ограничений.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM06:2026 · Диагностический признак сбоя

Масштабирование без верхнего предела может превращать рост AI-нагрузки в продолжающиеся инфраструктурные расходы владельца.

Граница связи: Карточка относится к оплачиваемой ёмкости инфраструктуры, а не непосредственно к токенам модели или цене внешних инструментов.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	193	58
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	194	79
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	447	123
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	685	333

Недостаток · AISAFE2-WEAK-076

Отсутствие ресурсной изоляции арендаторов в разделяемой среде

Краткое описание

Потребление одного арендатора не ограничивается его долей общих ресурсов, поэтому его нагрузка ухудшает качество обслуживания других арендаторов.

Подробное описание

Общие вычислительные ресурсы инференса, очереди или каналы не имеют эффективной пер-арендаторной квоты, резерва либо справедливого планирования. Повышенная активность одного арендатора занимает общую ёмкость, увеличивает задержки и долю незавершённых запросов у остальных, хотя их собственная нагрузка не изменилась. Для применения ограничения потребление должно быть измеримо по субъекту; корректность выставления счетов оценивается отдельно.

Область применения

- Общая безопасность

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-040 · Производительность, ёмкость и управление перегрузкой

Прямые причинные воздействия

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Отсутствие ресурсной изоляции арендаторов позволяет нагрузке одного арендатора занимать общие очереди и вычислительные ресурсы, из-за чего другие арендаторы не получают результат в установленный директивный интервал.

Условия проявления

1. **Предпосылка.** Несколько арендаторов обслуживаются на общих вычислительных ресурсах, очередях или каналах.
2. **Триггер.** Один арендатор легитимно или намеренно создаёт повышенную нагрузку.
3. **Разрешающий пробел.** Потребление арендатора не ограничено выделенной долей общего ресурса или справедливым обслуживанием.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM06:2026** · Диагностический признак сбоя
>
> Без квоты или справедливого распределения нагрузка одного арендатора может вытеснить запросы остальных из общей среды инференса.
>
> **Граница связи:** Требуется измеримая конкуренция за общий ресурс; ошибка изоляции не доказывает намеренное злоупотребление или перерасход денег.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	195	37
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	196	58
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	197	75
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	448	127
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	686	333

Недостаток · AISAFE2-WEAK-077

Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку

Краткое описание

Вычислительная ёмкость и механизм её наращивания рассчитаны без запаса на пиковую нагрузку, поэтому штатный рост обращений выводит время ответа за установленный предел.

Подробное описание

Заданные целевые времена ответа и профиль пиковой нагрузки не учитываются при расчёте ёмкости. Масштабирование отсутствует, опирается на нерелевантные показатели или запаздывает относительно роста спроса, а процесс планирования не пересматривает ресурсы по фактической нагрузке. В результате предсказуемый штатный пик, без атаки и отказа компонентов, выводит задержки за установленный предел.

Область применения

- Общая безопасность

Происхождение

- Процесс
- Реализация

Классы недостатка

- AISAFE2-WC-040 · Производительность, ёмкость и управление перегрузкой

Прямые причинные воздействия

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку может привести к недоступности системы или непредоставлению результата в директивный срок, потому что ёмкость и механизм масштабирования рассчитаны без запаса на пиковую нагрузку, и штатный рост обращений выводит время ответа за директивные пределы.

Условия проявления

1. **Предпосылка.** Планирование не корректирует ресурсы по фактическому профилю нагрузки.
2. **Триггер.** Штатная нагрузка растёт или достигает пиковых значений.
3. **Разрешающий пробел.** Ёмкость рассчитана без запаса, а масштабирование отсутствует, опирается на нерелевантные показатели или отстаёт от спроса.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	198	58
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	199	79
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	449	127
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	687	333

Недостаток · AISAFE2-WEAK-078

Недостатки управления нагрузкой, очередями и приоритетами при перегрузке

Краткое описание

Система не локализует перегрузку: приём, очереди и планирование выполнения не рассчитаны на спрос выше ёмкости, очереди растут неограниченно, а деградация распространяется на всех потребителей.

Подробное описание

При спросе выше доступной ёмкости система продолжает принимать запросы без ограничения интенсивности, ограниченной очереди, сброса избыточной нагрузки и приоритетов для критичных потребителей. Неограниченные очереди и несправедливое планирование удерживают ресурсы, увеличивают задержки и распространяют деградацию на весь контур вплоть до отказа в обслуживании. Недостаток проявляется после насыщения и сам по себе не свидетельствует о неверно спланированной ёмкости.

Область применения

- Общая безопасность

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-040 · Производительность, ёмкость и управление перегрузкой

Прямые причинные воздействия

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Недостатки управления нагрузкой, очередями и приоритетами при перегрузке могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что при спросе выше ёмкости приём не ограничивается, очереди растут неограниченно, и деградация распространяется на всех потребителей вплоть до отказа в обслуживании.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Условия проявления

1. **Предпосылка.** Автоматический контур принимает запросы без ограничения, использует неограниченные очереди и не задаёт приоритеты потребителей.
2. **Триггер.** Совокупный входящий поток превышает ёмкость обработки из-за штатного пика, ресурсоёмкого клиента или атаки.
3. **Разрешающий пробел.** Отсутствует механизм сброса избыточной нагрузки и защиты критичных потребителей.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM06:2026 · Диагностический признак сбоя

Неограниченный приём и рост очереди при перегрузке удерживают ресурсы и распространяют задержки на весь AI-сервис.

Граница связи: Карточка описывает обработку перегрузки после насыщения; она не устанавливает источник спроса и не задаёт бюджет отдельного инференса.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	200	35
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	201	58
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	202	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	203	75
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	450	127
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	688	333

Недостаток · AISAFE2-WEAK-079

Зависимость от единственного поставщика сервисов или площадки размещения

Краткое описание

Функционирование AI-системы ограничено доступностью единственного поставщика моделей, инструментов, инфраструктуры, канала связи или площадки размещения, для которого не подготовлены замена и переключение.

Подробное описание

Архитектура зависит от единственного поставщика модели, инструмента, инфраструктуры, канала связи или площадки: совместимая альтернатива не подготовлена, переносимость данных, конфигурации и интеграций не проверена, а переключение не отрепетировано. Отказ, прекращение обслуживания, изменение условий или регламентные работы внешней стороны останавливают функцию на срок, который владелец не может сократить заменой.

Область применения

- Общая безопасность

Происхождение

- Проектирование
- Процесс

Классы недостатка

- AISAFE2-WC-041 · Зависимость от поставщика и переносимость

Прямые причинные воздействия

AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала

Зависимость от единственного поставщика сервисов или площадки размещения может привести к недоступности системы или непредоставлению результата в директивный срок, потому что доступность ограничена единственным поставщиком без подготовленной замены и переключения, и его авария или регламентные работы означают простой на неконтролируемый владельцем срок.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения
> 2. Промежуточное воздействие: AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала
> 3. Итоговое воздействие: AISAFE2-IMP-026 · Нарушение операционной деятельности, организационных процессов или бизнес-процессов
>
> **Механизм переходов**
>
> 1. Недоступность единственного внешнего поставщика без подготовленного переключения делает обязательную функцию системы технически недоступной.
> 2. Техническая недоступность превышает допустимый интервал именованного обязательного шага без эффективной замены и вызывает наблюдаемое отклонение бизнес-процесса.
>
> **Условия применимости**
>
> - Фактически недоступен единственный поставщик, а подготовленного переключения на резервного поставщика или работоспособного обходного пути нет.
> - Назван обязательный процессный шаг, его допустимый интервал превышен и зафиксировано фактическое отклонение процесса.
>
> **Исключения**
>
> - Сама архитектурная зависимость без фактической недоступности не образует путь.

- Сбой в допустимом интервале либо доступный эффективный обходной путь исключают финальный переход.

Условия проявления

1. **Предпосылка.** Функция AI-системы зависит от единственного поставщика или площадки размещения.
2. **Триггер.** Поставщик становится недоступен из-за аварии, регламентных работ, изменения условий или прекращения обслуживания.
3. **Разрешающий пробел.** Нет подготовленной альтернативы и переносимости данных, конфигурации и интеграций, поэтому переключение в допустимый срок невозможно.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	204	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	205	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	206	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	207	79
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	451	127
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	689	492

Недостаток · AISAFE2-WEAK-080

Ограниченная полнота внешнего семантического контроля вывода

Краткое описание

Внешний фильтр, основанный на конечном наборе форм или признаков, не обеспечивает полный контроль произвольного смыслового ограничения в открытом тексте: эквивалентное содержание может быть выражено иначе.

Подробное описание

Одно и то же содержание можно выразить парафразом, переводом, кодированием или косвенным описанием. Поэтому правило или детектор, распознающий известные формы, может пропустить семантически эквивалентный вывод. Это ограничение относится к произвольным смысловым свойствам открытого текста и не означает невозможность контроля формализуемых свойств: синтаксис, типы, фиксированные словари, структурные схемы и детерминированные права могут проверяться принудительно.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-013 · Проверка вывода и контекстная изоляция

Прямые причинные воздействия

AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы

Недопустимое содержание выражается парафразом, переводом или кодированием, не распознаётся конечным набором признаков внешнего фильтра и фактически достигает пользователя.

Ограничение: Только для конкретного воздействия на пользователя по заранее утверждённым нормам, с указанной затронутой группой и подтверждённым вредом.

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Преобразованное чувствительное содержание из инструментального источника не распознаётся семантическим фильтром и выдаётся неуполномоченному получателю.

Ограничение: Только при фактической выдаче или воспроизводимой реконструкции чувствительных сведений из инструментального источника неуполномоченным получателем.

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Системная инструкция передаётся парафразом, переводом или кодированием, которые внешний фильтр не связывает с защищаемым исходным содержанием.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Чувствительные сведения в преобразованной форме обходят внешний семантический фильтр и фактически пересекают границу доверия.

Ограничение: Только если преобразованный вывод всё ещё раскрывает чувствительные сведения получателю, которому их передавать нельзя

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода
> 2. Промежуточное состояние: Семантически эквивалентная опасная формулировка проходит конечный внешний фильтр и принимается управляющим контуром агента
> 3. Итоговое воздействие: AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core
>
> **Механизм переходов**
>
> 1. Ограниченный набор формулировок внешнего фильтра не распознаёт семантически эквивалентный обход, поэтому опасный текст достигает управляющего контура агента.
> 2. Agent Core принимает прошедшую формулировку как управляющий сигнал, и трасса фиксирует фактическое изменение подцели, инструмента, аргументов, последовательности либо состояния.

Условия применимости

- Для выхода задана точная семантическая политика, а проверка подтверждает фактический обход альтернативной формулировкой.
- Прошедший текст потребляется управляющим контуром Agent Core, и трасса показывает изменение плана, действия либо состояния.

Исключения

- Текст, который видит только пользователь и который не входит в управляющий контур агента, не образует этот путь.
- Неожиданная формулировка без доказанного изменения состояния или действия Agent Core недостаточна.

Условия проявления

1. **Предпосылка.** Требование к открытому тексту задано только в терминах смысла, например как нераскрытие определённых сведений.
2. **Триггер.** Субъект, влияющий на генерацию, выражает то же содержание парафразом, переводом, кодированием или косвенным описанием.
3. **Разрешающий пробел.** Проверка вывода распознаёт только конечный набор известных форм выражения защищаемого смысла.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	208	37
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	209	77
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	452	98
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	453	106
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	454	107
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	455	130
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	690	320
AISAFE2-MIT-056 · Исполнение версионируемой политики допустимости содержимого	691	529

Недостаток · AISAFE2-WEAK-081

Неаутентифицированный внешний вход в AI-приложение

Краткое описание

Внешний клиент достигает AI-функции или её внутреннего сервиса без проверенного способа установить вызывающего субъекта до обработки запроса.

Подробное описание

Публичный, партнёрский или иной внешний маршрут принимает запрос без проверки учётных данных либо оставляет прямой путь к внутреннему сервису, обходящий аутентификацию шлюза. В результате квоты, авторизация данных, аудит и ограничения действий связываются с общей или заявленной самим клиентом идентичностью. Явно анонимная функция допустима только в отдельно утверждённой области с собственными ограничениями; допуск к ней не заменяет авторизацию последующих операций.

Область применения

- Общая безопасность

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-042 · Аутентификация субъектов

Условия проявления

1. **Предпосылка.** Опубликованный либо прямой маршрут внутреннего AI-сервиса доступен внешнему клиенту до установления доверенной идентичности.
2. **Триггер.** Атакующий обращается к этому маршруту без проверяемых учётных данных либо подставляет непроверенный идентификатор.
3. **Разрешающий пробел.** Приложение продолжает модельный вызов, чтение данных или другое значимое действие без проверки субъекта на первом доверенном входе и обходных путях.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	210	35
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	211	72
AISAFE2-MIT-058 · Аутентификация на внешнем входе AI-приложения	692	535

Недостаток · AISAFE2-WEAK-082

Недостаточная стойкость аутентификации привилегированных человеческих учётных записей

Краткое описание

Привилегированная человеческая учётная запись принимает единственный или недостаточно стойкий фактор, не соответствующий последствиям её компрометации.

Подробное описание

Высокорисковый административный доступ защищён паролем, повторно используемым фактором, исключением из MFA или слабым восстановлением. Кража, фишинг или повтор учётных данных позволяют атакующему действовать с корректно назначенными привилегиями, поэтому минимальность роли и периодический пересмотр сами по себе не закрывают путь.

Область применения

- Общая безопасность

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-042 · Аутентификация субъектов

Прямые причинные воздействия

AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и фактически прочитать системный промпт либо инструкцию инструмента через штатную конфигурацию.

Ограничение: Только если атакующий, принятый за привилегированного пользователя, фактически прочитал системный промпт либо инструкцию инструмента через конфигурацию

AISAFE2-IMP-011 · Компрометация баз знаний

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и фактически экспортировать существенную часть базы знаний, её метаданных либо индекса через функции управления.

Ограничение: Только если атакующий под видом привилегированного пользователя экспортировал существенную часть базы знаний, её метаданных или индекса

AISAFE2-IMP-012 · Кража модели

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и несанкционированно скопировать веса либо связанные артефакты модели.

Ограничение: Только если атакующий под видом привилегированного пользователя фактически скопировал веса или связанные артефакты модели

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и успешно прочитать либо изменить хранимый объект вне собственных полномочий.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и успешно вызвать инфраструктурную функцию вне собственных полномочий.

Ограничение: Только если подмена привилегированного пользователя привела к успешному вызову функции, недоступной атакующему

Условия проявления

1. **Предпосылка.** Привилегированный человек входит в административный или служебный интерфейс, управляющий конфигурацией, моделями, знаниями или данными AI-системы.
2. **Триггер.** Атакующий подбирает, похищает или воспроизводит фактор входа посредством перебора, фишинга или утечки.
3. **Разрешающий пробел.** Привилегированный вход принимает один недостаточно стойкий фактор, не ограничивает попытки или допускает путь без контекстной усиленной проверки.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	212	61
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	213	79
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	456	107
AISAFE2-IMP-011 · Компрометация баз знаний	457	109
AISAFE2-IMP-012 · Кража модели	458	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	459	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	460	118
AISAFE2-MIT-059 · Усиленная аутентификация привилегированных человеческих учётных записей	693	538

Недостаток · AISAFE2-WEAK-083

Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей

Краткое описание

Служебный или административный интерфейс AI-компонента принимает соединения из недоверенной либо избыточно широкой сети вместо только утверждённых источников и маршрутов.

Подробное описание

Административные консоли, внутренние точки инференса, серверы инструментов, реестры моделей и интерфейсы базы знаний доступны из Интернета либо широкого внутреннего сегмента без разрешающего списка источников. Прямой путь обходит прикладной шлюз, ограничения частоты и иные контроли обычного маршрута, а ошибка аутентификации, настройки или кода становится удалённо достижимой.

Область применения

- Общая безопасность

Происхождение

- Реализация
- Проектирование

Классы недостатка

- AISAFE2-WC-038 · Доверенные зоны и потоки данных

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Сетевая достижимость служебного хранилища из недоверенной сети превращает отдельно установленный дефект аутентификации или конфигурации в удалённый путь, по которому атакующий фактически получает существенную часть базы знаний.

Ограничение: Только при отдельно установленном дефекте аутентификации или конфигурации, фактическом удалённом получении существенной части базы знаний и доказательстве, что недоверенный сетевой путь был необходимым условием доступа.

AISAFE2-IMP-012 · Кража модели

Сетевая достижимость внутренней точки инференса или хранилища из недоверенной сети позволяет использовать отдельно установленный дефект доступа или отсутствие штатного ограничения и довести кампанию до получения весов либо функционального извлечения модели.

Ограничение: Только при отдельно установленном дефекте доступа или обходе штатного ограничения, фактическом использовании недоверенного сетевого пути и подтверждённом получении весов либо функциональном извлечении модели.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Сетевая достижимость служебного интерфейса из недоверенной сети делает отдельно установленный дефект аутентификации или конфигурации удалённо применимым, после чего атакующий успешно читает либо изменяет хранимый объект.

Ограничение: Только при отдельно установленном дефекте аутентификации или конфигурации, успешной удалённой операции чтения либо изменения и доказательстве, что недоверенный сетевой путь сделал её достижимой.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Сетевая достижимость административного интерфейса из недоверенной сети делает отдельно установленный дефект аутентификации, авторизации или кода удалённо применимым, после чего атакующий успешно вызывает запрещённую функцию.

Ограничение: Только при отдельно установленном дефекте аутентификации, авторизации или кода, успешном удалённом вызове запрещённой функции и доказательстве, что недоверенный сетевой путь сделал её достижимой.

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Фактически достижимая из недоверенной сети внутренняя точка инференса позволяет обходить штатный внешний контур ограничения частоты; при отдельно установленном недостатке внутренних квот по субъекту или совокупного бюджета повторные прямые вызовы доводят потребление вычислений, памяти, сети либо ускорителей собственной среды выше утверждённого порога.

Ограничение: Только если доступный извне внутренний интерфейс обошёл внешний лимитер, а недостаточные внутренние квоты позволили повторными вызовами измеримо превысить бюджет ресурсов; токенный предел отдельного запроса может сохраняться

Условия проявления

1. **Предпосылка.** Служебный интерфейс AI-компонента предназначен для ограниченного административного или внутреннего сетевого доступа.
2. **Триггер.** Недоверенный узел устанавливает прямое соединение со служебным интерфейсом из Интернета или иной неутверждённой сети.
3. **Разрешающий пробел.** Эффективный сетевой рубеж не ограничивает источники и не принуждает утверждённый промежуточный маршрут доступа.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	214	37
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	215	44
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	216	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	217	58
AISAFE2-IMP-011 · Компрометация баз знаний	461	109

Карточка	Связи	Стр.
AISAFE2-IMP-012 · Кража модели	462	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	463	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	464	118
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	465	123
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	694	531

Недостаток · AISAFE2-WEAK-084

Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов

Краткое описание

Для точных версии и среды AI-компонента нет утверждённого эталона защищённых настроек либо соответствие ему не проверяется при развёртывании и после изменений.

Подробное описание

Для серверов инференса, реестров моделей, базы знаний, агентных сред, инструментов и телеметрии не заданы версия эталона, владелец, обязательные настройки и допустимые исключения. Развёртывание не сравнивает эффективную конфигурацию с эталоном, а эксплуатация не выявляет дрейф. В результате небезопасные настройки — анонимный доступ, отладочные функции или избыточные привилегии процессов — могут сохраняться незамеченными.

Область применения

- Общая безопасность

Происхождение

- Процесс
- Реализация

Классы недостатка

- AISAFE2-WC-035 · Безопасный выпуск и управление уязвимостями

Прямые причинные воздействия

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Отсутствие применимого эталона оставляет в рабочей среде конкретную небезопасную настройку, например анонимный доступ; через неё субъект без полномочий успешно читает либо изменяет хранимые данные.

Ограничение: Только для точной небезопасной настройки, которая действует из-за отсутствия применимого эталона и фактически приводит к успешному чтению либо изменению хранимого объекта субъектом без полномочий.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Отсутствие применимого эталона оставляет открытой конкретную служебную функцию или избыточное право процесса; через эту настройку субъект без полномочий успешно вызывает инфраструктурную функцию.

Ограничение: Только для точной открытой функции или избыточного права процесса, которые действуют из-за отсутствия применимого эталона и фактически приводят к успешному запрещённому вызову.

Условия проявления

1. **Предпосылка.** Компонент AI-инфраструктуры работает с настройками по умолчанию или с отклонениями, накопленными при поддержке и диагностике.
2. **Триггер.** Небезопасная настройка — анонимный доступ, отладочный режим, открытая служебная функция, избыточные права процесса — достижима для источника воздействия.
3. **Разрешающий пробел.** Утверждённый эталон защищённой конфигурации отсутствует либо соответствие ему не проверяется при развёртывании и периодически.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	218	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	219	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	220	44
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	221	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	222	58
AISAFE2-COMP-013 · Телеметрия и обнаружение	223	64
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	224	67
AISAFE2-COMP-023 · Подсистема формирования базы знаний	225	86
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	466	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	467	118
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	695	419
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	696	514
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	697	564

Недостаток · AISAFE2-WEAK-085

Отсутствие изолированных неизменяемых версий критичных AI-активов

Краткое описание

Для критичных наборов данных, артефактов модели, базы знаний, индексов и конфигурации нет версионированной копии вне рабочих полномочий, из которой проверенно восстанавливается точная версия.

Подробное описание

Основной экземпляр и копии критичного AI-актива доступны одной учётной записи, автоматизации или границе отказа, перезаписываются последней версией либо не проверяются восстановлением. Атакующий с рабочими полномочиями может удалить, зашифровать или повредить все экземпляры, а известное безопасное состояние остаётся незафиксированным или недостижимым. Восстановленная версия должна быть идентифицируема, но это не доказывает её смысловую корректность.

Область применения

- Общая безопасность

Происхождение

- Реализация
- Процесс

Классы недостатка

- AISAFE2-WC-031 · Готовность реагирования и восстановления

Прямые причинные воздействия

AISAFE2-IMP-031 · Утрата датасета в результате атаки

Недостаток позволяет атакующему изменить или удалить датасет обучения, тестирования либо валидации вместе с доступными доверенными копиями, вследствие чего требуемая версия датасета становится невосстановимой.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения доступных доверенных копий.

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Недостаток позволяет атакующему изменить или удалить веса, базу знаний, индекс либо критичную конфигурацию вместе с доступными доверенными версиями, вследствие чего требуемый AI-артефакт становится невосстановимым.

Ограничение: Только для вызванной атакой невосстановимой утраты критичного AI-артефакта, кроме датасета, когда отсутствует доступная доверенная версия или воспроизводимый рецепт.

Условия проявления

1. **Предпосылка.** Наборы данных, базы знаний, веса и конфигурации существуют только в пределах досягаемости полномочий рабочего контура.

2. **Триггер.** Субъект или процесс с правами на рабочее хранилище выполняет удаление, шифрование или порчу содержимого.
3. **Разрешающий пробел.** Версионированные копии, защищённые от изменения и удаления, не созданы либо доступны с теми же учётными данными и из той же сети.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(c) · Диагностический признак сбоя

Отсутствие изолированной пригодной копии AI-данных создаёт риск, что доступ к персональным данным не удастся восстановить после инцидента.

Граница связи: Модельные веса и иные артефакты не всегда содержат персональные данные; требуемые сроки и остальные пути восстановления оцениваются отдельно.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Затронутый актив содержит персональные данные либо необходим для восстановления доступа к ним.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-008 · Данные для подсистемы обучения	226	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	227	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	228	67
AISAFE2-COMP-023 · Подсистема формирования базы знаний	229	86
AISAFE2-IMP-031 · Утрата датасета в результате атаки	468	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	469	141
AISAFE2-MIT-009 · Версионирование и откат базы знаний	698	346
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	699	448
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных	700	503
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	701	568

Недостаток · AISAFE2-WEAK-086

Утрата атрибутов доступа исходного материала при индексировании базы знаний

Краткое описание

Фрагмент или производный объект индекса не сохраняет актуальные атрибуты доступа исходного материала, поэтому извлечение не может проверить право текущего субъекта на выдачу.

Подробное описание

При разбиении и индексировании не сохраняются связь с источником, разрешения, классификация, владелец и состояние отзыва; изменение прав или удаление материала не синхронизируется с фрагментами и производными объектами. Общий индекс затем отбирает данные только по релевантности, и даже известная идентичность инициатора не позволяет авторизовать окончательный фрагмент.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-010 · Контроль доступа и авторизация
- AISAFE2-WC-016 · Управление и происхождение данных

Прямые причинные воздействия

AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов

Утрата атрибутов доступа исходных документов при включении в базу знаний может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что фрагменты индексируются без переноса атрибутов доступа исходных документов, извлечение не может фильтровать выдачу по правам субъекта, и закрытые для него документы попадают в контекст и ответ.

Условия проявления

1. **Предпосылка.** В общий индекс базы знаний включены документы из источников с разными правами доступа.
2. **Триггер.** К системе допущены субъекты с разными правами на исходные документы, а извлечение отбирает фрагменты только по релевантности запросу.
3. **Разрешающий пробел.** Права, владелец и актуальность исходных документов не переносятся в метаданные фрагментов и не синхронизируются с источником.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C5.2.2 · Диагностический признак сбоя

Потеря атрибутов доступа при индексировании лишает извлечение сведений, нужных для проверки прав пользователя на фрагмент.

Граница связи: Риск возникает при общих индексах с разными правами; независимая проверка по источнику может его снять.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверку прав конечного пользователя на каждом этапе поиска и сборки контекста.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-023 · Подсистема формирования базы знаний	230	86
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	470	106
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	702	342

Недостаток · AISAFE2-WEAK-087

Ненаследование чувствительности исходных данных векторными представлениями

Краткое описание

Векторное представление получает более слабый класс обработки, чем исходные данные, хотя при доступной модели и дополнительных данных оно может раскрывать содержание, связь с записью или чувствительные атрибуты.

Подробное описание

Класс чувствительности источника не переносится на эмбеддинг при вычислении, хранении, передаче, экспорте и резервировании. Векторы попадают во внешние сервисы, общие индексы или аналитические копии как якобы нечувствительные числовые данные. Риск зависит от модели, вспомогательных данных и доступа атакующего и включает реконструкцию, вывод атрибутов, связывание с источником и установление присутствия записи; обратимость не заявляется универсально.

Происхождение

- Проектирование

Классы недостатка

- AISAFE2-WC-016 · Управление и происхождение данных
- AISAFE2-WC-018 · Минимизация данных и приватность

Прямые причинные воздействия

AISAFE2-IMP-011 · Компрометация баз знаний

Ослабленная защита производного векторного представления позволяет неразрешённому субъекту по результатам теста для точной модели и конфигурации получить либо воспроизводимо реконструировать существенную часть базы знаний или индекса.

Ограничение: Только при тесте, относящемся к конкретному представлению, с точными моделью и версией, формой вектора или параметрами квантования, доступом атакующего, вспомогательными данными, корпусом, методом и измеренным порогом существенного восстановления. Универсальная обратимость векторных представлений не заявляется.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Когда эмбеддинг чувствительного источника ошибочно обрабатывается как нечувствительный, он может быть передан получателю, которому исходный класс данных не разрешён.

Ограничение: Только при доказанной связи с чувствительным источником, наследовании класса обработки, фактическом переходе вектора через неутверждённую границу доверия и отсутствии разрешения у точного получателя. Универсальное восстановление исходных данных из векторных представлений не требуется и не заявляется.

Условия проявления

1. **Предпосылка.** Векторные представления построены по чувствительному содержимому и сохраняют связь с исходными записями или их атрибутами.
2. **Триггер.** Получатель векторов использует доступную модель или вспомогательные данные для реконструкции, связывания с источником, вывода атрибутов либо проверки присутствия записи.
3. **Разрешающий пробел.** Векторы хранятся, передаются или предоставляются третьим сторонам по более слабой политике, чем исходные данные, без подтверждённой оценки применимых способов раскрытия.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM09:2026** · Диагностический признак сбоя
>
> Если векторное представление не наследует класс чувствительности источника, его могут передать или экспортировать по более слабым правилам, хотя по нему могут быть восстановлены сведения об исходных данных.
>
> **Граница связи:** Не каждый вектор обратим; риск зависит от модели, вспомогательных сведений и доступа к представлению.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | В приложении используются векторное хранилище, поиск по сходству или векторные представления данных. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	231	48
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	232	69
AISAFE2-COMP-023 · Подсистема формирования базы знаний	233	86
AISAFE2-IMP-011 · Компрометация баз знаний	471	109
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	472	130
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	703	356
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	704	519
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	705	547
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	706	550
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	707	571

Недостаток · AISAFE2-WEAK-088

Отсутствие проверяемой оценки этического воздействия AI-системы

Краткое описание

Для сценария не заданы утверждённые этические критерии, затрагиваемые группы, измеримые подтверждения и владелец решения либо система не оценивается по ним до выпуска и при пересмотре.

Подробное описание

Этические принципы остаются декларацией: не определены недопустимые эффекты, затрагиваемые группы, метод и пороги оценки, владелец остаточного риска и канал обращений. Выпуск или существенное изменение допускаются без связанных с версией результатов, а эксплуатационный пересмотр не проверяет устойчивое неблагоприятное воздействие. Обнаружение пользователем после события не заменяет предварительную и периодическую оценку.

Происхождение

- Процесс

Классы недостатка

- AISAFE2-WC-036 · Соответствие и допустимое использование

Прямые причинные воздействия

AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы

Отсутствие проверяемой этической оценки позволяет выпустить точную версию, не проверенную по утверждённым нормам и затрагиваемым группам; эта версия затем фактически причиняет пользователю конкретный этически недопустимый вред.

Ограничение: Только когда отсутствие обязательной оценки позволяет выпустить точную версию, а затем зафиксировано конкретное воздействие этой версии на пользователя по заранее определённым нормам и для указанной группы.

AISAFE2-IMP-028 · Нарушение этических норм

Утверждённая этическая оценка обнаружила бы конкретный эффект и заблокировала, сузила или исправила бы выпуск, но версия выходит без оценки и фактически причиняет устойчивый запрещённый либо вредный эффект группе.

Ограничение: Только если показано, что обязательная оценка обнаружила бы эффект и изменила решение о выпуске, а версия без неё фактически причинила вред затронутой группе

Условные причинные пути

Это условная последовательность переходов, а не прямая связь недостатка с итоговым воздействием и не подтверждение произошедшего инцидента.

> **Путь 1**
>
> 1. Исходный недостаток: AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы
> 2. Промежуточное воздействие: AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы
> 3. Промежуточное состояние: Фактически реализованный этический вред получает публичную огласку и атрибутируется организации или продукту
> 4. Итоговое воздействие: AISAFE2-IMP-029 · Репутационный ущерб
>
> **Механизм переходов**
>
> 1. Отсутствующая утверждённая оценка, которая обнаружила бы точный эффект и изменила решение о выпуске, позволяет версии выйти и причинить фактический вред определённой группе.
> 2. Фактически реализованный этический вред публикуется для значимой аудитории и однозначно связывается с организацией либо продуктом.

3. Публичная атрибуция фактического вреда причинно снижает заранее выбранную и измеренную метрику доверия.

Условия применимости

- Утверждённая оценка обнаружила бы конкретный эффект и заблокировала, сузила либо исправила выпуск, но точная версия вышла без неё.
- Вред определённой группе фактически наступил, получил публичность и атрибуцию, после чего метрика доверия измеримо снизилась.

Исключения

- Одно отсутствие оценки или оценка, которая не изменила бы решение о выпуске, не образуют путь.
- Отсутствие фактического вреда, частное или неатрибутированное событие и публичность без измеримого результата исключаются.

Условия проявления

1. **Предпосылка.** Сценарий применения затрагивает пользователей или группы, для которых возможны недопустимое обращение или систематически неравное воздействие.
2. **Триггер.** Точная версия выпускается, существенно изменяется или продолжает эксплуатацию без целевой и состязательной оценки воздействия.
3. **Разрешающий пробел.** Проверяемые этические критерии и затрагиваемые группы не определены, а владелец критериев, периодический пересмотр и канал обращений затронутых лиц отсутствуют.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	234	67
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	473	98
AISAFE2-IMP-028 · Нарушение этических норм	474	135
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	708	415
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	709	488

Недостаток · AISAFE2-WEAK-089

Избыточные функции и полномочия инструментов агента

Краткое описание

Эффективный набор инструментов, операций и полномочий агента не сужается до задачи, роли, сессии и этапа, поэтому ошибка или инъекция достигает непредусмотренного действия.

Подробное описание

Агент чтения получает функции изменения или удаления, а узкая задача — универсальную оболочку, сетевой вызов, отправку сообщения либо платёжную операцию. Один широкий набор доступен на всех шагах с полномочиями, не выведенными из текущей роли и задачи. Поэтому последствия отклонения определяются полной возможностью подключённых инструментов, а не необходимым эффектом сценария.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-010 · Контроль доступа и авторизация

Прямые причинные воздействия

AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core

Избыточный состав функций и полномочий инструментов, доступных агенту, может привести к обходу ограничений и предусмотренной логики агента, потому что набор инструментов и операций не ограничен задачей, ролью и этапом выполнения, и ошибка модели или инъекция приводит к действиям, которых сценарий не предусматривал.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Избыточный состав функций и полномочий инструментов, доступных агенту, может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что инструменты подключены с максимальными правами и доступны на всех этапах выполнения, и вызов через агента выполняет операции с правами выше прав инициатора.

AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов

Избыточные полномочия дают агенту вызвать тарифицируемые инструменты и внешние сервисы, не нужные задаче; ошибка или инъекция приводит к подтверждённым начислениям вне утверждённого бюджета.

Ограничение: Только когда цикл или действие содержит тарифицируемый инструмент либо внешний сервис и создаёт подтверждённое начисление вне утверждённого бюджета.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Избыточный состав функций и полномочий инструментов, доступных агенту, может привести к утечке чувствительных данных в недоверенный контур, потому что агенту доступны инструменты отправки данных во внешние каналы независимо от задачи, и ошибка модели или инъекция выводит данные наружу.

AISAFE2-IMP-031 · Утрата датасета в результате атаки

Внедрённая атакующим инструкция использует избыточно доступную разрушительную операцию для уничтожения точного набора данных и всех пригодных доверенных копий или путей воспроизводимого восстановления.

Ограничение: Только если избыточно доступный инструмент использован атакой для уничтожения точного датасета и всех пригодных копий или путей восстановления; невосстановимость подтверждена

AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки

Внедрённая атакующим инструкция использует избыточно доступную разрушительную операцию для уничтожения весов, базы знаний, индекса либо критичной конфигурации и всех пригодных копий или путей воспроизводимого восстановления.

Ограничение: Только если избыточно доступный инструмент использован атакой для уничтожения точного критичного артефакта, кроме датасета, и всех пригодных путей восстановления; невосстановимость подтверждена

Условия проявления

1. **Предпосылка.** Агенту подключены функции изменения, удаления, передачи, оплаты или универсальные исполнители, не требуемые текущей задачей.
2. **Триггер.** Выбор действий зависит от модели, на которую могут повлиять ошибка или внедрённая инструкция.
3. **Разрешающий пробел.** Состав инструментов и операций не сокращается по задаче, роли пользователя, сессии и этапу выполнения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

> **C9.3.1** · Диагностический признак сбоя
>
> Избыточные функции и полномочия инструмента расширяют доступ за пределы задачи, роли и этапа агента.
>
> **Граница связи:** Связь охватывает избыточные полномочия исполняемого инструмента; доступность лишнего инструмента в списке сама по себе не подтверждает нарушение изоляции.
>
Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает изоляцию инструмента с минимальными полномочиями по его риску.
Условие связи с карточкой	Инструмент исполняется с полномочиями сверх необходимых для его задачи и разрешённой области.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	235	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	236	44
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	237	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	238	77

Карточка	Связи	Стр.
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	475	101
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	476	118
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	477	121
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	478	130
AISAFE2-IMP-031 · Утрата датасета в результате атаки	479	137
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	480	141
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	710	375

Недостаток · AISAFE2-WEAK-090

Недостаточная изоляция исполнения инструментов и сгенерированного кода

Краткое описание

Недоверенный инструмент, обработчик или сгенерированный код исполняется с доступом к файловой системе, сети, секретам, соседним процессам или ресурсам сверх явно предоставленной области.

Подробное описание

Интерпретатор кода, браузер, парсер документа, сервер инструмента или подагент наследует хост, переменные окружения, внутреннюю сеть и полномочия ядра агента без отдельных лимитов. Инъекция, вредоносный файл или уязвимость исполнителя поэтому выходит за локальный процесс, раскрывает секреты и данные, достигает внутренних сервисов либо исчерпывает общий ресурс.

Происхождение

- Реализация
- Проектирование

Классы недостатка

- AISAFE2-WC-043 · Изоляция среды исполнения

Прямые причинные воздействия

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к несанкционированному чтению или изменению хранимых данных, потому что скомпрометированный инструмент на общем хосте получает доступ к файловой системе, секретам и данным других арендаторов с полномочиями ядра агента.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что исполнители работают с полномочиями ядра агента и общего хоста, и компрометация одного компонента (запуск произвольного кода) распространяется на функциональность всей системы.

AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к чрезмерному потреблению вычислительных ресурсов системы, потому что на исполнителей не наложены ограничения вычислений, памяти и времени выполнения, и скомпрометированный или сбойный компонент исчерпывает ресурсы общего хоста.

AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к утечке чувствительных данных в недоверенный контур, потому что скомпрометированный исполнитель, обладая сетевым доступом хоста, выводит данные во внешние каналы за пределы контролируемой зоны.

Условия проявления

1. **Предпосылка.** Их процессы выполняются на общем хосте с полномочиями ядра агента: доступом к секретам, файловой системе, внутренней сети и ресурсам других арендаторов.
2. **Триггер.** Инструменты, интерпретаторы кода, браузеры и обработчики документов работают с недоверенным содержимым или исполняют сформированный моделью код.
3. **Разрешающий пробел.** Полномочия, сетевой выход и потребление ресурсов этих процессов не ограничены, и компрометация одного из них не локализуется.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

> **C9.3.1** · Диагностический признак сбоя
>
> Исполнение инструмента с доступом к файлам, сети и секретам сверх разрешённой области показывает недостаточную изоляцию.
>
> **Граница связи:** Уровень изоляции выбирают по риску конкретного инструмента; карточка не доказывает эксплуатацию.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает изоляцию инструмента с минимальными полномочиями по его риску.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	239	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	240	58
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	481	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	482	118
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	483	123
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	484	130
AISAFE2-MIT-017 · Изолированное исполнение инструментов	711	379
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	712	438

Недостаток · AISAFE2-WEAK-091

Исполняемая десериализация артефактов модели

Краткое описание

Загрузчик принимает формат артефакта модели, десериализация или сопровождающий код которого может исполнить произвольные операции с полномочиями процесса обучения или инференса.

Подробное описание

Артефакт рассматривается как данные, хотя его формат допускает исполняемую десериализацию, пользовательские слои или обработчики. Загрузка происходит без обязательного безопасного формата и до изоляции кода. Подписанный и полученный от доверенного источника артефакт всё равно сохраняет исполняемость: целостность подтверждает байты, но не превращает активный формат в данные.

Происхождение

- Реализация
- Проектирование

Классы недостатка

- AISAFE2-WC-032 · Цепочка поставки, недоверенная среда и физическая защита

Прямые причинные воздействия

AISAFE2-IMP-012 · Кража модели

Использование форматов артефактов модели, допускающих исполнение кода при загрузке, может привести к краже модели, потому что подменённый или вредоносный артефакт исполняет код при загрузке с полномочиями сервиса, и из инфраструктуры похищаются модели.

AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)

Загрузчик исполняет код, встроенный в артефакт модели, непосредственно при десериализации; код действует с полномочиями сервиса и успешно читает или изменяет достижимый хранимый объект вне полномочий атакующего.

AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)

Загрузчик исполняет код, встроенный в артефакт модели, непосредственно при десериализации; код действует с полномочиями сервиса и вызывает достижимую инфраструктурную функцию вне полномочий атакующего.

Условия проявления

1. **Предпосылка.** Артефакт модели принимается в формате с исполняемой десериализацией либо с сопровождающим пользовательским кодом.
2. **Триггер.** Источник артефакта может быть подменён или скомпрометирован: поставщик, хранилище, канал поставки.
3. **Разрешающий пробел.** Загрузка выполняется с полномочиями сервиса, без проверки содержимого и без изолированной среды.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM04:2026** · Диагностический признак сбоя
>
> Артефакт модели с исполняемой десериализацией может выполнить код поставщика при загрузке и нарушить границу между данными и программой.
>
> **Граница связи:** Исполняемый формат опасен и у подписанного артефакта; карточка не доказывает наличие вредоносной нагрузки или других рисков цепочки поставки.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	241	37
AISAFE2-COMP-009 · Конвейер обучения и дообучения	242	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	243	56
AISAFE2-IMP-012 · Кража модели	485	111
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	486	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	487	118
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	713	359

Недостаток · AISAFE2-WEAK-092

Недостаточно ограниченные делегированные полномочия пользователя у агента

Краткое описание

Делегированное агенту полномочие пользователя не ограничено получателем, аудиторией, операцией, ресурсом, задачей и сроком либо не может быть своевременно отозвано.

Подробное описание

Агент или коннектор получает широкое либо долгоживущее полномочие внешнего сервиса и повторно использует его между задачами и получателями. Ограничение существует только в инструкции модели, а целевая сторона не проверяет аудиторию, ресурс, операцию и срок. Компрометация или ошибочный шаг поэтому достигает всего объёма корректно выданного пользовательского права.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-011 · Передача идентичности и A2A-доверие

Условия проявления

1. **Предпосылка.** Пользователь передал агенту учётные данные внешних сервисов или платёжные средства для действий от его имени.
2. **Триггер.** Операции от имени пользователя инициируются без его участия: из-за ошибки модели, внедрённой инструкции или компрометации агента.
3. **Разрешающий пробел.** Область действия, лимиты и срок не сужены до задачи, использование полномочия пользователю не видно, а быстрый отзыв недоступен.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI03** · Диагностический признак сбоя
>
> Недостаточно ограниченное делегирование может передать агенту больше прав пользователя, чем нужно для задачи.
>
> **Граница связи:** Избыточное полномочие создаёт условие риска; злоупотребление, обход проверки и причинённый вред подтверждаются отдельно.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент использует собственную, делегированную или унаследованную идентичность и связанные права.
Условие связи с карточкой	Агент наследует права пользователя без ограничений по задаче, получателю, ресурсу или сроку.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	244	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	245	44
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	246	61
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	714	434

Недостаток · AISAFE2-WEAK-093

Неверная атрибуция тарифицируемого потребления между арендаторами

Краткое описание

Тарифицируемое потребление разделяемой среды связывается не с создавшим его арендатором, поэтому чужая нагрузка начисляется добросовестному плательщику.

Подробное описание

Пробел находится в цепочке измерения, сквозного сопоставления арендатора с операцией и формирования начисления. Метрика теряет фактического инициатора, вынужденные повторы или сорванные чужой перегрузкой обращения учитываются на другой учётной записи, а начисление нельзя сверить с создавшим потребление субъектом и точной операцией.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-022 · Пользовательский контроль тарифицируемых операций

Прямые причинные воздействия

AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы

Операция одного арендатора ошибочно получает платёжную идентичность другого; её измеренное потребление начисляется другому плательщику сверх заранее показанной или подтверждённой суммы.

Ограничение: Только если установлены фактический инициатор, ошибочно начисленный плательщик, точная операция и реальное списание выше согласованной суммы без отдельного согласия плательщика.

Условия проявления

1. **Предпосылка.** Несколько арендаторов используют общий тарифицируемый контур, где расход должен сохранять связь с фактическим инициатором и операцией.
2. **Триггер.** Чужая нагрузка, повтор или сорванное обращение создают измеренное потребление, которое попадает в расчёт начисления.
3. **Разрешающий пробел.** Связь арендатора с операцией теряется или подменяется до формирования начисления, поэтому потребление относится другому плательщику без проверяемого подтверждения.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	247	72
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	488	125
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	715	333

Недостаток · AISAFE2-WEAK-094

Небезопасное хранение и распространение делегированных учётных данных пользователя

Краткое описание

Сырые пользовательские токены, ключи или сессии сохраняются за пределами нужной задачи либо попадают в модель, журналы или нецелевые инструменты.

Подробное описание

Агент или коннектор принимает делегированное учётное значение внешнего сервиса, но не удерживает его в минимально необходимой области обработки. Значение хранится дольше задачи, копируется между сессиями, передаётся модели и несвязанным инструментам либо записывается в журналы, резервные копии и диагностические данные.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-028 · Управление ключами и секретами

Условия проявления

1. **Предпосылка.** Пользователь передаёт агенту или коннектору исходный токен, ключ или сеанс внешнего сервиса для действия от своего имени.
2. **Триггер.** Учётное значение сохраняется, копируется или маршрутизируется после нужного обмена либо достигает модели, журнала или нецелевого инструмента.
3. **Разрешающий пробел.** Не задана минимально необходимая область обработки по задаче, получателю, сроку, месту хранения и разрешённым потребителям исходного значения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI03** · Диагностический признак сбоя
>
> Попадание пользовательского токена в память, журнал или нецелевой инструмент создаёт путь для повторного использования полномочий агента.
>
> **Граница связи:** Нужно установить доступность токена и возможность его применения; компрометация полномочий не следует из одного факта хранения.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент использует собственную, делегированную или унаследованную идентичность и связанные права.
Условие связи с карточкой	Делегированные токены или сессии сохраняются в контексте, журнале либо не относящемся к задаче инструменте.

> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	248	61
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	716	554

Недостаток · AISAFE2-WEAK-095

Недостаточно ограниченные делегированные платёжные полномочия

Краткое описание

Доступное агенту платёжное полномочие не ограничено задачей, получателем, суммой, частотой и сроком и позволяет выполнять расходы сверх намерения пользователя.

Подробное описание

Агент или коннектор удерживает платёжное средство либо право на денежную операцию без принудительно проверяемого ограничения по получателю, назначению, сумме, частоте и времени. Ограничение остаётся только в инструкции модели или интерфейсе, поэтому ошибочный или внедрённый шаг достигает полного платёжного полномочия пользователя.

Происхождение

- Проектирование
- Реализация

Классы недостатка

- AISAFE2-WC-011 · Передача идентичности и A2A-доверие

Прямые причинные воздействия

AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы

Делегированное платёжное полномочие без лимитов суммы и частоты, привязанных к задаче, позволяет агенту фактически выполнить списания выше показанной или подтверждённой пользователем границы.

Ограничение: Только при фактически исполненном списании выше заранее показанной либо подтверждённой суммы или лимита и при отсутствии отдельного подтверждения точных параметров. Наличие платёжного средства без списания недостаточно.

AISAFE2-IMP-025 · Прямые финансовые потери

Неограниченное по сумме и частоте делегированное платёжное полномочие позволяет агенту выполнить денежную операцию от имени пользователя, после которой остаётся точный атрибутированный невозвратный чистый убыток.

Ограничение: Только если операция без нужного платёжного лимита исполнена и оставила подтверждённый чистый убыток после возвратов и оспаривания; потенциальные суммы не учитываются

Условия проявления

1. **Предпосылка.** Пользователь делегирует агенту платёжное средство или право инициировать денежные операции от своего имени.
2. **Триггер.** Ошибка модели, внедрённая инструкция или компрометация инициируют одну или несколько платёжных операций вне намерения пользователя.
3. **Разрешающий пробел.** Получатель, назначение, сумма, частота и срок не ограничены принудительно проверяемой политикой до исполнения каждой операции.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	249	44
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	489	125
AISAFE2-IMP-025 · Прямые финансовые потери	490	133
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	717	434

Меры защиты

Мера защиты · AISAFE2-MIT-001

Проверка входных данных по схемам и правилам

Краткое описание

Контроль проверяет вход по утверждённой схеме и отклоняет недопустимую структуру до передачи модели, инструменту или бизнес-логике.

Область применения

- Общая безопасность

Механизмы

- Валидация

Защитный механизм

Валидатор получает пользовательские запросы, параметры API, данные от другого агента, результат модели и аргументы инструмента. Он проверяет структуру, размер, типы, перечисления, допустимые сочетания и класс операции, после чего формирует нормализованный объект и решение «разрешить», «отклонить» или «передать на отдельную проверку».

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация определяет охватываемые синхронные, асинхронные, пользовательские и машинные интерфейсы по происхождению входа и последствиям ошибки. Для высокорисковых путей предпочтительны строгий контракт и отказ от обработки при нарушении; для низкорисковых могут быть достаточны нормализация или безопасный маршрут. Мера не проверяет истинность текста, полномочия субъекта и допустимость цели операции; для этого применяются отдельные средства авторизации и политики действий.

Предпосылки

- Нужны инвентарь всех охватываемых входных интерфейсов, версионируемый контракт для каждого из них и принудительная проверка до модели, инструмента или бизнес-логики.
- Для неизвестного поля, неверного типа и запрещенного сочетания заранее задается однозначное решение.

Реализация

Точки принудительного контроля выбираются по архитектуре и возможному эффекту: на внешнем API, перед сборкой контекста, перед вызовом инструмента или перед передачей в последующий компонент. Все пути, включенные в область меры, должны обеспечивать сопоставимую проверку.

Проверка

Проверкой должны быть охвачены все интерфейсы, включенные в область меры. Негативный набор подтверждает отклонение полей, значений и сочетаний, которые выбранный контракт определяет как недопустимые. Отдельно измеряются доля ложных отказов и задержка; до выпуска их пороги фиксируются в версии правил проверки. Проверяются штатный сценарий, фаззинг схемы, попытка обхода через служебный путь, отказ хранилища схем и повышенная нагрузка. Мера неэффективна, если найден путь без заявленной проверки либо невалидный по выбранному контракту объект достигает модели или инструмента.

Негативный тест

На каждом охваченном интерфейсе отдельно передают объекты с полем, запрещённым выбранным контрактом, с отсутствующим обязательным полем, неверным типом и запрещённым сочетанием значений. Каждый вариант повторяют через служебный обход: ни один объект, который контракт считает недопустимым, не должен достичь модели или инструмента.

Подтверждения

Подтверждения включают инвентарь охваченных интерфейсов, версии схем и правил, конфигурацию точек контроля, отчет негативных и нагрузочных тестов, журналы решений и метрики ложных отказов и задержки.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При отказе проверки высокорисковая операция блокируется или переводится на заранее определённый безопасный маршрут. Низкорисковый запрос может перейти в ограниченный режим, если он не интерпретирует объект за пределами подтверждённой части контракта; продолжение с непроверенным объектом запрещено.

Исправление и повторная проверка

Исправляют схему, правило нормализации или точку принудительной проверки, через которую объект с неизвестным полем дошёл до модели или инструмента. Затем тем же объектом повторяют вызов через исходный и обходной маршруты: оба должны завершиться отказом до интерпретации; сохраняют версию схемы, решение валидатора и трассировку границы.

Побочные эффекты

Жесткие правила могут блокировать легитимные новые сценарии и создавать нагрузку на поддержку.

Остаточный риск

Семантически вредоносный запрос может полностью соответствовать схеме. После проверки остается возможность злоупотребления допустимыми значениями; ее ограничивают авторизация, политика действий и мониторинг.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex API Gateway

Прямая поддержка. Валидатор API Gateway проверяет тело, параметры HTTP-запроса и заголовки по схемам OpenAPI до передачи запроса интеграции.

Ограничение: Валидатор проверяет только HTTP-запросы через настроенный шлюз и не оценивает смысл допустимых значений.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-APP-WAF1

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Проверяет структуру, размер, типы, обязательные поля и допустимые сочетания до передачи данных модели, инструменту или бизнес-логике.

Ограничение: Допустимые кодировки и значения задаются в контракте каждого интерфейса; соответствие схеме не доказывает смысловую безопасность запроса.

Условие применения: На пользовательских и машинных входах ИИ-приложения в Yandex Cloud до модельного вызова или другого действия.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет формат, размер и структуру входного запроса.

Ограничение: Не аутентифицирует внешний вход и не исключает обход публичного пути.

Условие применения: На публичном или партнёрском входе ИИ-приложения до передачи запроса внутреннему сервису.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Приказ ФСТЭК России от 11.04.2025 № 117 Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных

информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений

Редакция: Приказ № 117 от 11.04.2025 с изменениями по приказу № 137 от 08.05.2026

п. 61(а), шаблон запроса · Поддерживает часть требования

Проверка входа по утверждённой схеме помогает отклонять запросы, не соответствующие заданному шаблону.

Граница связи: Шаблон и перечень охваченных интерфейсов утверждает оператор; проверка структуры не устанавливает допустимость смысла запроса.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	Пользователи охваченной приказом № 117 системы выполняют свои обязанности через AI-сервис в формате строго заданных шаблонов запросов и ответов.
Условие связи с карточкой	Для запросов установлен шаблон, соответствие которому проверяется до обращения к AI-сервису.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 61 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	500	100
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	662	246
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	855	35
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	856	44
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	857	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	858	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	859	82
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1196	494
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1203	328
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1292	429
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1328	462

Мера защиты · AISAFE2-MIT-002

Проверка и безопасная обработка AI-вывода

Краткое описание

Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Механизмы

- Валидация

Защитный механизм

Контроль получает необработанный вывод, назначение приемника и контекст действия. Он разбирает структуру, проверяет допустимые поля и эффекты, выполняет контекстное экранирование и формирует безопасный результат либо отказ.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация определяет глубину обработки по типу приемника и последствиям ошибки. Для активного исполнения и решений с существенными последствиями предпочтительны строгая проверка и отдельный маршрут независимой проверки решения; для низкорискового текстового представления могут быть достаточны безопасное экранирование и явное обозначение ограничений. Мера не подтверждает фактическую правильность, полноту и правомерность содержания и не контролирует ручное исполнение пользователем вне системы.

Предпосылки

- Нужны инвентарь всех активных приемников AI-вывода и отдельный контракт безопасного разбора, кодирования или исполнения для каждого контекста.
- Должен отсутствовать прямой путь от необработанного вывода к средству отображения, интерпретатору, инструменту или исполнителю.

Реализация

Точки контроля размещаются непосредственно перед компонентом отображения, интерпретатором, хранилищем, API и преобразованием ответа в действие. Повторное использование сохраненного вывода требует новой проверки по правилам нового приемника.

Проверка

Все приемники, включенные в область меры, должны иметь подходящий контроль; прямой путь от модели к активному исполнению недопустим. Мера неэффективна при любом успешном исполнении заблокированного содержимого или наличии непроверенного активного приемника в заявленной области.

Негативный тест

Передают корректный вывод и варианты, нарушающие контракт конкретного приёмника, включая как неисчерпывающие примеры лишнее поле, XSS/HTML для контекста разметки или командную инъекцию для интерпретатора; затем повторяют вредоносный вариант через прямой вызов приёмника и при отказе валидатора. Корректный вывод проходит штатно, остальные варианты завершаются отказом до интерпретации и без внешнего эффекта.

Подтверждения

Подтверждениями служат перечень приемников и их контрактов, версии правил разбора и экранирования, результаты сквозных негативных тестов, трассировки решений и журналы блокирования активного содержимого.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Активное отображение и исполнение блокируются. Допускается показать безопасный текстовый результат с явной отметкой о невозможности проверки, если это не запускает действие и разрешено владельцем риска.

Исправление и повторная проверка

Устраняют непроверенный активный приемник либо исправляют разбор и контекстное экранирование, позволившие тестовой инъекции исполниться. Тот же AI-вывод повторно подают через основной и обходной пути к средству отображения или исполнителю: активное содержимое не исполняется, а подтверждением служат результат теста и журнал блокирования.

Побочные эффекты

Строгая обработка может портить полезную разметку и увеличивать задержку.

Остаточный риск

Эффективность проверки и экранирования ограничена полнотой контракта конкретного приёмника. Формально валидный вывод всё ещё может быть ложным, вредным или манипулятивным и привести к ошибочному решению пользователя.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex API Gateway

Частичная поддержка. API Gateway может проверять HTTP-ответы по схеме до их передачи клиенту.

Ограничение: Проверяется схема HTTP-ответа; контекстное экранирование, фактическую правильность и допустимость действия обеспечивает приложение.

Yandex Cloud AI Studio

Частичная поддержка. Responses API поддерживает структурированный вывод по JSON Schema и строгий режим соответствия схеме.

Ограничение: JSON Schema ограничивает форму ответа; истинность, безопасность для получателя и допустимость исполнения проверяются отдельно.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CONTENT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Разбирает и проверяет вывод для конкретного получателя, экранирует контекст и блокирует недопустимое отображение, сохранение или исполнение.

Ограничение: Проверка вывода и экранирование не устанавливают полномочия вызывающего субъекта на целевой объект и действие; такую авторизацию выполняют отдельно.

Условие применения: При отображении, сохранении или исполнении вывода модели; для инертного текста сохраняется требование безопасного отображения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Приказ ФСТЭК России от 11.04.2025 № 117 Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений

Редакция: Приказ № 117 от 11.04.2025 с изменениями по приказу № 137 от 08.05.2026

п. 61(а), шаблон ответа · Поддерживает часть требования

Проверка вывода для конкретного получателя позволяет отклонять ответы вне установленного шаблона.

Граница связи: Оператор задаёт шаблон и охват получателей; соответствие формату не подтверждает достоверность содержания.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	Пользователи охваченной приказом № 117 системы выполняют свои обязанности через AI-сервис в формате строго заданных шаблонов запросов и ответов.
Условие связи с карточкой	Оператор установил шаблон ответа для охваченного получателя; проверка выполняется до выдачи результата.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 61 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

п. 61(б), формат ответа · Поддерживает часть требования

Проверка AI-вывода может ограничить ответ заданным форматом для охваченного получателя.

Граница связи: Карта не определяет разрешённые темы ответа и не подтверждает правильность содержания; эти правила и остальные получатели проверяются отдельно.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	Пользователи охваченной приказом № 117 системы выполняют свои обязанности через AI-сервис в свободной текстовой форме.
Условие связи с карточкой	Для ответа оператор установил формат; связь охватывает контроль этого формата, а не допустимых тематик.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 61 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM10:2026 · Поддерживает часть требования

Проверка структуры и экранирование вывода для конкретного получателя снижают риск исполнить недоверенный результат модели как команду или разметку.

Граница связи: Защита действует на путях передачи вывода, где применяется контроль; запросы к внешним адресам, права на действие и ручное использование вывода проверяются отдельно.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется получателем, инструментом или приложением для последующего действия.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	501	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	508	101
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	573	146
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	575	148
AISAFE2-WEAK-060 · Отсутствие фильтрации сообщений об ошибках на внешней границе	669	254
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	690	286
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	860	35

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	861	37
AISAFE2-COMP-003 · Управление контекстом общения с моделью	862	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	863	44
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	864	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	865	77
AISAFE2-COMP-024 · Артефакты работы AI-системы	866	88
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1197, 1350	488
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	1209	331
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1253	389

Мера защиты · AISAFE2-MIT-003

Разделение доверенных инструкций и недоверенного контента при сборке контекста

Краткое описание

Сборщик контекста сохраняет происхождение, назначение и уровень доверия каждого фрагмента и не позволяет содержимому с более низким уровнем доверия перейти в канал инструкций с более высоким приоритетом.

Механизмы

- Изоляция
- Управление конфигурацией

Защитный механизм

Каждый значимый фрагмент получает тип, источник, уровень доверия и назначение. Сборка может быть централизованной или распределенной, если системные правила помещаются только в предназначенный для них канал, недоверенные фрагменты передаются как данные, а метки происхождения сохраняются после суммаризации, RAG, памяти и вызова инструмента.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик

Применимость

Глубина маркировки и число точек контроля определяются архитектурой и последствиями потери границы доверия. Мера не создает криптографической изоляции внутри модели и не гарантирует, что модель не последует смыслу вредоносного текста.

Предпосылки

- Обязательная предпосылка выбранного контроля — охват всех заявленных путей сборки контекста с сохраняемыми метками происхождения и сопоставимыми границами доверия.

Реализация

Любой путь сборки в заявленной области должен обеспечивать сопоставимые гарантии; самостоятельная конкатенация, теряющая границы доверия, означает обход контроля. Решения о доступе и действиях остаются за детерминированными механизмами вне модели.

Проверка

Трассируют каждый заявленный путь сборки, подставляя доверенную инструкцию, недоверенный фрагмент и вариант с потерянной меткой после каждого преобразования; сравнивают тип, источник и канал фрагмента на входе модели и перед действием. Мера неэффективна, если приложение формирует контекст вне контролируемого пути или недоверенный текст меняет полномочия без внешнего решения.

Негативный тест

Собирают контрольный контекст из доверенной инструкции и недоверенного фрагмента с прямым или косвенным внедрением вредоносных инструкций, затем теряют метку после суммаризации и повторяют через обход сборщика. Штатный размеченный контекст должен сохраниться, а недоверенный фрагмент не должен попасть в привилегированный канал или изменить полномочия ни при потере метки, ни через обходной путь.

Подтверждения

Подтверждения включают схему типов и происхождения, конфигурацию сборщика, выборочные трассировки контекста, инвентарь путей сборки и результаты тестов прямого, косвенного и многошагового обхода.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Фрагмент без типа или происхождения исключается. Если невозможно собрать минимальный доверенный контекст, высокорисковая операция блокируется; допустим ответ без RAG, памяти и инструментов с явным ограничением функции.

Исправление и повторная проверка

Возвращают обязательную метку происхождения и тип фрагмента либо закрывают путь, поместивший недоверенный текст в привилегированный канал. Повторяют то же прямое или косвенное внедрение вредоносных инструкций через исходную цепочку сборки: фрагмент остаётся данными или исключается; сохраняют трассировку собранного контекста без пользовательских секретов.

Побочные эффекты

Дополнительная структура сокращает доступный контекст и может ухудшать качество.

Остаточный риск

Эффективность границы зависит от полного охвата сборщиков, преобразований и межагентских передач. Даже при корректной разметке модель может интерпретировать данные как инструкцию; переход такого влияния в значимое действие должен блокироваться независимой политикой.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud AI Studio

Частичная поддержка. API разделяет сообщения по ролям, включая системную инструкцию и пользовательский ввод, что позволяет не смешивать их в одной строке.

Ограничение: Роли являются инструкциями модели, а не границей изоляции. Они не сохраняют автоматически происхождение RAG-фрагментов и не подтверждают, что приложение не конкатенирует недоверенный текст с системной инструкцией.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT-GOAL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Не позволяет содержимому пользователя, RAG, инструмента или агента стать доверенными управляющими инструкциями.

Ограничение: Нет типизированного версионируемого объекта цели и авторизации её изменения.

Условие применения: Когда цель многошагового агента может испытывать влияние пользовательского текста, RAG, инструмента или другого агента.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-APP-INJECT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Сохраняет происхождение и доверие фрагментов и не позволяет внешнему содержимому стать системной инструкцией.

Ограничение: Разделение доверия при сборке контекста не ограничивает само последующее действие агента и не охватывает все пути выбора и вызова инструмента.

Условие применения: Когда модель получает RAG-фрагменты, файлы, веб-страницы, сообщения или результаты инструментов из недоверенного источника.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM01:2026 · Поддерживает часть требования

Разделение инструкций и внешних данных при сборке контекста снижает шанс повысить внедрённый фрагмент до уровня управляющего правила.

Граница связи: Разметка фрагментов не создаёт непроницаемой границы внутри модели; преобразования и действия после ответа требуют отдельного контроля.

Субъект	Аналитик угроз AI-приложения
Условие применения	Недоверенное содержимое попадает в контекст модели и может пересечь границу между данными и инструкциями.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	502	100
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	517	104
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	578	154
AISAFE2-COMP-003 · Управление контекстом общения с моделью	867	40
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	868	42
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1198, 1204	328
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	1199	339
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	1200	342
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1201	494
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	1202	561
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	1242	375
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1293	429
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1354	497

Мера защиты · AISAFE2-MIT-004

Выявление и обработка внедрения вредоносных инструкций и обхода ограничений

Краткое описание

Версионируемый детектор выявляет признаки внедрения вредоносных инструкций и обхода ограничений и передаёт результат в заданную политику реакции.

Механизмы

- Обнаружение
- Валидация

Защитный механизм

Детектор получает доступные признаки текста, происхождения, языка, контекста сессии и класса последующего действия. В зависимости от архитектуры проверка может выполняться до передачи модели, после добавления RAG-контента или перед высокорисковым действием.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Целесообразность определяется оценкой риска, наблюдаемыми каналами внедрения вредоносных инструкций и обхода ограничений и доступностью достаточного качества обнаружения. Мера не заменяет разделение контекста, авторизацию, изоляцию инструментов и контроль исходящих данных. Если риск приемлем без вероятностного детектора при наличии детерминированных границ, мера может не применяться.

Предпосылки

- Нужны зафиксированная модель угроз, репрезентативный набор атак и чистых запросов для охватываемых языков и каналов и утвержденные пороги полноты, ложных срабатываний и задержки.
- Для недоступности детектора и потери сигнала заранее определяется безопасный маршрут.

Реализация

Решение детектора является сигналом для политики и мониторинга, но не выдает полномочия. Все входы, включенные в область меры, должны учитываться при проверке возможного обхода.

Проверка

До выпуска в версии политики для каждого канала фиксируются минимальная полнота, допустимая доля ложных срабатываний и максимальная задержка. Проверка должна охватывать все применимые точки контроля. Мера работает частично при выполнении порогов

только для части языков или каналов. Мера неэффективна при превышении порога пропусков, отсутствии версии либо потере сигнала без обнаружения. Приёмочные значения полноты, ложных срабатываний и задержки проверяются на зафиксированном оценочном наборе для каждого канала.

Негативный тест

Подают чистый запрос и тот же сценарий с многоязычным либо обфусцированным внедрением вредоносных инструкций, затем повторяют атаку при недоступности детектора и под расчётной нагрузкой. Чистый запрос должен остаться ниже порога, зафиксированного для его канала в версии политики, атака — получить предусмотренный маршрут до высокорискового действия, а отказ детектора — перевести запрос в заранее заданный безопасный режим.

Подтверждения

Подтверждениями являются точная версия детектора и политики, перечень охваченных каналов и языков, версия оценочного набора, метрики полноты, ложных срабатываний и задержки, журналы решений и результат проверки резервного режима.

Поведение при отказе

Безопасная деградация

Сценарий отказа

При недоступности детектора или потере его сигнала высокорисковый запрос переводится на резервный детектор, ручную проверку или в режим без инструментов и чувствительных данных. Низкорисковый диалог может продолжиться только по решению, зафиксированному для этого маршрута, с уведомлением и усиленным мониторингом.

Исправление и повторная проверка

Перенастраивают или заменяют версию детектора и правило маршрутизации для пропущенного класса атаки, не ослабляя детерминированные границы. На зафиксированном наборе повторяют исходную попытку обхода ограничений и чистые запросы: метрики возвращаются в диапазон, зафиксированный для канала в версии политики, а сигнал доходит до требуемой реакции; сохраняют версию набора, порог и отчёт сравнения.

Побочные эффекты

Детектор способен блокировать легитимный контент, добавляет задержку и требует регулярного обновления оценочных наборов.

Остаточный риск

Детектор вероятностный и может быть обойден новой формой атаки. После проверки остается риск нераспознанной инструкции; поэтому детектор не является самостоятельной авторизацией или гарантией предотвращения.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CONTENT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Обнаруживает внедрение инструкций и обход ограничений и применяет реакцию политики.

Ограничение: Детектор охватывает выбранные классы атак; общую политику категорий входного и выходного содержимого и действия для остальных категорий задают отдельно.

Условие применения: На входном или выходном маршруте, где утверждённая политика содержимого требует выявлять внедрение инструкций или обход ограничений.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM01:2026 · Поддерживает часть требования

Детектор может обнаружить заданные признаки внедрения инструкции и передать сигнал политике реакции до опасного продолжения.

Граница связи: Новая кодировка или формулировка может пройти детектор; его результат не заменяет авторизацию действия и границу доверия.

Субъект	Аналитик угроз AI-приложения
Условие применения	Недоверенное содержимое попадает в контекст модели и может пересечь границу между данными и инструкциями.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	503	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	509	101
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	525	107
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	556	130
AISAFE2-COMP-003 · Управление контекстом общения с моделью	869	40
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	870	77
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1198, 1204	324
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	1203	316
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	1205	371
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	1206	438
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1207	494

Карточка	Связи	Стр.
AISAFE2-MIT-056 · Исполнение версионируемой политики допустимости содержимого	1208	529
AISAFE2-MIT-017 · Изолированное исполнение инструментов	1245	379
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1268	405
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1294	429

Мера защиты · AISAFE2-MIT-005

Отображение происхождения и статуса подтверждения AI-вывода

Краткое описание

Интерфейс связывает существенный вывод с фактически использованными источниками и явно показывает статус независимого подтверждения.

Механизмы

- Наблюдаемость
- Человеческий контроль

Защитный механизм

Механизм представления и маршрутизации получает ответ, использованные фрагменты, сведения о модели и результат проверки. Он связывает утверждения с источниками, обозначает неподтвержденные части, добавляет предупреждение либо направляет ответ на независимую проверку для решений с существенными последствиями.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Мера обычно применима, когда AI-вывод показывается человеку или включается в материал, на который человек может опереться. Она не делает источник истинным, не оценивает его юридическую силу, не предотвращает вредный вывод и не заменяет независимую проверку «Независимая проверка решений с существенными последствиями».

Предпосылки

- Нужна фактическая трассировка использованных извлечённых фрагментов, ответов инструментов или иных источников до каждого представления и экспорта результата.
- Владелец решения задает случаи, где отсутствие подтвержденного происхождения блокирует действие или публикацию.

Реализация

Точки контроля выбираются перед отображением, экспортом и использованием ответа в рабочем процессе. Предупреждение снижает ошибочное доверие, но само по себе не разрешает решение или действие.

Проверка

Для ответа и его экспорта сопоставляют показанные источники и статус подтверждения с фактическим извлечением и трассировкой вызовов инструментов, затем повторяют проверку без одного источника и при отказе трассировки. Мера неэффективна при ложной атрибуции или автоматическом использовании неподтвержденного вывода там, где требуется проверка.

Негативный тест

Формируют ответ из реально использованного источника и варианты без источника, с устаревшей или вымышленной ссылкой, затем экспортируют их при отказе трассировки. Контрольный ответ должен показывать источник, совпадающий с фактическим извлечением и трассировкой вызовов инструментов; остальные варианты не должны отображаться как подтвержденные или закрывать значимое решение без независимой проверки.

Подтверждения

Подтверждения включают правила представления, сквозные тесты отображения и экспорта, фактические трассировки извлечения или вызова инструмента, выборку сопоставлений утверждений с источниками и журнал обязательной эскалации.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Ответ помечается как неподтвержденный. Значимое действие или публикация, для которых происхождение является условием решения, блокируются либо передаются человеку; в остальных сценариях допустимость ограниченного представления определяется владельцем риска. Выдавать отсутствие происхождения за подтвержденную атрибуцию нельзя.

Исправление и повторная проверка

Исправляют потерянную трассировку либо представление, которое показало вымышленный или неиспользованный источник как подтвержденный. Повторяют исходный ответ во всех охваченных представлениях и экспортах: ссылка соответствует фактически использованному источнику, а отсутствие трассировки явно обозначено; сохраняют сопоставление ответа с событием извлечения или вызова инструмента.

Побочные эффекты

Избыточные предупреждения вызывают привыкание и ухудшают интерфейс.

Остаточный риск

Корректно процитированный источник может быть ошибочным, устаревшим или вредоносным. Остаточный сценарий — пользователь принимает убедительный, но неверный вывод даже при видимой атрибуции.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud AI Studio

Частичная поддержка. File Search возвращает аннотации с идентификатором и именем файла, которые приложение может показать как происхождение использованного фрагмента.

Ограничение: Аннотация указывает использованный файл, но не подтверждает утверждение и не сообщает результат независимой проверки.

Yandex Search API

Частичная поддержка. Генеративный ответ Search API может содержать ссылки на веб-источники, которые приложение показывает рядом с ответом как сведения о происхождении.

Ограничение: Ссылки показывают источники ответа, но не подтверждают их качество, истинность утверждений и результат независимой проверки.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	491	93
AISAFE2-WEAK-002 · Отсутствие привязки утверждений AI-ответа к источникам	572	145
AISAFE2-WEAK-005 · Недостаточная маркировка и предупреждения в интерфейсе	577	151
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели	657	242
AISAFE2-WEAK-073 · Отсутствие независимого проверочного материала для AI-результата	683	274
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	871	35
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	872	48
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	1209	320
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	1210, 1214	339
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1211	389
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1212, 1351	488
AISAFE2-MIT-009 · Версионирование и откат базы знаний	1218	346

Мера защиты · AISAFE2-MIT-006

Ограничение потребления ресурсов и стоимости

Краткое описание

Сквозные счётчики и единая политика жёстких бюджетов ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Область применения

- Общая безопасность

Механизмы

- Контроль ресурсов

Защитный механизм

Счетчики получают идентичность, класс операции, текущую стоимость и остаток бюджета. Шлюз API ограничивает входной поток, а планировщик агента и обёртки инструментов контролируют внутренние шаги, глубину цепочки, число вложенных вызовов и платные операции.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация выбирает охватываемые модели, инструменты, очереди и внешние платные операции по стоимости, доступности и возможности злоупотребления. Для систем с существенными последствиями предпочтителен сквозной бюджет через границы компонентов; для ограниченных сценариев могут применяться локальные пределы. Мера не ограничивает неучтенный канал и не заменяет планирование общей емкости инфраструктуры.

Предпосылки

- Нужны сквозная идентификация субъекта и цепочки, инвентарь всех учитываемых моделей, инструментов, очередей и платных операций и единая политика бюджетов.
- Отказ счетчика должен приводить к ограниченному режиму, а не к неограниченному продолжению.

Реализация

Бюджет сохраняется при переходе между моделью, RAG, агентом, инструментом и внешним сервисом и не сбрасывается на границе компонента. Для расходов, оплачиваемых пользователем, политика до существенной операции показывает доступную оценку или верхнюю границу, применяет пользовательский бюджет и отправляет предупреждение при заданном пороге; в режиме жёсткого денежного лимита до платной операции по применимому тарифу определяется и атомарно резервируется верхняя граница затрат; если определить её нельзя, операция блокируется, а фактическое списание сверяется с резервом. Результатом является разрешение, замедление, отказ или безопасное завершение цепочки. Служебные учетные записи и пакетные процессы не должны обходить бюджет без отдельного профиля. В бюджет включается предпринятая внутренняя работа, в том числе отклоненные или завершившиеся ошибкой операции: расширение и нормализация входа, сравнения, окна модели, потоки выполнения, память, временное хранилище, очередь, параллелизм и повторные попытки. До запуска следующей ограничиваемой операции проверяется и, где применимо, резервируется достаточный остаток бюджета. Исчерпание бюджета не должно превращаться в обычный разрешающий, частичный или молча пустой успешный результат.

Проверка

Каждая операция в области учёта, зафиксированной в версии бюджетной политики, должна отражаться в соответствующем бюджете; фактическое превышение не должно выходить за допуск из этой версии. Мера неэффективна, если бюджет можно сбросить сменой пути или при отказе счетчика выполнение становится неограниченным.

Негативный тест

Запускают контрольную операцию ниже лимита и ту же цепочку с превышением каждого бюджета, параллельными запросами, новым сеансом и отказом счетчика. Контрольная операция должна завершиться, а превышение и обход — остановиться до следующей учитываемой или платной операции; отказ счетчика не должен открывать неограниченный режим.

Подтверждения

Подтверждениями служат версии бюджетных политик, конфигурация счетчиков и шлюзов, трассировки сквозного расходования бюджета, результаты конкурентных и отказных тестов, журналы ограничений и метрики стоимости, задержки и ложных блокировок.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При отказе счётчика новые операции с существенным ресурсным или финансовым воздействием блокируются, а активная цепочка завершается в безопасной точке. Для остальных сценариев допускается только заранее определённый минимальный режим с локальным лимитом; неограниченное продолжение запрещено.

Исправление и повторная проверка

Закрывают путь сброса бюджета, исправляют общий счетчик или переводят отказавший счетчик в ограниченный режим. Повторяют исходную распределенную или конкурентную нагрузку с тем же субъектом: суммарный расход достигает предела, зафиксированного в версии бюджетной политики, и дальнейшие операции прекращаются; сохраняют конфигурацию бюджета и сквозную трассировку списаний.

Побочные эффекты

Жесткие лимиты ухудшают доступность и могут непропорционально затрагивать отдельные группы пользователей.

Остаточный риск

Контроль не устраняет общую нехватку мощности и недостаточную отказоустойчивость. Медленные распределённые запросы через разные идентичности или неучтённый компонент могут оставаться ниже каждого локального порога.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud AI Studio

Частичная поддержка. Responses API позволяет задать максимальное общее число вызовов встроенных инструментов, обрабатываемых в одном ответе.

Ограничение: Предел действует для встроенных инструментов в одном ответе; внешние вызовы, повторы и общий бюджет операции остаются вне него.

Yandex Serverless Containers

Частичная поддержка. Настройка задает максимальное число одновременных вызовов контейнера в одной зоне доступности.

Ограничение: Ограничения действуют только на контейнеризованный путь и не охватывают токены, шаги агента, повторные попытки, внешние инструменты или суммарную стоимость рабочего процесса.

Yandex Smart Web Security

Частичная поддержка. Advanced Rate Limiter ограничивает число HTTP-запросов по времени и параметрам трафика и может блокировать превышение.

Ограничение: Ограничиваются входящие HTTP-запросы; токены, внутренние шаги агента, вызовы инструментов и общая стоимость не учитываются.

Yandex Cloud Billing

Косвенная поддержка. Бюджет позволяет контролировать расходы на ресурсы в Yandex Cloud.

Ограничение: Бюджет показывает достижение порога расходов, но не останавливает потребление и не завершает AI-цепочку.

Yandex Monitoring

Косвенная поддержка. Monitoring принимает пользовательские метрики потребления и уведомляет при достижении заданных порогов, поэтому может сигнализировать об исчерпании операционного бюджета.

Ограничение: Оповещение требует прикладных метрик; Monitoring не останавливает цепочку и не объединяет расходы модели, RAG и инструментов.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Применяет вне модели лимиты шагов, времени, стоимости, параллелизма, повторов и внутренних или внешних бюджетов.

Ограничение: Счётчики не создают обязательное для всех исполнителей состояние аварийной остановки и не определяют остановку при отсутствии прогресса или отзыве задания.

Условие применения: Для многошагового агента с ограничиваемыми шагами, временем, стоимостью, параллелизмом или повторами.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Применяет для каждой идентичности лимиты запросов, шагов, времени, параллелизма и бюджета.

Ограничение: Не является изоляцией конфиденциальности файлов и состояния.

Условие применения: Если нагрузки нескольких арендаторов или систем совместно потребляют вычислительные ресурсы.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Применяет серверные лимиты частоты, времени, параллелизма и совокупного потребления.

Ограничение: Лимиты запросов не задают домен, аутентификацию, журналирование и запрет прямого обхода публичного шлюза к внутреннему сервису.

Условие применения: На публичных и партнёрских маршрутах ИИ-приложения, где требуется ограничить частоту и объём работы.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.1.2 · Поддерживает часть требования

Сквозной бюджет ограничивает глубину агентной цепочки, расход токенов и стоимость одной операции во время исполнения.

Граница связи: Конкретные пределы и охват дочерних вызовов нужно задать и проверить в системе.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает лимиты глубины рекурсии, токенов и расходов на одно выполнение.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM06:2026 · Поддерживает часть требования

Жёсткие бюджеты на токены, шаги, вызовы инструментов, параллелизм и стоимость ограничивают разрастание расхода в охваченной цепочке.

Граница связи: Неучтённые платные маршруты и нехватку общей мощности такие бюджеты не устраняют; пределы надо настроить и проверить.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-017 · Неограниченное потребление токенов при обращении к языковым моделям	543	120
AISAFE2-IMP-018 · Незапланированные расходы на использование платных инструментов	544	121
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	545	123
AISAFE2-IMP-020 · Неожиданно высокие расходы пользователей AI-системы	548	125
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	549	126
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	551	127
AISAFE2-WEAK-052 · Неограниченные внутренние циклы и усиление нагрузки	655	237
AISAFE2-WEAK-062 · Отсутствие сквозной координации ресурсных и стоимостных бюджетов	671	257
AISAFE2-WEAK-063 · Недостаточные пользовательские средства управления расходами и прозрачность тарификации	672	260
AISAFE2-WEAK-075 · Отсутствие совокупных пределов масштабирования и расходов инфраструктуры	685	277
AISAFE2-WEAK-076 · Отсутствие ресурсной изоляции арендаторов в разделяемой среде	686	280
AISAFE2-WEAK-077 · Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку	687	281
AISAFE2-WEAK-078 · Недостатки управления нагрузкой, очередями и приоритетами при перегрузке	688	282
AISAFE2-WEAK-093 · Неверная атрибуция тарифицируемого потребления между арендаторами	715	311
AISAFE2-COMP-002 · Модель и интерфейс инференса	873	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	874	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	875	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	876	58
AISAFE2-COMP-013 · Телеметрия и обнаружение	877	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	878	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	879	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	880	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	881	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	882	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	883	82

Карточка	Связи	Стр.
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1213	485
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1329	462

Мера защиты · AISAFE2-MIT-007

Допуск проверенных источников в контур RAG

Краткое описание

В рабочий RAG-контур допускаются только идентифицированные источники, прошедшие установленную проверку происхождения, целостности и качества.

Механизмы

- Валидация
- Целостность

Защитный механизм

Шлюз допуска получает источник, метаданные владельца, правила допустимости и версии обработчика, разбиения и векторизации. Он проверяет происхождение, тип и размер, целостность, вредоносные вложения, запрещенные классы данных и требуемые метаданные.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Организация выбирает проверки и способ изоляции по происхождению, доверию, доступности источника атакующему и последствиям ошибки. Мера не подтверждает истинность содержания и не гарантирует, что ранее допустимый источник не будет скомпрометирован после проверки.

Предпосылки

- Нужны инвентарь каналов загрузки и прямой записи, стабильная идентификация источника и версии и принудительная граница между непроверенными и рабочими данными.
- Уровень проверки и условия повторного допуска устанавливаются до загрузки источника.

Реализация

Результат допуска связывается с точной версией источника и конвейера; подписанный манифест является усиленным вариантом такой связи. Контроль применяется до предоставления источника рабочему извлечению и повторно при значимом изменении. Если используется карантин или черновой индекс, рабочая цепочка не должна получать из него данные в обход решения о допуске.

Проверка

Перечисляют все пути загрузки и прямой записи, проводят допустимый и отклоненный источники через каждый путь, изменяют ранее допущенный объект и выполняют рабочий запрос по их уникальным маркерам. Мера неэффективна, если источник вне действующего решения о допуске извлекается в рабочем запросе.

Негативный тест

Загружают допустимый контрольный источник и источник с отсутствующими метаданными или вредоносным вложением, затем изменяют отклонённый объект и пытаются записать его прямо в индекс при отказе обработчика. Контрольный источник должен стать доступен извлечению, а все недопущенные варианты — остаться вне рабочего индекса и сформировать наблюдаемое решение об отказе.

Подтверждения

Подтверждения включают записи допуска или, для усиленного варианта, подписанные манифесты, версии обработчика и модели векторизации, конфигурацию разделения рабочих и непроверенных данных, журналы допуска, результаты негативных загрузок и тест прямого обхода.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Новый или измененный источник не предоставляется рабочему извлечению до выбранного решения о допуске; при наличии карантина он остается в нем. Система может продолжать использовать последнюю проверенную версию. Для высокорисковых источников автоматическая публикация при недоступности шлюза не соответствует назначению меры.

Исправление и повторная проверка

Удаляют обход прямой записи, возвращают отсутствующие метаданные либо повторно обрабатывают измененный источник через установленный допуск. Тот же источник повторно загружают штатным и обходным способом: он остается вне рабочего извлечения до решения, после допуска доступна только утвержденная версия; сохраняют запись допуска и идентификатор версии.

Побочные эффекты

Проверка увеличивает задержку публикации и стоимость обработки, а карантин может временно снижать полноту базы знаний.

Остаточный риск

Допущенный материал может быть ложным, устаревшим или стать вредоносным после компрометации источника. Остаточный сценарий — семантически вредный текст проходит формальные проверки и влияет на модель; его ограничивают разделение контекста и внешняя политика действий.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Security Deck

Частичная поддержка. Сигнал DSPM о вредоносном содержимом в данных можно включить в решение о допуске источника в RAG.

Ограничение: Сервис не создает сам по себе шлюз продвижения в рабочий индекс и не подтверждает безопасность проверенного источника. Охват зависит от поддерживаемого хранилища, профиля и настроенного процесса допуска.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-RAG-INGEST1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет происхождение, версию и обязательные метаданные источника и отделяет недопущенный материал от рабочего извлечения.

Ограничение: Проверка и карантин источника не назначают загрузчику минимальные права и не разделяют пути загрузки и поиска.

Условие применения: До включения нового или изменённого источника в рабочий RAG-корпус или поисковый индекс.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI06** · Поддерживает часть требования
>
> Проверка происхождения и целостности источника до допуска в RAG снижает риск отравления повторно используемого контекста через непроверенный источник.
>
> **Граница связи:** Допущенный текст может оставаться ложным или вредоносным; запись в память вне RAG проверяется отдельно.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.
Условие связи с карточкой	Постоянный контекст агента формируется из охваченного контролем пути RAG-источника.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	518	104
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	589	173
AISAFE2-COMP-023 · Подсистема формирования базы знаний	884	86
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1199	324
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	1210, 1214	331
AISAFE2-MIT-009 · Версионирование и откат базы знаний	1215, 1219	346
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1216, 1307	444

Мера защиты · AISAFE2-MIT-008

Авторизация извлекаемых данных на минимальной значимой гранулярности

Краткое описание

Каждый запрос к хранилищу авторизуется на минимальной значимой гранулярности до извлечения документа, фрагмента или записи.

Механизмы

- Авторизация

Защитный механизм

Механизм извлечения передает в механизм политики идентичность запрашивающего субъекта, цель, источник, метки доступа и кандидаты. Проверка повторяется после ранжирования, расширения, чтения из кэша и других этапов, которые могут изменить окончательный набор или область прав.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация определяет гранулярность и точки повторной проверки по структуре прав, преобразованиям и последствиям утечки. Все заявленные пути поиска, расширения и кэширования должны сохранять выбранную гарантию, но мера не проверяет правдивость, качество и безопасность смысла уже разрешенных данных.

Предпосылки

- Обязательными предпосылками выбранного контроля являются устойчивая идентичность субъекта, цель запроса и актуальные метки доступа на минимальной значимой единице авторизации.

Реализация

Контроль может быть централизованным или распределенным, если финальные данные связаны с решением актуальной политики. Например, при общем индексе двух арендаторов запрос пользователя арендатора A не должен получить фрагмент арендатора B, даже если тот появился только после повторного ранжирования, расширения цитаты или чтения из кэша. Тот же принцип применяется к сохраненным результатам и повторному использованию данных. Сервисная идентичность не должна заменять пользовательский контекст без явного правила.

Проверка

Авторизация должна применяться на выбранной гранулярности после последнего преобразования, способного изменить состав или права данных перед контекстом. В межпользовательских и межконтурных негативных тестах число неразрешенных единиц данных равно нулю. Проверяются разрешенное извлечение, устаревшая группа, смена цели, расширение цитаты, кэш, потеря идентичности, отказ сервиса политики и пакетный путь. Мера неэффективна при любой успешной выдаче запрещенных данных или отсутствии решения для единицы авторизации.

Негативный тест

От имени пользователя без права запрашивают запрещенную единицу, затем повторяют после повторного ранжирования, расширения контекста и через кэш: во всех вариантах число неразрешенных единиц в контексте и ответе должно быть равно нулю.

Подтверждения

Подтверждения включают версию политики, схему идентичности, гранулярности и меток, журналы решений, трассировки после повторного ранжирования и расширения, результаты межпользовательских и межконтурных негативных тестов и проверку кэшей.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Извлечение блокируется. Допускается режим без RAG или с публичным заранее проверенным корпусом, если он отделен технически; использование общей сервисной идентичности как резервного пути запрещено.

Исправление и повторная проверка

Исправляют гранулярность меток, перенос идентичности или повторную авторизацию после повторного ранжирования, из-за которых запрещенный фрагмент попал в контекст. Повторяют тот же межпользовательский запрос и обращение через кэш: число неразрешенных единиц данных равно нулю; сохраняют решение политики и трассировку выбранных единиц.

Побочные эффекты

Проверка повышает задержку и может уменьшать полноту извлечения, особенно при сложной политике и частых изменениях групп.

Остаточный риск

Ошибочная метка, устаревшая группа, неучтённый кэш или производный объект могут обойти корректно работающую авторизацию. Даже разрешённый фрагмент может содержать чувствительный вывод другого уровня или вредоносную инструкцию.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Identity and Access Management

Частичная поддержка. IAM проверяет полномочия субъекта на операции с облачными ресурсами и поддерживает роли, сервисные аккаунты и запрещающие политики доступа.

Ограничение: Гранулярность облачного ресурса может быть крупнее документа или фрагмента RAG. IAM не переносит пользовательский контекст через повторное ранжирование и кэши; наличие ролей не подтверждает минимальность выданных прав.

Yandex Managed Service for OpenSearch

Частичная поддержка. Если OpenSearch используется как RAG-хранилище, отдельные пользователи и уровни доступа могут ограничить круг субъектов, работающих с данными кластера.

Ограничение: Сервис сам не переносит пользовательский контекст и контекст арендатора в извлечение, не гарантирует ACL документа или фрагмента после повторного ранжирования и не изолирует прикладные кэши.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Авторизует документы RAG до и после извлечения.

Ограничение: Не покрывает публичную сетевую доступность каждого хранилища.

Условие применения: При извлечении RAG-документов из областей с различающимися правами доступа.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-RAG-ACL2

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Использует проверенные права пользователя и арендатора до извлечения и повторно проверяет документы после ранжирования или кэша перед контекстом модели.

Ограничение: Гарантия зависит от актуальных меток и прав и охвата кэшей, расширения и агентного пути. Разрешённый документ всё ещё может содержать ложные сведения или вредоносную инструкцию.

Условие применения: Если RAG содержит документы разных пользователей, арендаторов или иных областей доступа.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 8 (правила доступа) · Поддерживает часть требования

Проверка прав на фрагмент до выдачи поддерживает установленное правило доступа к персональным данным в RAG-контуре.

Граница связи: Мера не задаёт правила доступа для остальных хранилищ и операций ИСПДн и не регистрирует все действия с персональными данными.

Субъект	Оператор персональных данных
Условие применения	RAG-контур извлекает персональные данные, обрабатываемые в ИСПДн.
Условие связи с карточкой	Правило доступа оператора применяется к каждой выдаваемой единице персональных данных после последнего преобразования набора результатов.

Открыть положение в источнике

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C5.2.2 · Поддерживает часть требования

Проверка прав на каждый извлекаемый фрагмент и повторная проверка после преобразований поддерживают авторизацию пользователя при сборке контекста.

Граница связи: Нужны актуальные атрибуты прав и охват всех путей сборки; другие каналы доступа карта не покрывает.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверку прав конечного пользователя на каждом этапе поиска и сборки контекста.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM09:2026 · Поддерживает часть требования

Проверка прав на выдаваемые фрагменты и повторная проверка после ранжирования и кэша ограничивают попадание чужого содержимого в контекст модели.

Граница связи: Мера не требует ограничить сам поиск по общему индексу. Число результатов, оценки сходства и время ответа могут раскрывать данные другого арендатора без выдачи фрагмента; инверсия и отравление эмбеддингов тоже вне её охвата.

Субъект	Аналитик угроз AI-приложения
Условие применения	В приложении используются векторное хранилище, поиск по сходству или векторные представления данных.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	522	106
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	603	184
AISAFE2-WEAK-086 · Утрата атрибутов доступа исходного материала при индексировании базы знаний	702	298
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	885	48
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	886	61
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	887	75
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1200	324
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1217	444
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	1223	349
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	1227	356
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1261	400

Мера защиты · AISAFE2-MIT-009

Версионирование и откат базы знаний

Краткое описание

Версионирование связывает рабочую базу знаний с точным составом и позволяет атомарно восстановить последнюю допустимую версию.

Механизмы

- Жизненный цикл
- Восстановление
- Целостность

Защитный механизм

Реестр связывает рабочее содержимое с однозначно определенным выпуском, а процедура возврата восстанавливает проверенную версию и очищает несовместимые кэши.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Организация определяет способ идентификации, контроля целостности и возврата по частоте изменений, допустимому времени восстановления и риску смешения данных. Мера не предотвращает публикацию содержательно плохого выпуска, если проверки допуска его не выявили. Для неизменяемого встроенного набора без отдельного выпуска мера не применяется.

Предпосылки

- Обязательные предпосылки выбранного контроля — однозначная идентификация выпуска, целостность метаданных и проверенная возможность вернуть согласованное рабочее состояние.

Реализация

Конвейер фиксирует версии источников, преобразований, модели векторизации и состав индекса; при повышенных требованиях целостность дополнительно подтверждается хешами. Продвижение должно сохранять согласованный состав выпуска и исключать неуправляемое смешение несовместимых версий. Возврат может быть реализован атомарным переключением рабочей ссылки или иным проверенным механизмом, который очищает несовместимые кэши и восстанавливает согласованное состояние. Точки контроля выбираются при сборке, продвижении, выборе рабочей версии и восстановлении.

Проверка

Для каждого рабочего фрагмента должна восстанавливаться цепочка «источник -> преобразование -> фрагмент -> вектор -> выпуск» на выбранной гранулярности происхождения. Мера неэффективна, если рабочую версию нельзя однозначно определить или восстановить. Откат должен укладываться во время восстановления, установленное для базы знаний, и не оставлять несовместимых версий.

Негативный тест

Продвигают контрольную полную версию и варианты с измененным источником, неполной сборкой или поврежденным объектом, включая прямой путь и отказ хранилища. Контрольная версия должна стать единственной активной, остальные не должны смешаться с рабочим состоянием; возврат обязан восстановить точную предыдущую версию и очистить несовместимый кэш.

Подтверждения

Подтверждениями служат записи выпусков и происхождения, а при использовании усиленных механизмов — манифесты и контрольные суммы, журнал смены рабочей версии, отчет учения по возврату, метрика времени восстановления и результат проверки очистки кэшей.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Новый выпуск не продвигается; система сохраняет последнюю проверенную версию. При повреждении рабочего выпуска извлечение блокируется либо переключается на проверенную резервную версию с уведомлением.

Исправление и повторная проверка

Восстанавливают отсутствующий манифест, несовместимую версию либо механизм переключения, из-за которого откат оставил смешанное состояние. Повторяют продвижение поврежденной сборки и возврат к выбранной версии: извлечение использует только целевой выпуск, а кэши не содержат иной версии; сохраняют журнал смены и отчет времени восстановления.

Побочные эффекты

Хранение версий увеличивает стоимость и может продлевать хранение чувствительных данных; ошибочный откат способен временно снизить актуальность знаний.

Остаточный риск

Потеря зависимого индекса, кэша или внешнего источника может сделать откат неполным. Даже корректно версионированный, но содержательно вредный выпуск остаётся активным до сигнала мониторинга.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Managed Service for OpenSearch

Частичная поддержка. Резервные копии и восстановление позволяют вернуть состояние индекса, если OpenSearch используется как рабочее RAG-хранилище.

Ограничение: Резервная копия кластера не фиксирует версии корпуса, разбиения на фрагменты, модели эмбеддингов, преобразований и кэшей и не делает откат всего выпуска RAG атомарным.

Yandex Object Storage

Частичная поддержка. Версионирование бакета сохраняет историю объектов и позволяет восстановить предыдущую версию файла базы знаний.

Ограничение: Бакет с выключенным версионированием не даёт версий; даже включённые версии файлов не фиксируют состав индекса, разбиение, эмбеддинги и кэши, поэтому атомарный откат выпуска RAG требует отдельного механизма.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI06** · Поддерживает часть требования
>
> Откат к проверенной версии базы знаний может исключить отравленные данные из рабочего RAG-контура.
>
> **Граница связи:** Нужны обнаружение и пригодная чистая версия; мера не предотвращает первоначальный допуск и не очищает другую память агента.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.
Условие связи с карточкой	Отравление затронуло версионируемую базу знаний; пригодная версия известна и повреждение обнаружено.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	568	141
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	698	296
AISAFE2-COMP-023 · Подсистема формирования базы знаний	888	86
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	1215, 1219	339
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	1218	331
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1220	410
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1221	444
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1222	492
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1316	456

Мера защиты · AISAFE2-MIT-010

Минимизация и фильтрация чувствительных данных до передачи в AI-контур

Краткое описание

Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Область применения

- Общая безопасность

Механизмы

- Минимизация данных
- Валидация

Защитный механизм

Контроль получает данные, цель обработки, получателя и политику. Он удаляет ненужные поля, маскирует или токенизирует значения, блокирует секреты и формирует решение с перечнем преобразований.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера обычно применима, если данные могут содержать чувствительные сведения и передаются в модель, RAG, память, обучение или внешний AI-компонент. К чувствительным относятся не только персональные данные и секреты, но и коммерческие, технологические, служебные и косвенные признаки, если их раскрытие создаёт установленный риск. Она не гарантирует обнаружение всех смысловых или косвенных признаков и не заменяет авторизацию и контроль исходящих данных.

Предпосылки

- Нужны классификация охватываемых данных, карта всех путей до AI-границы и правило обработки неизвестного класса.
- Преобразование должно выполняться до передачи модели, журналирования внешнего вызова и формирования кэша.

Реализация

Точки принудительного контроля располагаются перед каждой новой границей доверия и новым местом хранения, а не только на внешнем API. Повторно извлеченные и сгенерированные данные проходят контроль снова. Контроль повторяется перед выдачей сгенерированного результата, записью в долговременную память, созданием временного артефакта среды выполнения и формированием телеметрии. В обычные логи, трассировки, записи ошибок и диагностические выгрузки передается только минимальный набор данных для заявленной цели.

Проверка

Проверка должна охватывать все чувствительные входы, хранилища и исходящие пути, включенные в область меры. Контрольные секреты и запрещенные тестовые записи не должны достигать приемника. До выпуска в версии политики для каждого класса данных фиксируются допустимые ложноположительные и ложноотрицательные показатели. Для вероятностного детектора используется репрезентативный синтетический набор по заявленным классам, языкам, форматам и путям; ложноположительные и ложноотрицательные результаты фиксируются раздельно. Неизвестное поле, тайм-аут или ошибка преобразования дают безопасный отказ до следующей границы. Контрольное значение не должно появляться при

штатном запросе, ошибке, повторе и диагностическом экспорте. Проверяются маскирование, вложенные и кодированные значения, изображение или файл, повторное извлечение, отказ классификатора и высокая нагрузка. Мера неэффективна при успешной передаче контрольного секрета или наличии заявленного пути без проверки.

Негативный тест

В каждую охваченную точку входа помещают уникальный контрольный секрет и запрещенную синтетическую запись, включая неизвестное поле и отказ преобразования: значение не должно появиться за AI-границей или в журнале.

Подтверждения

Подтверждения включают схему потоков и классов данных, версию политики и преобразований, результаты тестов с контрольными секретами и запрещенными записями, метрики ошибок классификации, журналы решений без исходного чувствительного значения и проверку обходных путей.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Для чувствительных потоков с существенными последствиями передача и запись запрещенного или неопределенного содержимого блокируются либо направляются на отдельную проверку. В иных сценариях допускается продолжение только с заранее выделенным нечувствительным набором данных; исходный заблокированный материал не должен появляться в обычных журналах.

Исправление и повторная проверка

Исправляют классификатор, преобразование или обходной канал, через который контрольный секрет достиг AI-границы. Повторяют передачу той же синтетической записи при штатной работе и отказе преобразования: запрещенное значение не покидает доверенный контур и не попадает в журнал; сохраняют версию политики и решение без исходного чувствительного значения.

Побочные эффекты

Маскирование способно ухудшать качество и создавать несправедливое влияние на отдельные группы; ложные срабатывания могут блокировать допустимую обработку.

Остаточный риск

Классификатор может пропустить редкое или контекстное значение; комбинация разрешенных признаков может раскрыть чувствительный факт. Остаточный сценарий — смысловая утечка без дословного секрета через разрешённый ответ или инструмент.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Lockbox

Частичная поддержка. Lockbox позволяет хранить пароли, токены и ключи отдельно от промптов и версионируемой конфигурации и получать их через контролируемый канал.

Ограничение: Lockbox хранит секрет, но не блокирует его передачу модельному вызову после выдачи приложению и не фильтрует другие чувствительные данные.

Yandex Security Deck

Косвенная поддержка. DSPM обнаруживает и классифицирует чувствительные данные в поддерживаемых хранилищах, помогая определить поля и источники для исключения или обезличивания.

Ограничение: DSPM находит чувствительные данные в хранилищах, но не фильтрует каждый запрос к модели, инструменту и журналу.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AUDIT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Удаляет, маскирует или токенизирует секреты и регулируемые данные до журналов и диагностики с безопасным запретом при сбое.

Ограничение: Преобразование до журналирования не задаёт обязательную схему события, правовое основание и порядок исключительного хранения полного содержимого.

Условие применения: Перед записью чувствительных значений в прикладной журнал, трассировку, сообщение об ошибке или диагностическую выгрузку.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY-PII1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Обнаруживает и минимизирует, маскирует или токенизирует данные до первой запрещённой границы доверия и блокирует остаточные данные или сбой преобразования.

Ограничение: Карточка не определяет допустимость обратимости, хранение таблиц соответствия и ключей, права на детокенизацию и оценку риска перебора при хешировании. Эти условия задают для выбранного метода.

Условие применения: Если утверждённое решение требует преобразовать персональные данные до передачи контуру, не вправе видеть исходные значения.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY-PII3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Блокирует ошибки разбора, неизвестные поля, тайм-ауты и сбои преобразователя и не допускает исходные чувствительные значения в последующие вызовы и журналы.

Ограничение: Безопасный отказ преобразователя не разделяет права редактора правил, пользователя ключа и таблицы, оператора обратного преобразования и читателя свидетельств.

Условие применения: При программном маскировании, токенизации, псевдонимизации или другом защитном преобразовании перед AI-обработкой.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY-PII4

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет средство обнаружения на репрезентативном наборе по классам, языкам, форматам и путям, раздельно измеряет пропуски и ложные срабатывания и задаёт безопасный отказ при неизвестном поле или ошибке.

Ограничение: Стандарт дополнительно требует периодической оценки качества и мониторинга ошибок и обходов в эксплуатации. В карточке эти процедуры и их периодичность не установлены.

Условие применения: Если в сценарии обработки персональных данных заявлен детектор, классификатор или другое средство их обнаружения.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA4

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Минимизирует, маскирует или токенизирует данные до модели, RAG, памяти, обучения, инструмента и журнала с безопасным запретом при сбое.

Ограничение: Преобразователь исполняет переданную политику, но не классифицирует каждое поле, не выбирает получателя и не решает, допустима ли обратимость.

Условие применения: До передачи чувствительных данных в промпт, RAG-контекст, инструмент, внешний вызов или журнал.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке

Редакция: Методические рекомендации Банка России № 3-МР от 16.06.2026

п. 5.6 · Поддерживает часть требования

Удаление лишних полей до передачи набора поставщику снижает объём чувствительных данных, доступных при обучении модели.

Граница связи: Маскирование или токенизация не равны автоматически обезличиванию; карта не создаёт синтетический набор и не контролирует дальнейшую обработку у поставщика.

Субъект	Участник финансового рынка из п. 1.1 рекомендаций
Условие применения	Поставщик услуг обучает модель ИИ на наборах данных, принадлежащих участнику финансового рынка из пункта 1.1.
Условие связи с карточкой	Организация очищает и минимизирует набор до его передачи поставщику для обучения; обычный запрос к готовой модели не образует этот сценарий.
Сроки и переходные условия	Методические рекомендации № 3-МР носят рекомендательный характер.

Открыть положение в источнике

Приказ ФСТЭК России от 11.04.2025 № 117 Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений

Редакция: Приказ № 117 от 11.04.2025 с изменениями по приказу № 137 от 08.05.2026

п. 60, абз. 2 · Поддерживает часть требования

Фильтрация до передачи в AI-контур может исключить часть информации ограниченного доступа из обращения к разработчику модели.

Граница связи: Нужно отдельно установить класс информации, получателя и все пути передачи; фильтр может пропустить косвенные признаки и не подтверждает исполнение запрета.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	Информация ограниченного доступа из системы, подпадающей под приказ № 117, передаётся лицу, разработавшему используемую модель.
Условие связи с карточкой	Фильтрация выполняется до передачи разработчику модели и охватывает установленные классы информации ограниченного доступа.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 60 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 5, ч. 4–5 · Поддерживает часть требования

Мера удаляет лишние поля до передачи персональных данных в AI-контур и тем поддерживает ограничение их состава заявленной целью обработки.

Граница связи: Законную цель и основание обработки определяют отдельно; фильтр действует лишь на охваченных путях и не доказывает достаточность минимизации во всём процессе.

Субъект	Оператор персональных данных
Условие применения	Оператор обрабатывает персональные данные для конкретной, заранее определённой законной цели; требования статьи 5 действуют независимо от наличия ИСПДн.
Условие связи с карточкой	До передачи в AI-контур исключаются именно те записи и поля, которые избыточны для установленной цели обработки.

Открыть положение в источнике

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 5(1)(c) · Поддерживает часть требования

Фильтрация до передачи в AI-контур удаляет или маскирует поля, не нужные для утверждённой цели обработки.

Граница связи: Цель и необходимость каждого оставленного поля определяет контролёр; фильтр может пропустить косвенные персональные признаки.

Субъект	Контролёр персональных данных по GDPR
Условие применения	Операция обработки персональных данных подпадает под действие GDPR.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	496	97
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	523	106
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	531	113
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	557	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	564	139
AISAFE2-WEAK-036 · Передача в AI-контур избыточных данных (нарушение минимизации)	630	209
AISAFE2-WEAK-040 · Отсутствие маскирования или анонимизации чувствительных данных	633	214
AISAFE2-WEAK-074 · Отсутствие независимого разграничения содержимого контекста и вывода модели	684	276
AISAFE2-COMP-002 · Модель и интерфейс инференса	889	37
AISAFE2-COMP-003 · Управление контекстом общения с моделью	890	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	891	44
AISAFE2-COMP-007 · Долговременная память	892	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	893	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	894	54
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	895	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	896	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	897	69

Карточка	Связи	Стр.
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	898	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	899	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	900	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	901	82
AISAFE2-COMP-023 · Подсистема формирования базы знаний	902	86
AISAFE2-COMP-025 · Источники данных для инструментов	903	90
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	1223	342
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	1224, 1305	441
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1225	444
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1226	478
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1330	462

Мера защиты · AISAFE2-MIT-011

Изоляция и ограничение экспорта векторных хранилищ

Краткое описание

Изоляция, авторизация и ограничения экспорта сужают пути доступа к векторным представлениям (эмбеддингам) и массового извлечения содержимого хранилища.

Механизмы

- Изоляция
- Авторизация

Защитный механизм

При создании индекса выбираются сегмент, ключ шифрования, сетевой контур и политика доступа. Запросы и экспорт проходят отдельную авторизацию, ограничения объема и журналирование.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик

Применимость

Мера обычно применима при наличии векторного хранилища, его снимков или резервных копий. Она не проверяет смысловую безопасность извлеченного текста и не устраняет возможность восстановления свойств данных по разрешенным векторам.

Предпосылки

- Нужны инвентарь рабочих, резервных и экспортных копий векторного хранилища, устойчивая идентичность субъекта и сегмента и принудительная проверка каждого запроса и экспорта.
- Общая резервная идентичность без сопоставимой сегментации не допускается.

Реализация

Административные, резервные и аналитические пути включаются в ту же границу. Результатом является разрешенный запрос или экспорт с привязанной идентичностью либо отказ.

Проверка

Каждый индекс и каждая копия должны иметь владельца, сегмент, политику и шифрование. Проверяются разрешенный запрос, административный путь, массовый запрос, снимок, восстановление резервной копии, отказ ключа, сетевой обход и тесты восстановления сведений по векторам. Мера неэффективна при доступе к чужому сегменту или неучтенной копии.

Негативный тест

Межконтурный запрос и неразрешенный экспорт в обязательном наборе должны завершаться отказом.

Подтверждения

Подтверждениями служат инвентарь индексов и копий, конфигурации доступа, сети, шифрования и ключей, журналы запросов и экспорта, результаты межконтурных и административных негативных тестов и отчет восстановления резервной копии.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Запрос и экспорт из затронутого сегмента блокируются. При недоступности ключа или политики автоматический переход к общей учетной записи запрещен; допустимо переключение на проверенную изолированную резервную копию.

Исправление и повторная проверка

Устраняют общую идентичность, ошибочную сегментацию либо неучтенную экспортную копию, позволившие межконтурный доступ. Повторяют тот же запрос и экспорт от исходного субъекта: оба завершаются отказом, при этом разрешенный запрос к своему сегменту работает; сохраняют политики доступа и события запрета.

Побочные эффекты

Сегментация повышает стоимость, усложняет поиск и может снизить полноту результатов между обоснованно разделенными корпусами.

Остаточный риск

Разрешенный пользователь может выполнять смысловое извлечение малыми запросами, а векторные представления сохраняют свойства исходных данных. Остаточный сценарий — утечка через разрешенное извлечение или скомпрометированную административную роль.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Key Management Service

Частичная поддержка. Для диска OpenSearch, используемого как векторное хранилище, можно выбрать пользовательский KMS-ключ и тем самым включить отдельный криптографический контур хранения.

Ограничение: Шифрование диска не задаёт авторизацию арендатора для запросов, лимиты экспорта или журналирование прикладного доступа и не защищает данные после разрешённого расшифрования.

Yandex Managed Service for OpenSearch

Частичная поддержка. Если OpenSearch используется как векторное хранилище, сеть, отсутствие публичного доступа и группы безопасности ограничивают пути подключения к кластеру.

Ограничение: Сетевая изоляция кластера не задаёт права на документы и арендаторов, лимиты экспорта и журналирование запросов.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Покрывает шифрование и авторизацию векторных хранилищ и их снимков.

Ограничение: Связь охватывает векторные хранилища и их копии; защищённый транспорт и шифрование других активов требуют отдельных настроек.

Условие применения: Для векторного хранилища и его снимков или резервных копий.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ограничивает доступ и экспорт векторного хранилища, учитывает его снимки и резервные копии и применяет выбранные сетевые и криптографические границы.

Ограничение: Не является общим контролем непубличности хранилищ.

Условие применения: Для векторного хранилища, его снимков, резервных копий и экспортных путей.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM09:2026 · Поддерживает часть требования

Разделение векторных хранилищ и ограничение запросов и экспорта сужают доступ к эмбеддингам и массовому извлечению.

Граница связи: Разрешённый поиск всё ещё может раскрывать сведения; мера не выявляет отравление, инверсию из уже утёкшей копии или все побочные сигналы общего индекса.

Субъект	Аналитик угроз AI-приложения
Условие применения	В приложении используются векторное хранилище, поиск по сходству или векторные представления данных.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-011 · Компрометация баз знаний	527	109
AISAFE2-WEAK-023 · Незашифрованное хранение	609	186
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	703	299
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	904	61
AISAFE2-COMP-023 · Подсистема формирования базы знаний	905	86
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	1227	342
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1228	444
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1269	405
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1331	462
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1372	525

Мера защиты · AISAFE2-MIT-012

Проверка целостности и безопасная загрузка артефактов модели

Краткое описание

Загрузчик проверяет идентичность, происхождение, целостность и безопасный формат артефакта модели до его открытия или исполнения.

Механизмы

- Целостность

Защитный механизм

Манифест связывает идентификатор модели, версию, формат, отпечаток, источник и результаты обязательных проверок. Шлюз допуска проверяет цепочку доверия до загрузки, продвижения и восстановления; среда выполнения фиксирует фактически загруженную версию.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера обычно применима, когда организация получает или загружает доступный ей артефакт, веса либо сериализованный объект. Организация выбирает подпись, доверенный отпечаток или эквивалентный механизм по цепочке поставки и последствиям подмены. Мера не защищает его конфиденциальность; несанкционированное получение ограничивается «Защита артефактов модели от несанкционированного получения». Для внешнего API, где артефакт недоступен организации, применяется лучшая доступная идентификация версии и проверяемые обязательства поставщика; это не эквивалентно локальной проверке артефакта.

Предпосылки

- Обязательные предпосылки выбранного контроля — доверенный источник подтверждения целостности и защищенная связь с точной версией; все заявленные пути загрузки принудительно проходят через проверяющий загрузчик, а известный административный или прямой путь в установленной области отсутствует.

Реализация

Плавающий псевдоним не используется как единственная ссылка. Прямое чтение произвольного URI и административная загрузка в обход контроля запрещаются. Подтверждение целостности связывается с байтами объекта, фактически открытого для загрузки после разрешения пути и ссылок. Замена объекта между проверкой и чтением, перенаправление на другой объект и автоматический переход к загрузчику с более слабой политикой блокируются.

Проверка

Измененный байт, недействительное подтверждение целостности, запрещенный формат, неизвестный URI и плавающий псевдоним должны блокироваться. Мера неэффективна при загрузке до решения или неизвестной версии в среде выполнения.

Негативный тест

Загружают контрольный артефакт с выбранным доказательством целостности и его копии с изменённым содержимым, неверной версией или недействительным доказательством (для подписи — с отозванным ключом), затем повторяют через прямой загрузчик и при отказе проверяющего. Контрольный артефакт должен загрузиться как указанная версия, а изменённый, неизвестный или непроверяемый — не должен быть разобран или исполнен.

Подтверждения

Подтверждения включают точный идентификатор артефакта и выбранное доказательство целостности — например подписанный манифест или доверенный отпечаток, — конфигурацию загрузчика, журнал фактической версии, результаты тестов изменения байта и запрещенного формата и отчет возврата.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Новая модель не загружается; используется последняя проверенная версия либо AI-функция блокируется. Автоматическое отключение проверки целостности для восстановления доступности запрещено.

Исправление и повторная проверка

Возвращают доверенную запись версии, исправляют проверку целостности или удаляют путь загрузки в обход нее. Повторно загружают исходный артефакт с тем же измененным байтом или запрещенным форматом: загрузчик отказывает до активации, а разрешенная точная версия загружается; сохраняют результат проверки и фактический идентификатор модели.

Побочные эффекты

Строгие форматы ограничивают совместимость, а управление ключами и изолированная загрузка увеличивают операционную сложность.

Остаточный риск

Подписанный артефакт может содержать скрытую закладку или быть получен из скомпрометированного обучения. Остаточный сценарий — доверенный поставщик подписывает вредоносную либо ошибочную модель, целостность которой формально подтверждена.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Container Registry

Частичная поддержка. Для модели, упакованной в контейнерный образ, контрольная сумма однозначно связывает загрузку с конкретной версией образа вместо плавающего тега.

Ограничение: Поддержка применима только к контейнеризованным артефактам и не проверяет внутреннюю безопасность сериализованной модели. Наличие контрольной суммы не подтверждает доверенный источник или принудительную загрузку по ней.

Yandex Key Management Service

Частичная поддержка. Пары ключей цифровой подписи KMS позволяют подписывать отпечаток артефакта и проверять подпись перед загрузкой.

Ограничение: KMS выполняет подпись и проверку; доверенный манифест и обязательную проверку перед загрузкой должно обеспечить приложение.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MODEL2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

До загрузки проверяет происхождение, точную версию, целостность и допустимый формат артефакта; запрещённый или изменённый объект не должен разбираться или исполняться.

Ограничение: Анализ сопутствующего кода и зависимостей и первый запуск в изолированной среде без продуктивных секретов и записи требуют отдельных проверок.

Условие применения: Перед первым использованием полученного извне модельного комплекта, доступного организации для загрузки.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает доступный организации модельный артефакт с проверяемым источником, точной версией, форматом и подтверждением целостности до загрузки.

Ограничение: Анализ сопутствующего кода, пользовательского загрузчика и зависимостей и изолированный первый запуск задают отдельно; поведенческая приёмка относится к AI-SUPPLY3.

Условие применения: Для модели, весов, токенизатора или кода загрузки, полученных вне утверждённого внутреннего конвейера.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом

рынке

Редакция: Методические рекомендации Банка России № 3-МР от 16.06.2026

п. 5.5 (целостность модели) · Поддерживает часть требования

Проверка отпечатка, источника и версии модели помогает выявить её подмену перед загрузкой в AI-систему.

Граница связи: Карта не охватывает всю цепочку данных поставщика и не подтверждает применение средств контроля целостности, прошедших указанную ФСТЭК оценку соответствия.

Субъект	Участник финансового рынка из п. 1.1 рекомендаций
Условие применения	Участник финансового рынка из пункта 1.1 использует данные или модели ИИ поставщика услуг.
Условие связи с карточкой	Организации доступен артефакт модели, целостность которого проверяется перед загрузкой; внешнее API без доступа к артефакту не подтверждает такую проверку.
Сроки и переходные условия	Методические рекомендации № 3-МР носят рекомендательный характер.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM04:2026 · Поддерживает часть требования

Проверка источника, отпечатка, версии и безопасного формата до загрузки снижает риск подменённого артефакта модели в цепочке поставки.

Граница связи: Подлинный подписанный артефакт может содержать закладку; внешний API без доступных весов не даёт той же глубины проверки.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

LLM05:2026 · Поддерживает часть требования

Проверка точной версии и безопасного формата модели может остановить подменённый или исполняемый артефакт до загрузки.

Граница связи: Она не распознаёт поведенческую закладку в подлинных весах и не контролирует отравление обучающих либо RAG-данных.

Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	534	116
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	538	118
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	584	167
AISAFE2-WEAK-091 · Исполняемая десериализация артефактов модели	713	308
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	906	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	907	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	908	61
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1229	448
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1230	473
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	1231	364
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1234	368
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1346	481

Мера защиты · AISAFE2-MIT-013

Проверка компонентов цепочки поставки перед выпуском

Краткое описание

Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки.

Область применения

- Общая безопасность

Механизмы

- Шлюз допуска
- Целостность

Защитный механизм

Конвейер сборки формирует состав, применяет проверки, подходящие типу компонента, сопоставляет результаты с порогами, определёнными владельцем безопасности для выбранных проверок, и создаёт подтверждение выпуска, связанное с точным артефактом. Это может включать проверку подписи и источника, анализ программных зависимостей, конфигурации, инструкции, модели или поведения без требования одинакового сканирования для всех объектов.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Организация определяет охватываемые компоненты и проверки по составу выпуска, происхождению, доступности атакующему и актуальным угрозам. Для значимых или высокорисковых изменений предпочтителен блокирующий шлюз; для малых изменений допустим сокращенный набор, если он сохраняет связь результата с точным выпуском. Мера не выявляет все неизвестные уязвимости, логические закладки и нежелательное поведение модели.

Предпосылки

- Нужны точный состав каждого выпуска, происхождение компонентов, правила применимых проверок и принудительный шлюз, связанный с неизмененным выпускаемым артефактом.
- Для отказа проверяющего механизма и срочного исправления заранее задается блокирующий или ограниченный аварийный процесс.

Реализация

Обход через ручную сборку, прямую замену образа или позднее добавление инструмента несовместим с выбранным шлюзом. Операционная работа с новыми уязвимостями после выпуска выполняется отдельным процессом управления уязвимостями AI-компонентов.

Проверка

Неустранённый результат выше порога, определённого владельцем безопасности для соответствующей проверки, либо отсутствующее подтверждение блокируют выпуск. Мера неэффективна, если рабочий артефакт отличается от проверенного или шлюз можно миновать.

Негативный тест

Собирают допустимый контрольный выпуск и варианты с компонентом, для которого выявлена находка выше утверждённого порога, неизвестным инструментом или подменой после проверки, затем повторяют через ручной обход и при отказе проверяющего механизма. Контрольный неизмененный артефакт должен получить подтверждение выпуска, а каждый иной вариант — блокироваться до развертывания и требовать новой сборки и проверки.

Подтверждения

Подтверждения включают состав выпуска, происхождение компонентов, версии и отчеты выбранных проверок, запись решения шлюза и результаты тестов подмены и ручного обхода; подписи и аттестация сборки включаются, если они являются частью реализации.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Выпуск блокируется. Срочное исправление допускается только через заранее определенный аварийный процесс с минимальным набором проверок, ограниченным исключением и обязательной полной проверкой после.

Исправление и повторная проверка

Обновляют состав выпуска, устраняют подмену после проверки либо закрывают ручной обход шлюза для проблемного компонента. Повторяют сборку с тем же компонентом с находкой выше порога либо неизвестным объектом: выпуск блокируется, а после замены проходит только связанный с новым составом артефакт; сохраняют состав и решение шлюза.

Побочные эффекты

Сканеры могут создавать ложные срабатывания; полный анализ увеличивает время выпуска, а отказ сканера может временно остановить поставку исправления.

Остаточный риск

Сканеры пропускают неизвестные закладки. Остаточный сценарий — вредоносный или логически небезопасный компонент не распознан известными проверками и проходит в подписанный выпуск.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

SourceCraft Security

Частичная поддержка. Анализ зависимостей SourceCraft Security выявляет транзитивные программные зависимости и связанные с ними известные уязвимости до выпуска.

Ограничение: Анализ может охватывать зависимости адаптеров и MCP-серверов, если они входят в проверяемый проект, но не проверяет их логику и настройки, модели, промпты или фактическую конфигурацию среды выполнения. Сам по себе он не доказывает наличие блокирующей политики выпуска.

Yandex Container Registry

Частичная поддержка. Встроенный сканер проверяет контейнерные образы, пакеты и зависимости по базам CVE; сканирование можно запускать при отправке образа или по расписанию.

Ограничение: Сканируются образы и зависимости; модели, промпты, конфигурация и блокирующее решение о выпуске остаются вне сканера.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI04** · Поддерживает часть требования
>
> Проверка происхождения, состава и результатов испытаний компонентов перед выпуском снижает риск принять подменённый или уязвимый артефакт в сборку агента.
>
> **Граница связи:** Шлюз выпуска не проверяет динамически подключённые инструменты, внешних агентов и изменения поставщика после выпуска.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент зависит от инструмента, манифеста, промпта или памяти, происхождение которых может быть нарушено.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-012 · Непроверяемая или скомпрометированная цепочка поставки AI	585	167
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	638	222
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	909	44
AISAFE2-COMP-009 · Конвейер обучения и дообучения	910	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	911	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	912	67
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	913	84
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	1231	359
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1232, 1282	419
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1233, 1310	451
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1235	368
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1317	456

Мера защиты · AISAFE2-MIT-014

Фиксация версии и базового профиля поведения модели

Краткое описание

Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки.

Механизмы

- Управление конфигурацией
- Шлюз допуска

Защитный механизм

До выпуска система запускает точную версию модели с зафиксированными системными инструкциями и параметрами инференса на версионируемой оценочной выборке. Полученные показатели и контрольные ответы образуют базовый профиль, однозначно связанный с этой конфигурацией и временем оценки.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Третья сторона

Применимость

Организация выбирает способ идентификации и глубину базового профиля по степени контроля над моделью, возможным последствиям скрытой смены и доступности измеримых показателей. Мера не заменяет эксплуатационный цикл обнаружения и обработки дрейфа «Обнаружение и обработка дрейфа, смещения данных и деградации» и не наблюдает поведение, для которого достоверный признак или репрезентативная выборка не определены.

Предпосылки

- Обязательные предпосылки выбранного контроля — точная или лучшая доступная идентификация модели и конфигурации, воспроизводимая оценочная выборка и надежная связь базового профиля с этой идентификацией.

Реализация

Базовый профиль формируют до первого выпуска и после каждого изменения модели, системной инструкции, параметров инференса или методики оценки. В записи сохраняют идентификаторы конфигурации, версию выборки, способ расчёта показателей и допустимое отклонение повторного прогона; профиль непрозрачного внешнего сервиса связывают с доступным идентификатором развертывания и временем оценки.

Проверка

Для каждого базового профиля должны быть определены версия или лучшая доступная идентификация модели, конфигурация, оценочная выборка, методика и результаты. Мера неэффективна, если профиль нельзя воспроизвести в пределах допуска, зафиксированного в записи базового профиля, либо однозначно связать с оцениваемой конфигурацией.

Негативный тест

Повторяют оценку точной контрольной конфигурации, затем меняют модель, системную инструкцию или параметр и пытаются сопоставить результат при недоступной выборке либо через неидентифицированную версию внешнего сервиса. Контрольный повтор должен соответствовать допуску, зафиксированному в записи базового профиля, а измененный или несопоставимый вариант — получить новый профиль либо явный статус невозможности сравнения, но не наследовать прежний результат.

Подтверждения

Подтверждениями служат доступные идентификаторы модели, системных инструкций и параметров, версия оценочной выборки и методики, базовые показатели, результаты повторной оценки и зафиксированные ограничения воспроизводимости.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Продвижение новой версии блокируется до формирования связанного с ней воспроизводимого базового профиля. Если рабочую конфигурацию нельзя идентифицировать, ее состояние считается неизвестным и для значимых решений применяется ограниченный режим или ручная проверка до восстановления идентификации.

Исправление и повторная проверка

Фиксируют недостающую версию модели, параметры инференса или репрезентативный базовый набор, из-за которых профиль нельзя воспроизвести. На той же конфигурации повторяют базовую оценку и сравнение: результат укладывается в допуск, зафиксированный в записи базового профиля и однозначно связан с активной версией; сохраняют параметры и отчет сравнения.

Побочные эффекты

Формирование и воспроизведение базового профиля увеличивают стоимость оценки, а неудачно выбранная выборка создает ложное представление о поведении модели.

Остаточный риск

Неполная выборка может не отражать редкие режимы поведения, а профиль устаревает при изменении данных и сценария. Сам по себе базовый профиль не обнаруживает последующие отклонения без отдельного эксплуатационного мониторинга.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud AI Studio

Частичная поддержка. Text Generation API возвращает фактическую версию модели вместе с результатом, что позволяет привязать контрольные ответы к наблюдаемой версии.

Ограничение: Номер версии модели не задаёт тестовую выборку, версию промпта, базовый профиль поведения и критерии допуска.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает точную версию модели и конфигурацию с версионируемой оценочной выборкой и воспроизводимым базовым профилем.

Ограничение: Не задаёт полный набор метрик качества или обязательную независимость проверяющего.

Условие применения: До первого выпуска или изменения модели, её конфигурации, системных инструкций либо методики оценки.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

> **MANAGE 1.1** · Поддерживает часть требования
>
> Базовый профиль поведения точной версии модели даёт сравнимые данные для оценки её пригодности к заявленной цели.
>
> **Граница связи:** Профиль не определяет все цели AI-системы и не принимает решение о продолжении разработки или выпуска.

Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	Владелец решает, достигает ли конкретная AI-система заявленной цели и допустим ли её выпуск, используя добровольный NIST AI RMF.
Условие связи с карточкой	Конкретная модель и конфигурация оцениваются против заявленных целей системы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	914	37
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	915	56

Карточка	Связи	Стр.
AISAFE2-COMP-013 · Телеметрия и обнаружение	916	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	917	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	918	79
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	1234	359
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	1235	364
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1236	448
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1237	451
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1238, 1318	456
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1239, 1335	466
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1240	469
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1241	473

Мера защиты · AISAFE2-MIT-015

Защита конфигурации системных промптов и инструкций инструментов

Краткое описание

Системные промпты, инструкции инструментов и правила сборки контекста управляются как защищённая версионируемая конфигурация.

Механизмы

- Управление конфигурацией

Защитный механизм

Конфигурация проходит выбранные проверки секретов и политик, получает однозначную версию и продвигается через контролируемый путь. Для систем с повышенными требованиями целостность может дополнительно подтверждаться криптографическим отпечатком, а секреты — передаваться среде выполнения по ссылке или краткоживущему полномочию из отдельного хранилища.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера обычно применима при наличии системных или разработческих промптов, описаний инструментов, шаблонов либо конфигурации правил. Она не может скрыть инструкции от косвенного восстановления по поведению модели и не защищает секрет, уже раскрытый до

контролируемой передачи.

Предпосылки

- Обязательными предпосылками являются управляемое хранилище конфигурации, отдельный канал секретов и доверенный способ проверки версии в среде выполнения.

Реализация

Значение секрета нельзя помещать в промпт, описание инструмента, параметры процесса, журнал или версионируемую конфигурацию. Среда выполнения фиксирует активную версию и, если выбран усиленный вариант, проверяет ее отпечаток. Административный, аварийный и иной путь изменения должен обеспечивать сопоставимый контроль либо исключаться. Данные отдельного запроса могут выбирать только разрешенную версию конфигурации в пределах утвержденного профиля. Они не должны заменять системную инструкцию, отключать обязательную проверку, ослаблять установленный предел или передавать исполняемый код как критическую для безопасности конфигурацию.

Проверка

Каждая активная конфигурация должна иметь однозначную одобренную версию и подтверждение выбранного способа контроля целостности; несанкционированное изменение и конфигурация с тестовым секретом должны блокироваться. Мера неэффективна при неизвестной активной версии, загрузке неподтвержденной конфигурации или наличии секрета в доступном модели контексте.

Негативный тест

Продвигают контрольную конфигурацию без секрета и варианты с секретом в инструкции, параметре или журнале, затем подменяют версию через административный канал и повторяют при отказе хранилища. Контрольная версия должна запуститься с отдельной ссылкой на секрет, а измененная или непроверяемая конфигурация — не попасть в среду выполнения и не раскрыть значение; возврат восстанавливает точную предыдущую версию.

Подтверждения

Подтверждения включают версию и, если применяется, отпечаток активной конфигурации, правила доступа и продвижения, результат поиска контрольного секрета, трассировку его выдачи во время выполнения, журналы загрузки и отчет теста несанкционированного изменения.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Новые изменения блокируются, система использует последнюю проверенную локально доступную версию. При нарушении целостности запуск или высокорисковая функция блокируются до восстановления.

Исправление и повторная проверка

Восстанавливают управляемую активную версию конфигурации, закрывают неавторизованную загрузку и удаляют секрет из контекста, доступного модели. Повторяют попытку применить неподтверждённую системную инструкцию и прочитать контрольный секрет: версия не активируется, секрет недоступен; сохраняют решение о продвижении и событие отказа.

Побочные эффекты

Строгий процесс замедляет срочные изменения и создает зависимость от хранилищ конфигурации и секретов.

Остаточный риск

Инструкции могут быть выведены косвенно, а одобренная конфигурация может содержать логическую ошибку. Остаточный сценарий — раскрытие или обход правил через поведение модели без изменения защищенной конфигурации.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Lockbox

Частичная поддержка. Lockbox хранит секреты отдельными неизменяемыми версиями, ограничивает доступ через IAM и позволяет переключиться на предыдущую версию.

Ограничение: Защищаются только значения, которые приложение хранит и читает из Lockbox; утверждение и продвижение версии системного промпта или описания инструмента остаются отдельными.

Yandex Audit Trails

Косвенная поддержка. Audit Trails фиксирует создание, изменение, ротацию и версии секретов Lockbox и может дать след изменения защищенной конфигурации секретов.

Ограничение: Audit Trails фиксирует операции с секретами Lockbox; содержимое промпта, его утверждение и блокировку неподтверждённого изменения он не проверяет.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT-GOAL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Версионирует и защищает изменения доверенной политики и конфигурации.

Ограничение: Не моделирует цель задачи и инициатора как объект авторизации.

Условие применения: Если доверенные цели или ограничения агента опираются на системные инструкции и защищённую конфигурацию.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CORE5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Хранит системные инструкции и политики инструментов как защищённую версионируемую конфигурацию с контролируемым продвижением.

Ограничение: Стандарт требует связать хеш артефакта инструкций с редакцией модели или приложения. В карточке контроль по отпечатку является дополнительным вариантом и должен быть включён для этой связи.

Условие применения: При наличии системных инструкций, шаблонов промптов, политик инструментов или другой доверенной конфигурации ИИ-приложения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM08:2026** · Поддерживает часть требования
>
> Проверка конфигурации на секреты и их выдача среде выполнения по отдельному каналу позволяют убрать секретные значения из скрытого контекста модели.
>
> **Граница связи:** Модель всё ещё может раскрыть сами инструкции или их смысл; защита файла конфигурации не блокирует извлечение через ответы.
>
Субъект	Аналитик угроз AI-приложения
Условие применения	Модель имеет доступ к скрытым инструкциям, контексту или описаниям инструментов, которые могут раскрыться.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-010 · Раскрытие содержимого системного промпта и инструкций инструментов	526	107
AISAFE2-IMP-014 · Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных	533	114
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	634	216
AISAFE2-COMP-003 · Управление контекстом общения с моделью	919	40
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	920	44
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	921	61
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	922	67
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	923	75
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	924	84
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1205	328
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1336	466
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1382	554

Мера защиты · AISAFE2-MIT-016

Ограничение доступных инструментов и проверка их аргументов

Краткое описание

Реестр ограничивает доступный агенту набор инструментов и проверяет версию, операцию и аргументы каждого вызова до исполнения.

Механизмы

- Авторизация
- Валидация

Защитный механизм

Реестр описывает инструмент, владельца, версию, классы чтения, записи, исполнения, внешней отправки, оплаты и администрирования, а также схему аргументов. Оркестратор формирует список доступных инструментов на выбранной гранулярности, валидирует предложенный моделью вызов и передает разрешенный объект внешней политике действий.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация выбирает гранулярность списка и проверки по риску действия, устойчивости контекста и доступным компенсирующим мерам. Общая проверка входных данных по схеме может проверять аргументы, но эта мера дополнительно связывает схему с конкретным инструментом, его версией и разрешённой операцией и применяет проверку непосредственно перед исполнением без обходного пути. Мера охватывает учтённые инструменты и версии, но не определяет допустимость бизнес-цели внутри формально разрешенного действия и не изолирует само исполнение.

Предпосылки

- Обязательные предпосылки выбранного контроля — актуальный реестр инструментов, одна или несколько принудительных точек проверки и отсутствие прямого доступа к исполнителю в обход них.

Реализация

Контроль может быть реализован единым диспетчером или эквивалентными распределенными точками, если между моделью и исполнителем нет обходного пути. Прямой вызов коннектора или устаревшей версии в обход выбранного контроля запрещается.

Проверка

Неучтённый инструмент, неизвестное поле, неверный тип и недопустимое сочетание не должны достигать исполнителя. Мера неэффективна при наличии пути от модели к исполнителю вне диспетчера.

Негативный тест

Вызывают разрешенный контрольный инструмент и варианты с исключённым инструментом, устаревшей версией, подмененным описанием или аргументом вне схемы, затем повторяют прямым вызовом исполнителя и при отказе реестра. Контрольный вызов должен пройти, а каждый недопустимый или непроверяемый вариант — завершиться до исполнения без внешнего эффекта.

Подтверждения

Подтверждения включают версию реестра и схем, классификацию эффектов, конфигурацию диспетчера, результаты негативных вызовов и прямого обхода, а также журналы выданного набора инструментов и решений по аргументам.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Агент продолжает работу без инструментов либо завершает задачу; открывать весь набор по умолчанию запрещено. Для обязательной операции возможна передача человеку через отдельный контролируемый процесс.

Исправление и повторная проверка

Удаляют лишний инструмент, сужают разрешённый класс операции или исправляют схему аргументов, пропустившую недопустимый вызов. Повторяют тот же вызов через планировщик и прямую конечную точку: действие не исполняется, разрешённый минимальный вызов сохраняет работоспособность; сохраняют версию списка инструментов и решение авторизации.

Побочные эффекты

Слишком узкий список снижает полезность агента, а частые изменения реестра создают операционную нагрузку и риск несовместимости версий.

Остаточный риск

Формально допустимый инструмент может быть использован для нежелательной цели. Остаточный сценарий — модель выбирает разрешенное действие с допустимыми параметрами, но вредным смыслом; его ограничивает «Независимая авторизация действий агента перед внешним эффектом».

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex API Gateway

Частичная поддержка. Если инструмент вызывается только через API Gateway, валидатор может проверять его HTTP-аргументы по контракту перед передачей исполнителю.

Ограничение: Валидируются HTTP-аргументы только на маршрутах через шлюз; перечень доступных инструментов и отсутствие обхода задаются отдельно.

Yandex Cloud AI Studio

Частичная поддержка. В Responses API запрос задаёт список доступных модели инструментов и схему аргументов FunctionTool; полномочия действия перед исполнением проверяет приложение.

Ограничение: Модель выбирает инструмент из списка в запросе, но полномочия действия и аргументы перед исполнением нужно проверять независимо от модели.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Регистрирует точные инструменты, версии, операции, эффекты и схемы аргументов и блокирует неизвестные или обходные вызовы.

Ограничение: В реестр дополнительно включают доступные данные, получателей и максимальные последствия каждого действия. Класса операции и схемы аргументов для этого недостаточно.

Условие применения: До продуктивного использования агентом инструментов, которые читают, изменяют или публикуют данные либо выполняют иные внешние действия.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CONTENT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ограничивает точные инструмент, версию, операцию и аргументы до исполнения.

Ограничение: Реестр инструментов и проверка аргументов не обеспечивают безопасный рендеринг вывода, запись в хранилище и авторизацию пользователя для последующего действия.

Условие применения: Когда вывод модели используется для выбора инструмента или формирования его аргументов.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MCP1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Реестр фиксирует инструмент, его версию, допустимые операции и схему аргументов и отклоняет неизвестный или некорректный вызов.

Ограничение: Реестр инструментов и проверка аргументов не назначают роли MCP Gateway и целевого сервиса.

Условие применения: При включении инструмента в MCP Gateway и перед передачей ему аргументов.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C10.2.4 · Поддерживает часть требования

Реестр формирует для агента ограниченный набор инструментов по выбранной области полномочий.

Граница связи: Нужно отдельно подтвердить, что MCP `tools/list` возвращает только разрешённые этому клиенту инструменты.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	В MCP-клиенте проверяется, какие инструменты доступны для обнаружения в пределах его полномочий по AISVS 1.0.
Условие связи с карточкой	Доступный агенту список формируется из областей доступа, разрешённых владельцем ресурса, и используется при ответе MCP tools/list.

Открыть положение в источнике

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI02 · Поддерживает часть требования

Разрешённый перечень инструментов и проверка операции, версии и аргументов перед вызовом отсекают часть ошибочных и навязанных действий агента.

Граница связи: Допустимый по схеме вызов может иметь вредную бизнес-цель; мера не проверяет права на объект и не исправляет уязвимость самого инструмента.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент вызывает легитимный инструмент и может применить его вне ожидаемого контракта.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	510	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	519	104
AISAFE2-WEAK-056 · Недостаточная валидация аргументов вызова инструментов	663	246
AISAFE2-WEAK-089 · Избыточные функции и полномочия инструментов агента	710	303
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	925	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	926	44

Карточка	Связи	Стр.
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	927	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	928	77
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1242	324
AISAFE2-MIT-017 · Изолированное исполнение инструментов	1243, 1246	379
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1244, 1295	429
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1254	389

Мера защиты · AISAFE2-MIT-017

Изолированное исполнение инструментов

Краткое описание

Недоверенный код, файлы и инструменты исполняются в отдельной среде с ограниченными ресурсами, сетью, секретами и состоянием.

Механизмы

- Изоляция

Защитный механизм

Диспетчер создает или выделяет изолированную среду, передает только разрешенные входы и необходимые учетные данные, применяет сетевую политику и лимиты и собирает результат через контролируемый канал. Для эфемерного режима состояние уничтожается, а для режима с сохранением состояния его хранение, повторное использование и очистка управляются отдельно.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик

Применимость

Мера обычно применима при исполнении кода, обработке недоверенного файла или запуске компонента, который можно изолировать. Организация выбирает эфемерный или режим с сохранением состояния и глубину изоляции по требуемому состоянию, последствиям выхода из среды и доступным платформенным гарантиям. Она не контролирует семантический эффект законного вызова внешнего сервиса и требует отдельной защиты для браузерного интерфейса. Для простого вызова внешнего API без локального исполнения мера может не применяться.

Предпосылки

- Обязательные предпосылки выбранного контроля — принудительный путь исполнения, идентичность и политики файловой системы, процессов, сети, ресурсов и жизненного цикла состояния, достаточные для заявленной изоляции.

Реализация

Точка принудительного контроля находится в среде запуска, а не в инструкции модели. Общая рабочая папка, выполнение в основной системе и наследование всех переменных окружения недопустимы, если они нарушают заявленную границу изоляции.

Проверка

Каждое исполнение в установленной области должно проходить через изолированную среду. Обязательные тесты не должны получать доступ к файлам основной системы, чужому процессу, невыданному секрету или запрещенному адресу. Мера неэффективна при любом прямом выполнении в основной системе или сохранении неразрешенного состояния.

Негативный тест

Выполняют безвредный контрольный инструмент и вариант с попыткой выхода из среды, исчерпанием лимита и записью контрольного значения, затем повторяют через обход изолированного исполнителя и при отказе оркестратора. Контрольный запуск должен вернуть только разрешённый результат, а вредоносный — не изменить основную систему или соседний процесс; контрольное значение не должно появиться в следующем запуске за другой границей доверия либо после очистки, обязательной по выбранной политике.

Подтверждения

Подтверждения включают профиль изоляции, идентичность и сетевые правила среды, лимиты ресурсов, трассировки запусков, результаты тестов выхода, доступа к секретам и сохранения состояния, а также журнал очистки.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Исполнение инструмента блокируется. Автоматический переход на основную систему или среду с более широкими правами запрещен; возможна ручная обработка в отдельном утвержденном контуре.

Исправление и повторная проверка

Закрывают прямое выполнение в основной системе, сужают границы изолированной среды или исправляют очистку сохраненного состояния. Повторяют исходный выход за границу и запуск следующей задачи: запрещенный ресурс недоступен, остаточное состояние не наблюдается у неуполномоченной следующей задачи; сохраняют конфигурацию изоляции и результаты обоих тестов.

Побочные эффекты

Изоляция увеличивает задержку и стоимость, ограничивает совместимость и создает зависимость от оркестратора среды.

Остаточный риск

Уязвимость изоляции или инфраструктуры может позволить выход, а разрешенное внешнее действие остается реальным. Остаточный сценарий — компрометация через неизвестную уязвимость среды либо вредный эффект внутри разрешенной сети.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Managed Service for Kubernetes

Частичная поддержка. NetworkPolicy ограничивает сетевые связи контейнеров приложений или сервисов, в которых исполняются инструменты, на уровне правил входящего и исходящего трафика.

Ограничение: Сетевая политика не является изолированной средой для недоверенного кода и не изолирует файловую систему, процессы, секреты или состояние; корректная конфигурация остаётся обязанностью владельца.

Yandex Serverless Containers

Частичная поддержка. Позволяет вынести инструмент в отдельную управляемую контейнерную среду без эксплуатации собственной виртуальной машины.

Ограничение: Отдельный контейнер не заявлен как среда для враждебного кода; сеть, секреты, очистку состояния и испытание границы изоляции настраивает владелец.

Yandex Virtual Private Cloud

Частичная поддержка. Группы безопасности позволяют ограничить входящие и исходящие сетевые связи среды исполнения инструмента.

Ограничение: Группа ограничивает только сеть; исходящий трафик в группе по умолчанию разрешён, а процессы, файлы и секреты остаются за другими мерами.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA-ISOL3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Изолирует исполнение кода и инструментов, ограничивает процессы, файлы, сеть и ресурсы и управляет переданными полномочиями и сроком состояния.

Ограничение: Для выполнения AI-INFRA-ISOL3 отдельно задают параметризацию и политику SQL, обязательное завершение по тайм-ауту, очистку временных данных и запрет выдачи продуктивных секретов ради удобства.

Условие применения: Когда модель или агент формирует код, команду или иной запрос, исполняемый в управляемой организацией среде.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.3.1 · Поддерживает часть требования

Изолированная среда ограничивает файлы, сеть, секреты и ресурсы инструмента при локальном исполнении.

Граница связи: Вызовы внешних API и допустимость их эффектов требуют отдельной авторизации; фактическую изоляцию проверяют по каждому инструменту.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает изоляцию инструмента с минимальными полномочиями по его риску.

Открыть положение в источнике

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI05 · Поддерживает часть требования

Изолированная среда ограничивает ресурсы, сеть, секреты и состояние, доступные непредвиденно исполняемому коду агента.

Граница связи: Изоляция ограничивает последствия исполнения кода, но сохраняется риск выхода через уязвимость среды и вредных действий по разрешённым внешним каналам.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент способен получить или выполнить код через управляемую им среду либо инструмент.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	539	118
AISAFE2-WEAK-004 · Отсутствие контекстного кодирования или изоляции	576	148
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	711	306
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	929	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	930	58
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	1243, 1246	375

Карточка	Связи	Стр.
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1245	328
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1247, 1283	419
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	1248, 1304	438
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1249	383
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1296	429

Мера защиты · AISAFE2-MIT-018

Управление машинными идентичностями и минимальными полномочиями

Краткое описание

Каждая AI-нагрузка получает отдельную машинную идентичность и минимальные краткоживущие полномочия для конкретных получателей, ресурсов и операций.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Жизненный цикл

Защитный механизм

Сервис идентификации выдаёт известной нагрузке краткоживущее полномочие, связанное с получателем, операцией, ресурсом и областью применения. Учётные данные доставляются через аутентифицированный канал, привязанный к идентичности нагрузки; делегирование может только сохранять или сужать права. Фактические прямые и унаследованные права, срок, отзыв и отсутствие самостоятельного продления проверяются до использования и при пересмотре.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Мера применяется, если AI-компонент, загрузчик, задание, адаптер, инструмент или иная рабочая нагрузка действует от машинной идентичности, получает либо делегирует полномочия. Период пересмотра определяется критичностью и изменениями архитектуры.

Полностью локальный компонент без доступа к ресурсам и внешним действиям может быть вне области.

Предпосылки

- Нужны инвентарь машинных идентичностей с владельцем, назначением, потребителями и границей доверия и минимальная матрица «субъект—операция—ресурс—получатель».
- Система выдачи должна поддерживать проверяемые срок, получателя и отзыв либо документированный эквивалент с ограниченным периодом пересмотра.
- Канал доставки должен аутентифицировать нагрузку и не раскрывать полномочие соседней нагрузке, пользователю образа или неаутентифицированному пути к сервису метаданных.

Реализация

Полномочия назначаются на минимальный ресурс, а области действия ключей и токенов задаются явно. Нагрузка не управляет другими идентичностями, не назначает роли и не выпускает себе постоянные учётные данные. Краткоживущий токен доставляется через аутентифицированный канал идентичности нагрузки или экземпляра с проверкой получателя; неаутентифицированный устаревший путь к сервису метаданных, пользовательские данные экземпляра, образ, открытая переменная, журнал и исходный пользовательский токен не используются как резервный канал. Пересмотр охватывает прямые, групповые и унаследованные права и запускается после изменения нагрузки, роли, конечной точки, делегирования или владельца.

Проверка

Для выбранной нагрузки восстанавливают эффективные прямые и унаследованные права и сопоставляют каждое с утверждённой операцией. Выполняют разрешённую и соседнюю запрещённую операцию, затем повторяют вызов с истёкшим, отозванным, предназначенным другому получателю и самостоятельно продлённым полномочием. Другая нагрузка и неаутентифицированный клиент сервиса метаданных не должны получить учётные данные; события выдачи, доставки и отзыва связываются с нагрузкой.

Негативный тест

Из соседней нагрузки запрашивают учётные данные через рабочий и устаревший пути к сервису метаданных, затем используют их для другой операции и после отзыва: значение не выдаётся чужой нагрузке, а все запрещённые операции отклоняются.

Подтверждения

Сохраняются инвентарь идентичностей, граф эффективных прав, метаданные роли, области, получателя и срока без значения секрета, схема доставки, события выдачи и отзыва и результаты разрешённого и запрещённых тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Полномочие не выдаётся и действие блокируется. Общий долгоживущий ключ и неаутентифицированный путь к сервису метаданных не включаются как резервный вариант; допускается только отдельная человеческая эскалация.

Исправление и повторная проверка

Разделяют общую идентичность, сужают права, отключают небезопасный путь доставки и отзывают прежние учётные данные. Повторяют попытки чужой нагрузки, соседней операции и вызова после отзыва; они отклоняются, штатная минимальная операция проходит.

Побочные эффекты

Разделение идентичностей и короткий срок увеличивают число объектов, зависимость от сервиса выдачи и требования к синхронизации времени и аварийному восстановлению.

Остаточный риск

Мера не управляет прикладными секретами, не заменяет авторизацию конечного пользователя и не задаёт поставщико-специфичные роли, области или максимальный срок. Сужение полномочий теряется, если следующий компонент игнорирует идентичность, подменяет её общей учётной записью или выпускает новое широкое полномочие вне охваченного пути; ошибочная матрица или скомпрометированная нагрузка также могут злоупотребить формально разрешённым правом до истечения или отзыва.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Identity and Access Management

Прямая поддержка. Сервисные аккаунты и роли позволяют отделить машинную идентичность агента и выдать ей только необходимые права на конкретные ресурсы Yandex Cloud.

Ограничение: IAM ограничивает права на ресурсы Yandex Cloud, но не назначает минимальный набор за владельца, не охватывает внешние системы и не устраняет риск статических ключей.

Yandex Security Deck

Частичная поддержка. CIEM позволяет отзывать у субъектов избыточные права доступа.

Ограничение: CIEM находится на стадии предварительного доступа и не проектирует минимальные роли, не выпускает краткоживущие токены и не подтверждает ограничение получателя, задачи и срока полномочия.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM-LEASTPRIV1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Выделяет отдельную рабочую идентичность, связывает её с минимальными операциями, ресурсами и получателями и проверяет прямые и унаследованные права.

Ограничение: Машинная идентичность не подтверждает полномочия конечного пользователя. Проверку пользователя в IAM или бэкенде и конкретные документированные роли задают отдельно.

Условие применения: Для продуктивных машинных идентичностей AI-компонентов и их доступа к ресурсам Yandex Cloud.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM-LEASTPRIV3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Даёт каждому инструменту или рабочей нагрузке отдельную машинную идентичность, узкую область операций, ресурсов и получателей и короткий срок действия.

Ограничение: Для каждого инструмента отдельно фиксируют вызывающего пользователя или контекст делегирования, целевой сервис и сетевой путь. Машинная идентичность не заменяет проверку пользователя в бэкенде.

Условие применения: Для инструмента агента, который обращается к целевому ресурсу под машинной идентичностью.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM4

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Отдельная машинная идентичность получает краткоживущее полномочие через аутентифицированный канал доставки, связанный с нагрузкой, получателем, операцией и ресурсом; чужое, истёкшее или отозванное полномочие отклоняется.

Ограничение: Выдача машинного полномочия не выбирает IAM-токен или API-ключ и минимальную роль для конкретного вызова AI Studio.

Условие применения: При аутентификации машинного клиента в конкретном API AI Studio.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Пересмотр восстанавливает прямые, групповые и унаследованные права каждой машинной идентичности по расписанию и после изменений, а проверка срока и отзыва исключает сохранение прежнего полномочия после решения.

Ограничение: Пересмотр машинных прав не охватывает доступ людей и внутренние роли всех сред AI-системы.

Условие применения: Для продуктивных машинных идентичностей, в том числе при изменении роли, группы, интеграции или делегирования.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Разделяет идентичности арендаторов и рабочих нагрузок и ограничивает операции и ресурсы.

Ограничение: Машинные полномочия не обеспечивают очистку памяти и временных объектов и не устанавливают все сетевые границы общей среды.

Условие применения: Если несколько систем, арендаторов или классов данных используют общую вычислительную среду.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-RAG-INGEST1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Отдельная идентичность загрузчика получает учётные данные по аутентифицированному каналу и полномочия только на утверждённые операции, источник, индекс и получателя; прямые и унаследованные права пересматриваются по расписанию и после изменений.

Ограничение: Минимальные права загрузчика не проверяют происхождение и содержимое пакета и не принимают решение о его публикации.

Условие применения: Для машинной идентичности загрузчика, который обновляет RAG-корпус или поисковый индекс.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SECRET1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ограничивает машинные учётные данные по получателю, операции, ресурсу и сроку действия.

Ограничение: Не покрывает все типы секретов, их хранение, обнаружение и резервные копии.

Условие применения: Для машинных учётных данных AI-компонента, загрузчика, задания или инструмента.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI03 · Поддерживает часть требования

Отдельная машинная идентичность, краткий срок полномочия и запрет расширять его при делегировании сужают доступ агента к ресурсам.

Граница связи: Полномочия конечного пользователя, человеческие учётные записи и прикладные секреты требуют отдельных проверок.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент использует собственную, делегированную или унаследованную идентичность и связанные права.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM03:2026 · Поддерживает часть требования

Краткоживущие полномочия для конкретных ресурсов и операций ограничивают доступный агенту набор действий при ошибке или злоупотреблении.

Граница связи: Мера не сокращает функциональность инструмента, не проверяет цель разрешённого действия и не заменяет авторизацию от имени пользователя.

Субъект	Аналитик угроз AI-приложения
Условие применения	AI-агент получает инструменты или полномочия, достаточные для действия вне границ конкретной задачи.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	540	118
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	594	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	604	184
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	614	190
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	931	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	932	44
AISAFE2-COMP-009 · Конвейер обучения и дообучения	933	54
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	934	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	935	61
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	936	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	937	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	938	82
AISAFE2-COMP-023 · Подсистема формирования базы знаний	939	86
AISAFE2-MIT-017 · Изолированное исполнение инструментов	1249	379
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	1250, 1302	434
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1251	497
AISAFE2-MIT-064 · Управляемый жизненный цикл секретов AI-системы	1252	554
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1270	405
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1297	429

Карточка	Связи	Стр.
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1358	514

Мера защиты · AISAFE2-MIT-019

Подтверждение человеком высокорисковых действий до исполнения

Краткое описание

Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению.

Область применения

- Общая безопасность

Механизмы

- Человеческий контроль
- Авторизация

Защитный механизм

Блокирующий шлюз показывает уполномоченному человеку точные параметры высокорискового действия, связывает решение с ними и проверяет его актуальность непосредственно перед исполнением. Изменение параметров или повторное использование решения требует нового подтверждения.

Формы контроля

- Техническая
- Организационная
- Выполняемая человеком

Ответственность

- Заказчик

Применимость

Целесообразность определяется классом риска, обратимостью, стоимостью, чувствительностью данных, внешним эффектом и доступными детерминированными ограничениями.

Предпосылки

- Обязательные предпосылки — классификация действия, достоверная идентичность подтверждающего и криптографически либо транзакционно связанный предпросмотр.

Реализация

Система классифицирует действие и формирует предпросмотр с целью, объектом, получателем, данными, стоимостью и последствиями в объёме, зафиксированном для класса действия в формате предпросмотра. Для выбранных классов решение ограничивается сроком

и связывается с неизмененным действием или однозначно определенной группой однородных операций. Повторное использование вне этой области и обход через другой инструмент запрещаются.

Проверка

Проверяются одобрение, отказ, истечение, вводящий в заблуждение предпросмотр, изменение параметров после решения, повтор и альтернативный инструмент. Мера неэффективна, если действие выполняется без действительного решения или решение можно повторно использовать вне связанного с ним действия.

Негативный тест

Попытка выполнить действие без подтверждения, повторно использовать решение или изменить параметры после подтверждения должна быть отклонена до внешнего эффекта.

Подтверждения

Подтверждения включают матрицу действий и ролей, формат предпросмотра, подписанные или транзакционно связанные решения, журналы исполнения и результаты тестов изменения параметров, повтора и обхода.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Высокорисковое действие блокируется. Допускается переход на независимый ручной процесс с отдельной идентичностью и журналом; автоматическое одобрение при недоступности человека запрещено.

Исправление и повторная проверка

Закрывают путь исполнения без решения, связывают подтверждение с неизменяемыми параметрами и запрещают повторное использование. Повторяют исходное высокорисковое действие без подтверждения и после изменения аргумента: внешнего эффекта нет; новый вызов исполняется только после отдельного действующего решения, которое сохраняют как подтверждение.

Побочные эффекты

Шлюз увеличивает задержку и требует доступности уполномоченного подтверждающего. Риски перегрузки и формального одобрения в массовом процессе ограничиваются мерой «Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений».

Остаточный риск

Шлюз защищает только параметры, показанные человеку и принудительно связанные с решением. Даже уполномоченный пользователь может ошибиться, не понять контекст или подтвердить вредное действие под влиянием социальной инженерии.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT-HITL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Показывает уполномоченному человеку параметры и последствия действия, связывает решение с неизменённым предложением и блокирует исполнение без действительного подтверждения.

Ограничение: Стандарт дополнительно требует одноразовую привязку к запрашивающему субъекту, сеансу, сроку и хешу параметров и запрет самоподтверждения. Карточка допускает подтверждение группы операций и не устанавливает все эти условия.

Условие применения: Для действий агента, которым организация назначила обязательное подтверждение человека по классу риска.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.2.1 · Поддерживает часть требования

Блокирующий шлюз связывает высокорисковое действие с явным решением уполномоченного человека до исполнения.

Граница связи: Классы действий, полномочия проверяющего и обходные маршруты определяются и проверяются для конкретной системы.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверенное подтверждение человека до привилегированного или необратимого действия.

Открыть положение в источнике

C9.2.8 · Поддерживает часть требования

Привязка решения к неизменённым параметрам действия и запрет повторного использования поддерживают часть проверки подтверждения по C9.2.8.

Граница связи: Карточка допускает транзакционную привязку. Для C9.2.8 нужна именно криптографическая связь с параметрами, инициатором, контекстом и одноразовым значением.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Для уровня L3 AISVS 1.0 подтверждение криптографически связывается с параметрами действия, инициатором, контекстом исполнения и уникальным одноразовым значением.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	511	101

Карточка	Связи	Стр.
AISAFE2-WEAK-057 · Отсутствие обязательного подтверждения чувствительных действий	664	248
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	940	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	941	44
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	942	65
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	943	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	944	77
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	1211	331
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	1253	320
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	1254	375
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	1255	392
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1256	429
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	1257	438
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	1258, 1326	459
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1352	488

Мера защиты · AISAFE2-MIT-020

Аварийная остановка и возврат агента в безопасное состояние

Краткое описание

Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Механизмы

- Реагирование
- Восстановление

Защитный механизм

Она прекращает дальнейшие шаги автономного агента, отзывает его полномочия, локализует затронутое состояние и возвращает управляемые компоненты к проверенной конфигурации.

Формы контроля

- Техническая
- Организационная

Ответственность

- Заказчик

Применимость

Мера обычно применима к автономным или длительно работающим агентам и их дочерним процессам. Она не отменяет уже совершенный необратимый внешний эффект.

Предпосылки

- Обязательные предпосылки — независимый канал управления, инвентарь дочерних процессов и возможность отзыва полномочий и сетевых соединений.

Реализация

Аварийный контур принимает решение оператора или автоматический сигнал и применяет необходимые для конкретной архитектуры действия: прекращает планирование, останавливает исполнителей, отзывает полномочия, ограничивает исходящие соединения и сохраняет минимально необходимое состояние расследования. Возврат восстанавливает проверенные конфигурацию и зависимости. Точки контроля выбираются в оркестраторе, системе управления доступом, очередях и средах исполнения так, чтобы охватить заявленные пути; отключение только пользовательского интерфейса недостаточно, если агент продолжает действовать.

Проверка

Все компоненты, включенные в область аварийного контура, должны прекратить новые действия в пределах максимального времени, зафиксированного в схеме аварийного контура; отозванное полномочие и закрытая сессия не должны использоваться повторно. Мера неэффективна, если хотя бы один заявленный путь продолжает внешний эффект после этого времени. Восстановленное состояние должно совпадать с выбранной контрольной точкой.

Негативный тест

Во время контрольной цепочки запускают ручную и автоматическую остановку, затем пытаются продолжить через зависшую задачу, дочернего агента и прямой коннектор при недоступном компоненте или потере сети. Все охваченные шаги должны прекратиться или потерять полномочия до нового внешнего эффекта; повторный запуск разрешается только после явного восстановления проверенной конфигурации.

Подтверждения

Подтверждения включают схему аварийного контура, перечень точек остановки, конфигурации отзыва и сетевой изоляции, журналы запуска, отчет технического учения, измеренное время прекращения и результат безопасного повторного запуска.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Применяется инфраструктурная изоляция: блокирование сети, учетных записей и вычислительного ресурса. Новые агентные сессии запрещаются до восстановления аварийного контура; ручная локализация эскалируется немедленно.

Исправление и повторная проверка

Подключают пропущенный путь к аварийной остановке, отзывают сохранившееся полномочие или исправляют возврат состояния. Повторяют тот же сценарий неконтролируемого агента: новые шаги и внешние эффекты прекращаются в пределах максимального времени, зафиксированного в схеме аварийного контура, состояние возвращается к выбранной контрольной точке; сохраняют временную шкалу и сверку состояния.

Побочные эффекты

Остановка может нарушить легитимные процессы, привести к потере незавершенных данных и временно снизить доступность зависимых функций.

Остаточный риск

Остановка не предотвращает первое действие. Остаточный сценарий — внешний эффект уже возник либо неучтенный компонент продолжает работу; требуется последующее реагирование и компенсация.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Identity and Access Management

Частичная поддержка. Отзыв ролей сервисного аккаунта может быть частью аварийной остановки, блокируя последующие действия агента над ресурсами Yandex Cloud.

Ограничение: Отзыв ролей ограничивает последующие обращения к ресурсам, но не останавливает процесс агента, не отменяет уже выполненные действия и не восстанавливает состояние.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Независимый канал прекращает новые действия, отзывает полномочия и переводит охваченные компоненты агента в безопасное состояние.

Ограничение: Для AI-AGENT2 остановку связывают с серверным состоянием, обязательным для всех исполнителей. Лимиты шагов, времени, стоимости и повторов и отказ счётчиков задаются отдельно.

Условие применения: Для автономного или длительно работающего агента, его дочерних процессов и внешних исполнителей.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI08 · Поддерживает часть требования

Независимая остановка агента и отзыв его полномочий могут прервать дальнейшие шаги цепочки после обнаружения распространения ошибки.

Граница связи: Мера не обнаруживает исходный сбой, не отменяет совершённые действия и не останавливает неохваченные зависимости.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Взаимодействующие агенты и зависимые операции могут усилить исходный сбой или повтор.
Условие связи с карточкой	Распространение сбоя обнаружено, а независимый канал остановки охватывает агента и зависимые процессы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	512	101
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	546	123
AISAFE2-IMP-021 · Зацикливание AI-системы при обработке запросов	550	126
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	945	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	946	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	947	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	948	61
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	949	72
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	950	82
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1255	389
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1259, 1279	410
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1359	514

Мера защиты · AISAFE2-MIT-021

Управление доверием и сроком жизни долговременной памяти

Краткое описание

Долговременная память принимает только данные с известным происхождением, областью доверия, владельцем и сроком жизни.

Механизмы

- Жизненный цикл
- Авторизация

Защитный механизм

Она проверяет происхождение и право записи и чтения долговременной памяти, управляет уровнем доверия и сроком жизни и обеспечивает пересмотр, исправление и удаление записей.

Формы контроля

- Техническая

Ответственность

- Совместная ответственность

Применимость

Мера обычно применима к устойчивым предпочтениям, фактам, сводкам, состояниям задач и другим данным, переживающим текущую сессию. Организация определяет метаданные, срок жизни или условие пересмотра и глубину очистки по типу записи и последствиям её ошибочного повторного использования. Заказчик утверждает правила записи, чтения, пересмотра и удаления памяти и проверяет их исполнение. В управляемом сервисе поставщик отвечает за документированные механизмы хранения, изоляции и удаления; в самостоятельно управляемой среде эти обязанности остаются у заказчика.

Предпосылки

- Нужны схема долговременной записи с обязательным происхождением, классом доступа и условием пересмотра или истечения, а также инвентарь индексов, кэшей и копий.
- Чтение должно повторно авторизоваться, а удаление иметь проверяемый срок для каждого слоя хранения.

Реализация

Перед записью контроль проверяет субъект, класс данных, источник и политику; затем помещает запись в подходящий сегмент с заданными правами и сроком жизни либо условием пересмотра. Чтение повторно авторизуется и учитывает уровень доверия. Продвижение из карантина, если он используется, изменение доверия и удаление являются отдельными журналируемыми операциями. Очистка должна охватывать индексы, кэши и резервные копии в пределах установленной политики.

Проверка

Каждая устойчивая запись должна иметь метаданные, которые выбранная политика определяет обязательными, включая происхождение и управляемое условие пересмотра или истечения. Межпользовательское чтение и неавторизованное повышение доверия не допускаются; удаление должно завершаться в срок, зафиксированный в политике жизненного цикла. Мера неэффективна при записи без обязательного происхождения или сохранении доступной записи после срока удаления.

Негативный тест

Записывают разрешенную контрольную память и вариант с недостоверным происхождением, затем читают их в другом сеансе после истечения срока, через прямой индекс и при отказе политики. Разрешенная актуальная запись должна читаться только уполномоченным субъектом, а отравленная, просроченная или непроверяемая — не попадать в контекст; удаление должно охватить кэши и копии в срок, зафиксированный в политике жизненного цикла.

Подтверждения

Подтверждения включают схему записи памяти, политики доступа и жизненного цикла, выборку записей с происхождением, журналы чтения, изменения доверия и удаления, межпользовательские тесты и отчет полной очистки с учетом кэшей и резервных копий.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Запись памяти отключается, а система переходит в режим без долговременной памяти. При недоступности авторизации чтение блокируется; использование общей памяти как резервного пути запрещено.

Исправление и повторная проверка

Исправляют запись без происхождения, повторную авторизацию либо очистку индекса, кэша и копий, оставившую просроченную память доступной. Повторяют исходную запись, межсессионное чтение и удаление: запись без обязательных метаданных не создается, а удаленная запись недоступна в пределах политики; сохраняют события жизненного цикла.

Побочные эффекты

Жесткие правила ухудшают персонализацию, а короткий срок жизни увеличивает число повторных запросов и снижает непрерывность работы.

Остаточный риск

Повторные легитимные сообщения могут постепенно сформировать вредное состояние, а полное удаление из всех резервных копий может быть отложено политикой хранения. Остаточный сценарий — разрешенная, но ложная запись влияет на будущие ответы до обнаружения и исправления.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Managed Service for YDB

Частичная поддержка. TTL позволяет автоматически помечать устаревшие записи памяти и впоследствии удалять их по заданному времени жизни.

Ограничение: TTL удаляет запись с задержкой и не проверяет её происхождение, доверие или влияние на решение агента.

Yandex Object Storage

Частичная поддержка. Правила жизненного цикла могут автоматически удалять объекты долговременной памяти или их неактивные версии по сроку.

Ограничение: Правило действует только для памяти в Object Storage; удаление может задержаться, а доверие к содержимому оно не проверяет.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA-ISOL2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Покрывает намеренно сохраняемое именованное состояние с владельцем, авторизацией чтения и записи, доверием, сроком или пересмотром и удалением кэшей и резервных копий.

Ограничение: Жизненный цикл долговременной памяти не очищает временные файлы, рабочий кэш и память процесса после каждого задания.

Условие применения: Если между сессиями намеренно сохраняются предпочтения, факты, сводки или состояния задач.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает долговременную запись с владельцем, происхождением, правами и сроком, повторно авторизует чтение и очищает связанные индексы, кэши и копии в сроки установленной политики.

Ограничение: Не покрывает временное состояние среды выполнения и сеть.

Условие применения: Если общая среда сохраняет долговременную память пользователей, задач или арендаторов.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C8.2.3 · Поддерживает часть требования

Проверка происхождения и права записи не даёт принять ответ агента или инструмента в долговременную память без установленного источника.

Граница связи: Карта охватывает долговременную память; проверка каждого входящего ответа перед записью должна быть настроена отдельно.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения включает проверку источника до попадания вывода агента или инструмента в доверенную память.

Открыть положение в источнике

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI06 · Поддерживает часть требования

Проверка происхождения, прав записи и чтения и срока жизни долговременной памяти ограничивает закрепление подложных записей между сеансами.

Граница связи: Мера не охватывает разовый контекст, RAG и обучающие данные; разрешённая запись тоже может быть ложной.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	520	104
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	590	173
AISAFE2-COMP-007 · Долговременная память	951	50
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1260, 1308	444
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1319	456

Мера защиты · AISAFE2-MIT-022

Регистрация и сквозное связывание событий AI-цепочки

Краткое описание

Сквозная корреляция связывает значимые события AI-цепочки с субъектом, запросом, версиями компонентов и фактическим результатом.

Область применения

- Общая безопасность

Механизмы

- Наблюдаемость

Защитный механизм

На значимых границах AI-цепочки формируются события с идентичностью субъекта, версией компонента, решением, временем и ссылкой на объект. Сквозной идентификатор связи сохраняется между моделью, RAG, памятью, инструментами и дочерними агентами, а телеметрия доставляется в защищённое хранилище без необоснованной записи полной инструкции или ответа.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Организация выбирает события, детализацию, срок хранения, агрегацию и выборку по последствиям неатрибутируемого действия, требованиям расследования и риску сбора чувствительных данных. Мера не делает наблюдаемым компонент без интеграции и не является правилом обнаружения или процедурой реагирования. Для изолированного исследовательского прототипа без эксплуатационной AI-цепочки, требований расследования и решений, зависящих от событий, мера может не применяться.

Предпосылки

- Обязательные предпосылки — единый сквозной идентификатор связи, согласованные схемы событий, надежное время и защищенная доставка.

Реализация

Для высокочастотных низкорисковых операций допустимы агрегация или выборка, если они не разрушают требуемую атрибуцию и обнаружение. Точки инструментирования выбираются на значимых границах данных и действий. Для охваченных операций ведётся версионируемая матрица «операция → источник события → область → назначение → способ корреляции». Событие по умолчанию содержит идентичность, объект, операцию, решение, идентификатор корреляции и минимальные технические атрибуты, но не полную инструкцию, вывод,

аргументы инструментов, учётные данные или извлечённое содержимое. Сохранение содержимого допускается только в отдельном утверждённом профиле расследования с целью, сроком и кругом доступа.

Проверка

Каждое действие, для которого обязательна сквозная трассировка, должно восстанавливаться от исходного запроса до решения и эффекта; обязательные поля и версии не должны теряться. Проверяются штатная цепочка, отказ и блокировка, дочерний агент, RAG, память, исходящая передача, несогласованные часы, потеря коллектора, высокая нагрузка и маскирование. Мера неэффективна, если такое действие нельзя атрибутировать либо потеря событий не обнаруживается в срок, зафиксированный в конфигурации доставки для этого класса событий.

Негативный тест

Воспроизводят контрольную цепочку, затем теряют, изменяют и переставляют одно обязательное событие и повторяют путь при отказе доставки; отдельно помещают контрольное значение в поле, запрещенное для журналирования. Полная цепочка должна восстанавливаться, повреждённая — обнаруживаться не позже срока, зафиксированного в конфигурации доставки для этого класса событий, и не считаться полной, а контрольное значение не должно появиться в штатных логах, ошибках или диагностическом экспорте.

Подтверждения

Подтверждения включают версию схемы и перечень точек инструментирования, выборочные сквозные трассировки, конфигурации доставки, доступа, хранения и маскирования, метрики полноты и потери событий и результаты отказного и нагрузочного тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При недоступности доставки предусмотренный защищённый буфер сохраняет события; если буфер недоступен, потеря обнаруживается и фиксируется. Для высокорисковых действий потеря телеметрии, необходимой для атрибуции и обнаружения отказа, включает ограниченный режим или блокирование; продолжение требует отдельного временного решения.

Исправление и повторная проверка

Восстанавливают обязательное событие, идентификатор корреляции или безопасное журналирование поля, потерянное на конкретной границе AI-цепочки. Повторяют тот же штатный, ошибочный и повторный запрос: цепочка связывается без разрыва и контрольный секрет отсутствует во всех записях и диагностическом экспорте; сохраняют запросы к журналу и схему события.

Побочные эффекты

Журналы могут собрать чувствительные данные и создать дополнительную поверхность доступа; минимизация снижает детализацию расследования, а высокая детализация увеличивает стоимость хранения.

Остаточный риск

Остаточный сценарий — неинструментированный или скомпрометированный компонент выполняет действие без достоверного события. Нарушение целостности источника способно сформировать убедительную, но ложную трассировку.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud Logging

Частичная поддержка. Централизованно принимает логи сервисов и пользовательские записи, в которые приложение может включать идентификатор AI-цепочки и этапа.

Ограничение: Cloud Logging принимает записи, но идентификаторы цепочки, полноту событий и удаление чувствительных данных обеспечивает приложение.

Yandex Audit Trails

Косвенная поддержка. Настроенный журнал аудита собирает выбранные события уровня сервисов для документированных операций Responses, Foundation Models и MCP Hub. Схемы событий содержат сведения об аутентификации и авторизации, путь ресурса, метаданные запроса и поля конкретной операции.

Ограничение: Собираются только настроенные события облачных сервисов. Состав параметров запроса не подтверждён; внутренние шаги приложения и фактический эффект нужно журналировать отдельно.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AUDIT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ведёт версионируемую матрицу «операция → источник события → область → назначение → корреляция» и обнаруживает потерю обязательных событий.

Ограничение: Для каждой операции дополнительно указывают сервис, уровень события, точное имя события и фильтр источника. Если сервис не публикует нужное событие, его фиксирует приложение или хранилище.

Условие применения: Для существенных операций с моделями, ключами, хранилищами, индексами, MCP и средой исполнения ИИ-системы.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AUDIT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Формирует минимальную коррелируемую запись с идентичностью, объектом, операцией, решением, версией и техническим результатом.

Ограничение: Карточка задаёт минимальную запись и отдельный режим хранения содержимого, но не определяет правовое основание такого хранения и способ удаления; их фиксируют в утверждённом решении.

Условие применения: Для прикладных журналов каждой ИИ-системы; полное содержимое допускается только по отдельному утверждённому решению.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY-PII3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Обеспечивает корреляцию, идентичность, объект, операцию, решение, версию и технический результат.

Ограничение: Не требует всех полей преобразования и детокенизации и не отделяет читателей свидетельств.

Условие применения: Для событий защитного преобразования или обратного преобразования, включённых в журналы ИИ-системы.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MON-TRACE1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Трассирует субъект, запрос, модель, RAG, память, инструмент и эффект с точными версиями и корреляцией.

Ограничение: Стандарт дополнительно требует конкретные поля: рабочую идентичность компонента, целевую точку, код ошибки, длительность и связь платформенного ID с прикладным. Общая схема карточки не задаёт их все.

Условие применения: Для пользовательских и машинных запросов ИИ-приложения, включая RAG и агентные действия.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MON-TRACE2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает значимые события агентной цепочки, решения и результат с субъектом, запросом и версиями компонентов и обнаруживает потерю обязательного события.

Ограничение: Для каждой итерации отдельно задают номер шага, параметры после удаления чувствительных данных, версию плана или подцели и решение продолжить, остановить либо запросить подтверждение. Результат внешнего действия подтверждают событием целевой системы.

Условие применения: Для каждой итерации многошагового агентного цикла; при MCP — с привязкой к событиям шлюза и целевой системы.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает события внешней границы и целевого компонента по субъекту, запросу и идентификатору корреляции.

Ограничение: Корреляция событий даёт часть журнала, но не устанавливает аутентификацию, лимиты и защиту обходного пути публичного маршрута.

Условие применения: Для событий публичного или партнёрского входа и связанного с ним внутреннего компонента ИИ-приложения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 8 (регистрация действий) · Поддерживает часть требования

Сквозная трасса связывает субъект, компонент и операцию с персональными данными на охваченных границах AI-цепочки.

Граница связи: Журнал AI-цепочки не охватывает автоматически действия в БД, резервных копиях и вне приложения; состав событий и защита самих записей проверяются отдельно.

Субъект	Оператор персональных данных
Условие применения	AI-компоненты выполняют действия с персональными данными в ИСПДн.
Условие связи с карточкой	Для охваченных действий события связывают субъекта, объект и операцию; полнота регистрации остальных действий проверяется отдельно.

Открыть положение в источнике

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C12.1.1 · Поддерживает часть требования

Связанные события AI-цепочки позволяют восстановить сессию, участника, версии компонентов и результат действия.

Граница связи: Карточка допускает выбор событий, детализации и выборки. Полноту записи AI-взаимодействий и достаточность контекста сессии нужно проверить для конкретной системы.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	AI-приложение ведёт журнал взаимодействий с контекстом сессии и AI-специфичной телеметрией по AISVS 1.0.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	621	199
AISAFE2-COMP-013 · Телеметрия и обнаружение	952	64
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	1261	342
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1262	405
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1263	410

Карточка	Связи	Стр.
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1264, 1360	514
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1265	519
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1266	525
AISAFE2-MIT-065 · Защищённое сохранение материалов расследования	1267	558
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1355	497
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1390	573

Мера защиты · AISAFE2-MIT-023

Выявление цепочек AI-атак и злоупотреблений

Краткое описание

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции.

Механизмы

- Обнаружение

Защитный механизм

Версионируемый коррелятор получает события, идентичности, контекст сессии и сведения об активах, объединяет их во временном окне, зафиксированном в версии правила и передает сформированный сигнал вместе с исходными событиями по маршруту анализа и реагирования.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера назначается по оценке риска при существенной экспозиции и наличии прямых или косвенных сигналов, позволяющих обнаружить выбранный сценарий. При отсутствии достоверной разметки допускаются прокси-показатели качества, но их неопределенность и пределы обнаружения должны быть описаны явно. Без наблюдаемого сценария или источника сигналов мера не применяется. Она не предотвращает первое действие и не заменяет прямые меры против злоупотребления сервисом «Контроль злоупотребления AI-сервисом» и извлечения модели «Ограничение извлечения модели через интерфейс». Для изолированного исследовательского прототипа без эксплуатационного трафика и обязательства обнаруживать многошаговые злоупотребления мера может не применяться.

Предпосылки

- Обязательная предпосылка — сквозная телеметрия для всех событий, идентичностей и версий, на которых основан выбранный сценарий корреляции.

Реализация

Правило получает сквозно связанные события AI-цепочки, контекст идентичности, базовый профиль и сведения об активах, сопоставляет их во времени и формирует предупреждение с тяжестью, объяснением и маршрутом. Набор может включать последовательности внедрения вредоносных инструкций, злоупотребления штатной функциональностью, извлечения модели, отравления RAG или памяти, обхода полномочий и аномального потребления. Правила тестируются до включения, работают на потоковых и отложенных данных и обновляются после инцидентов. Отдельное ключевое слово без контекста не считается достаточной моделью обнаружения. Для путей, чувствительных к смешению контекстов, правило может сопоставлять проверенную привязку арендатора, субъекта и запроса с владельцем ответа, объекта, пространства имён кэша или контрольной метки среды выполнения и формировать сигнал при несовпадении. Такой сигнал относится только к инструментированному пути и не доказывает изоляцию остальных компонентов.

Проверка

Для каждой версии правила до выпуска фиксируются доступные показатели качества, допустимая доля ложных срабатываний, максимальная задержка и покрываемые источники. Если достоверная разметка отсутствует, используются сценарные повторы, экспертная выборка или косвенные показатели с явным описанием неопределенности. Для контроля смешения контекстов положительный тест подтверждает отсутствие сигнала для корректно связанного объекта. Мера работает частично при неполном покрытии и неэффективна, если выбранный обязательный сценарий не дает предупреждение либо потеря правила не обнаруживается. Близкий нормальный сценарий должен оставаться ниже порога ложных срабатываний, зафиксированного для версии правила.

Негативный тест

Воспроизводят утвержденную многошаговую атаку и близкий легитимный сценарий, затем разбивают атаку между сессиями, замедляют ее и повторяют при потере источника либо несовместимой схеме. Атака должна сформировать один связанный сигнал в пределах максимальной задержки из версии правила, легитимный сценарий — остаться ниже порога из этой версии, а потеря обязательного входа — стать наблюдаемым отказом, а не молчаливым пропуском.

Подтверждения

Подтверждения включают версию правила и связанный сценарий, перечень источников и полей, результаты воспроизведения, журналы предупреждений, доступные метрики качества и задержки и результат проверки резервного набора.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Включается резервный минимальный набор правил и уведомление. Для сценариев, связанных с высокорисковыми действиями, система переходит к усиленному ручному мониторингу или ограничивает соответствующие функции до восстановления.

Исправление и повторная проверка

Исправляют правило корреляции, окно или маршрут сигнала, из-за которых исходная многошаговая кампания осталась незамеченной либо не дошла до анализа. Повторяют ту же последовательность и близкий нормальный сценарий: кампания создает один связанный сигнал с требуемым контекстом, нормальный сценарий не пересекает порог из версии правила; сохраняют версию правила и временную шкалу.

Побочные эффекты

Шум способен перегрузить аналитиков, а строгие правила увеличивают ложные срабатывания и стоимость расследования.

Остаточный риск

Адаптивный субъект может действовать ниже порогов. Остаточный сценарий — вредная последовательность использует легитимные действия и не отличается от нормального поведения доступными сигналами.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex SIEM

Частичная поддержка. Правила анализируют поток событий безопасности и выявляют признаки угроз или аномалий.

Ограничение: Требуются доставка качественной AI-телеметрии и собственные правила; функциональность правил находится на стадии предварительного доступа и не гарантирует полноту обнаружения. Само подключение сервиса не подтверждает эффективность правил обнаружения.

Yandex Cloud Detection and Response

Косвенная поддержка. Сервис собирает данные облачной инфраструктуры для выявления аномалий.

Ограничение: YCDR видит аномалии облачной инфраструктуры; действия модели, промпты и вызовы инструментов требуют отдельной телеметрии и правил.

Yandex Monitoring

Косвенная поддержка. Алерт по пользовательской или сервисной метрике может дать отдельный сигнал об аномальном потреблении или поведении для последующей корреляции с AI-событиями.

Ограничение: Monitoring не содержит готовой модели цепочки AI-атаки, не связывает события промпта, идентичности, инструмента и ресурса и не заменяет SIEM-корреляцию.

Yandex Security Deck

Косвенная поддержка. Threat Detection автоматически обнаруживает подозрительную активность и уведомляет пользователя об обнаруженных угрозах.

Ограничение: Документация подтверждает обнаружение общей подозрительной активности; готовые правила для атак на модель, RAG и промпты не подтверждены. Прикладные события надо передать в сервис отдельно.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MON-ANOMALY2

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Коррелирует версионируемые сигналы с окном, порогом, владельцем и действием.

Ограничение: Порог и окно выбирают для конкретного сценария. Сигнал обнаружения не останавливает нагрузку без отдельного механизма реакции.

Условие применения: Для наблюдаемых сценариев злоупотребления, по которым доступны события или метрики ИИ-системы.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 6 · Поддерживает часть требования

Корреляция событий помогает выявить наблюдаемые цепочки несанкционированного доступа и атак в AI-контуре ИСПДн.

Граница связи: Мера обнаруживает только заданные и наблюдаемые сценарии; предупреждение атаки, ликвидацию последствий и реагирование выполняют другие процедуры.

Субъект	Оператор персональных данных
Условие применения	В AI-контуре ИСПДн возможны несанкционированный доступ к персональным данным и компьютерные атаки.
Условие связи с карточкой	Для выбранных сценариев доступны события, позволяющие обнаружить несанкционированный доступ или атаку.

Открыть положение в источнике

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

ASI10 · Поддерживает часть требования

Корреляция событий агента может выявить заданную цепочку злоупотребления и передать проверяемый сигнал для реакции.

Граница связи: Без наблюдаемого шаблона сигнала не будет; мера не устанавливает скрытую цель агента, не предотвращает первый вред и сама его не останавливает.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент способен сохранять состояние, действовать автономно или использовать полномочия после смены задачи.
Условие связи с карточкой	Для выбранного сценария отклонения агента есть связанные события и проверенное правило обнаружения.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	494	95
AISAFE2-IMP-019 · Чрезмерное потребление AI-системой собственных вычислительных ресурсов	547	123
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	565	139
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	622	199
AISAFE2-COMP-013 · Телеметрия и обнаружение	953	64
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1262	400
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1268	328
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	1269	356
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1270	383
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1271	410
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1272	419
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	1273	434
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	1274	441
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1275, 1332	462
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1276	485
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1277	494
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1278	514
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1391	573

Мера защиты · AISAFE2-MIT-024

Локализация и восстановление после AI-инцидентов

Краткое описание

Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Область применения

- Общая безопасность

Механизмы

- Реагирование
- Восстановление

Защитный механизм

Процедура связывает наблюдаемые признаки AI-инцидента с владельцами решений, командами локализации, сохранением подтверждений, критериями восстановления и обязательной проверкой устранения причины перед возвратом в рабочий режим.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Третья сторона

Применимость

Организация выбирает сценарии, команды, каналы связи и вид учения по риску и фактической архитектуре. Для высокорисковых и технически специфичных сценариев предпочтительно практическое испытание команд; для менее значимых может быть достаточна командно-штабная проверка. Мера не предотвращает первоначальное событие и не может отменить необратимое внешнее действие. Отдельный AI-сценарий не требуется, если архитектура не содержит AI-специфичных активов, сигналов или действий локализации; мера не заменяет реагирование общего назначения.

Предпосылки

- Нужны перечень AI-специфичных активов и сигналов, полномочия и команды локализации, проверенные точки восстановления и критерии отсутствия первопричины после возврата.
- Сценарий учения и предельное время локализации утверждаются до инцидента.

Реализация

После предупреждения или ручного решения сценарий выбирает действия: остановить агента, отозвать токены, изолировать коннектор, поместить источник RAG, память или обучающий набор в карантин, прекратить доступ к похищаемому артефакту, переключить модель или индекс, включить безопасный режим зависимости, заблокировать исходящую передачу и

сохранить состояние. Для подтвержденной утечки сценарий определяет затронутых пользователей и владельцев данных, порядок уведомления, отзыв или ротацию доступных системе учетных данных, токенов и ключей и проверку прекращения доступа; если требуется действие пользователя, уведомление должно назвать его явно. После локализации выполняются расследование, контролируемое восстановление, проверка безопасности и разбор причин с отслеживаемыми действиями. Общий план без технических команд и полномочий недостаточен.

Проверка

В каждом выбранном сценарии фиксируются целевые времена уведомления, локализации и восстановления. Вид и периодичность учения фиксируются для каждого сценария по его риску и должны подтверждать доступность ответственных, выполнимость предусмотренных команд и восстановление без повторной причины. Мера неэффективна, если локализацию нельзя выполнить в целевое время, зафиксированное для сценария или восстановление возвращает причину.

Негативный тест

Запускают учебный AI-инцидент и параллельный контрольный сценарий без признаков инцидента, затем исключают ответственного участника, внешнюю сторону или часть телеметрии и вызывают отказ возврата. Инцидент должен перейти в локализацию и безопасное состояние в срок, зафиксированный для сценария, контрольный сценарий не должен быть остановлен без основания, а восстановление запрещается до проверки устранения причины.

Подтверждения

Подтверждения включают сценарии, контакты и полномочия, исполняемые команды, отчеты проведенных учений, измеренные времена, результат проверки восстановленной среды и отслеживаемые действия по устранению причин.

Поведение при отказе

Блокировка процессом

Сценарий отказа

По умолчанию затронутая AI-функция изолируется и эскалируется руководителю реагирования. При недоступности специализированного сценария используется общий безопасный порядок: остановка новых действий, сохранение данных, отзыв доступа и ручное решение о восстановлении.

Исправление и повторная проверка

Устраняют конкретную причину задержки локализации или повторного появления инцидента, обновляют команду либо шаг восстановления. Повторяют исходное учение с тем же затронутым активом: локализация укладывается в целевое время, зафиксированное для сценария, восстановленный контур не воспроизводит первопричину; сохраняют временную шкалу, команды и проверку состояния.

Побочные эффекты

Локализация может нарушить доступность и уничтожить часть незавершенного состояния; внешняя сторона может отвечать медленнее требуемого.

Остаточный риск

Неполная телеметрия может не позволить доказать устранение причины; к началу локализации данные уже могли быть раскрыты или необратимое действие выполнено, поэтому требуются уведомление, компенсация и предотвращение повторения.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud Backup

Частичная поддержка. Позволяет восстановить защищенные ВМ, серверы и данные после локализации инцидента.

Ограничение: Восстанавливаются только заранее защищённые ресурсы; безопасную точку, целевые сроки и корректность AI-памяти нужно определить и испытать отдельно.

Yandex Cloud Detection and Response

Частичная поддержка. Поддерживает расследование облачного инцидента, анализ событий и работу с оповещениями при локализации AI-инцидента.

Ограничение: YCDR помогает расследовать облачный инцидент, но не останавливает агента, не отзывает все его полномочия и не восстанавливает AI-состояние.

Yandex Identity and Access Management

Частичная поддержка. Отзыв ролей у скомпрометированного субъекта прекращает его последующие разрешённые операции с ресурсами Yandex Cloud.

Ограничение: Отзыв роли не останавливает уже запущенный процесс, не отменяет выполненные действия и не локализует внешние коннекторы или учетные данные вне IAM.

Yandex Virtual Private Cloud

Частичная поддержка. Изменение правил группы безопасности может оперативно закрыть сетевой путь к скомпрометированному AI-компоненту или коннектору.

Ограничение: Сетевая изоляция не отзывает токены, не останавливает прикладной процесс, не помещает данные в карантин и не выполняет контролируемое восстановление.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-FORENSIC2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Реагирование на инцидент сохраняет относящиеся к делу подтверждения и состояние восстановления.

Ограничение: Нет изолированного архива содержимого дела, аудита чтения и экспорта и управляемого удаления.

Условие применения: Если при реагировании сохраняются сообщения, изменения контекста или другие материалы AI-инцидента.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-FORENSIC3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Сохраняет свидетельства инцидента и состояние во время локализации и восстановления.

Ограничение: Нет неизменяемого объекта свидетельств, манифеста контрольных сумм, цепочки хранения и разделения обязанностей.

Условие применения: При сохранении журналов и выгрузок во время локализации и расследования AI-инцидента.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 6 · Поддерживает часть требования

Процедура AI-инцидента связывает признаки атаки с локализацией, сохранением материалов и проверкой безопасного восстановления.

Граница связи: Мера не доказывает обнаружение всех инцидентов ИСПДн и не выполняет отдельные уведомления и взаимодействие с государственными системами.

Субъект	Оператор персональных данных
Условие применения	Компьютерный инцидент или атака затрагивает AI-компоненты ИСПДн и безопасность обрабатываемых персональных данных.
Условие связи с карточкой	Процедура задаёт исполнимые действия локализации, ликвидации последствий и безопасного восстановления затронутого AI-пути.

Открыть положение в источнике

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(c) · Поддерживает часть требования

Сценарий восстановления AI-путей после инцидента помогает вернуть проверенное состояние затронутой обработки.

Граница связи: Нужны отдельные сроки и испытания восстановления доступа ко всем затронутым персональным данным и сервисам.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Начало действия положения	2018-05-25

Открыть положение в источнике

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

MANAGE 2.3 · Поддерживает часть требования

Сценарий AI-инцидента задаёт локализацию, сохранение подтверждений и проверку восстановления при выявлении нового сбоя или злоупотребления.

Граница связи: Карта покрывает инцидентный путь; остальные способы реагирования на новый риск задаются отдельно.

Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	Для конкретной AI-системы определены процедуры реакции и восстановления при ранее неизвестном риске в рамках добровольного NIST AI RMF.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-004 · Передача потребителем чувствительных данных	497	97
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	552, 553	127
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	558	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	566	139
AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление	649	229
AISAFE2-COMP-002 · Модель и интерфейс инференса	954	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	955	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	956	44
AISAFE2-COMP-007 · Долговременная память	957	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	958	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	959	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	960	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	961	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	962	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	963	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	964	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	965	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	966	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	967	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	968	82

Карточка	Связи	Стр.
AISAFE2-COMP-023 · Подсистема формирования базы знаний	969	86
AISAFE2-MIT-009 · Версионирование и откат базы знаний	1220	346
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	1259, 1279	392
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1263	400
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1271	405
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1280	492
AISAFE2-MIT-065 · Защищённое сохранение материалов расследования	1383	558
AISAFE2-MIT-068 · Изолированные неизменяемые версии критичных AI-активов	1386	568
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1392	573

Мера защиты · AISAFE2-MIT-025

Блокирующий шлюз выпуска AI-системы

Краткое описание

Блокирующий шлюз разрешает выпуск только при выполнении утверждённых критериев и наличии свежих подтверждений для точного состава версии.

Область применения

- Общая безопасность

Механизмы

- Шлюз допуска

Защитный механизм

Блокирующий шлюз сопоставляет точный выпускаемый артефакт с обязательными результатами проверок, их сроками и порогами дефектов, зафиксированными в конфигурации шлюза, а также с обоснованными исключениями. При отсутствии действующего результата или изменении артефакта он не создает разрешение на выпуск.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Целесообразность шлюза определяется существенностью изменения, возможными последствиями, нормативными требованиями и доступными компенсирующими мерами. Для низкорисковых изменений может применяться сокращенная проверка или документированное исключение; после выбора шлюза его связь с точным выпуском и блокирующими результатами не должна размываться. Мера не заменяет технические меры защиты и не доказывает отсутствие неизвестных сценариев.

Предпосылки

- Обязательными предпосылками являются определенная область изменения, точный артефакт и актуальные результаты всех применимых мер.

Реализация

Для изменения формируется перечень широко применимых и обусловленных архитектурой или риском мер, точных версий и проверок, выбранных для конкретного выпуска. Для сценария для конечного пользователя в эти входы включаются утвержденные версионированные этические критерии и результаты их проверки; сам шлюз критерии не формирует. Шлюз сопоставляет результаты и исключения и выдает блокирующее решение. Неприменимость результата обосновывается, а для срочного или ограниченного случая может использоваться документированное исключение с владельцем риска и условиями повторной проверки. Точка принудительного контроля находится перед продвижением в рабочую среду; ручное развертывание и изменение объекта после решения несовместимы с выбранным шлюзом. Решение шлюза связывается с точной редакцией архитектуры и модели угроз, составом выпуска, применимыми проверками и действующими исключениями. Любой предусмотренный моделью угроз триггер пересмотра запускает повторное определение области и проверок до следующего выпуска.

Проверка

Каждая применимая обязательная проверка должна иметь актуальный результат; незакрытый результат выше порога из конфигурации шлюза либо отсутствующее подтверждение блокируют выпуск. Мера неэффективна, если рабочий выпуск возможен без успешного решения шлюза.

Негативный тест

Подают контрольный выпуск с полным действующим набором результатов и варианты с дефектом выше порога из конфигурации шлюза, просроченным тестом, необоснованным исключением из области меры или изменением артефакта, затем повторяют через аварийный и прямой административный путь. Только неизменённый контрольный артефакт должен получить разрешение; остальные не должны попасть в рабочий контур без нового полного решения.

Подтверждения

Подтверждения включают область и точный состав изменения, решение о применимости, актуальные результаты проверок, конфигурацию шлюза, привязку решения к артефакту, журнал исключений и результаты тестов прямого обхода, подмены после решения и аварийного выпуска.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Выпуск блокируется, предыдущая проверенная версия остается рабочей. Аварийный выпуск допускается только по ограниченному исключению с компенсирующими мерами, владельцем риска и сроком полной повторной проверки.

Исправление и повторная проверка

Закрывают обход выпуска, возвращают отсутствующий обязательный результат или связывают проверяемый артефакт с решением шлюза. Повторяют тот же выпуск через штатный и обходной маршруты: оба блокируются до устранения дефекта, после чего проходит только идентичный проверенный артефакт; сохраняют входы и итоговое решение шлюза.

Побочные эффекты

Шлюз замедляет выпуск и создает стимул занижать область изменения; недоступность одной проверки может задержать срочное исправление.

Остаточный риск

Шлюз подтверждает только объявленные результаты и их связь с артефактом, но не фактическое применение контроля в среде выполнения. Неизвестный или неохваченный тестами дефект может проявиться после корректно пройденного выпуска.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

SourceCraft Security

Частичная поддержка. Результаты проверок SourceCraft Security можно включить в блокирующую политику CI/CD перед выпуском программной части AI-системы.

Ограничение: Сервис не формирует полный перечень AI-мер, не проверяет модели, данные и конфигурацию среды выполнения и не исключает ручной выпуск за пределами настроенного конвейера.

Yandex Managed Service for GitLab

Частичная поддержка. Обязательные правила проверки могут блокировать интеграцию изменения в защищённую ветку до получения требуемых подтверждений.

Ограничение: Правила блокируют слияние в защищённую ветку, но не фактическое развёртывание и не изменения моделей, промптов и конфигурации вне репозитория.

Yandex Security Deck

Косвенная поддержка. Сканирование контейнерных образов может автоматически запускаться по документированным триггерам.

Ограничение: Модуль находится на стадии предварительного доступа, не блокирует выпуск самостоятельно и не проверяет модели, промпты или бизнес-риски изменения. Результат сканирования не доказывает безопасность выпуска.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-REL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Блокирует выпуск без точного манифеста артефактов и актуальных свидетельств и решений.

Ограничение: Для полного выполнения требования в состав выпуска явно включают версии данных или индекса, модельного комплекта, системных инструкций, конфигурации и образа и сверяют фактическое развёртывание с утверждённым манифестом. Шлюз не исключает неизвестных дефектов проверенной версии.

Условие применения: Перед продвижением точного комплекта ИИ-системы в продуктивную среду.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Блокирует выпуск сторонней модели без действующих результатов требуемой приёмки и связывает решение с точным выпускаемым комплектом.

Ограничение: Шлюз не выполняет проверку условий использования, анализ загрузчика и изолированные поведенческие тесты; содержание приёмки и требуемую независимость решения задают отдельно.

Условие применения: Перед первым продуктивным использованием сторонней модели и после существенного изменения её комплекта.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

> **MANAGE 1.1** · Поддерживает часть требования
>
> Шлюз выпуска связывает решение о развёртывании точной версии AI-системы с утверждёнными критериями и свежими результатами проверки.
>
> **Граница связи:** Он не определяет назначение системы и не заменяет оценку, следует ли продолжать её разработку.
>
Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	Владелец решает, достигает ли конкретная AI-система заявленной цели и допустим ли её выпуск, используя добровольный NIST AI RMF.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	498	98
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	639	222
AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы	708	301

Карточка	Связи	Стр.
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	970	67
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1281	444
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1284	419
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений	1287	423
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	1289	426
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1311	451
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1320	456
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1361	514
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1385	564
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1393	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1397	577

Мера защиты · AISAFE2-MIT-026

Управление уязвимостями AI-компонентов в эксплуатации

Краткое описание

Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Область применения

- Общая безопасность

Механизмы

- Жизненный цикл
- Реагирование

Защитный механизм

Актуальный инвентарь сопоставляется с уведомлениями, результатами сканирования и другими применимыми сигналами о дефектах. Владелец определяет тип результата, оценивает достижимость и тяжесть, назначает действие и срок, развертывает исправление или компенсирующую меру и подтверждает результат.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Организация определяет охватываемые компоненты, типы дефектов, каналы сведений и сроки по критичности и доступности исправлений. Для программных компонентов используются процессы управления уязвимостями, а для моделей, инструкций и конфигураций — соответствующие проверки поведения и изменений. Мера не покрывает неизвестные уязвимости, ошибки бизнес-логики и непрозрачное поведение модели. Мера не применяется к компоненту без управляемого жизненного цикла у внедряющей организации; договорные обязательства внешней стороны фиксируются отдельно.

Предпосылки

- Обязательная предпосылка — актуальный инвентарь компонентов и их владельцев.

Реализация

Периодичность пересмотра зависит от критичности и темпа изменений. Закрытие результата без проверки фактической рабочей версии несовместимо с назначением меры.

Проверка

Сверяют фактические версии рабочего состава с конкретным уведомлением или результатом сканирования, подтверждают достижимость, срок из задачи на исправление и развёрнутое исправление либо компенсацию, затем повторяют исходную проверку на той же версии. Мера неэффективна при просроченном результате, требующем устранения, без принятого решения. Для каждого учтённого компонента проверяются наличие владельца и канала уведомлений.

Негативный тест

Сопоставляют уведомление об уязвимости с точной затронутой версией и с незатронутым контрольным компонентом, затем моделируют недоступное исправление, несовместимое обновление, просроченное исключение и отказ источника. Контрольный компонент должен остаться в работе с зафиксированным обоснованием, а затронутая версия — получить срок, ограничение или блокировку и не закрываться до развертывания исправления либо компенсации и повторной проверки.

Подтверждения

Подтверждения включают актуальный инвентарь, каналы уведомлений и версии сканеров, решения о применимости, задачи и сроки исправления, записи исключений и результат сверки исправленной версии в рабочей среде.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Новые выпуски затронутого компонента блокируются; рабочая функция ограничивается либо защищается компенсирующей мерой. Потеря источника уведомлений переводит состояние в неизвестное и требует эскалации, а не молчаливого продолжения.

Исправление и повторная проверка

Обновляют инвентарь, устраняют уязвимый компонент либо оформляют ограниченное исключение с владельцем и сроком вместо молчаливого пропуска. Повторяют обнаружение на той же версии и попытку эксплуатации применимого сценария: уязвимый выпуск недоступен

или компенсирующая мера срабатывает; сохраняют версию, решение и результат повторной проверки.

Побочные эффекты

Исправление может вызвать регрессию или временный простой, а недоступный патч увеличивает срок и стоимость компенсации.

Остаточный риск

Поставщик может не выпустить патч, а процесс не выявляет неизвестные закладки. Остаточный сценарий — эксплуатируемая неизвестная либо принятая по исключению уязвимость остается доступной до компенсирующего действия.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

SourceCraft Security

Частичная поддержка. SAST, SCA и сканирование секретов дают сигналы о дефектах исходного кода и зависимостей эксплуатируемого AI-компонента.

Ограничение: Сервис не охватывает модель, промпт и фактическую конфигурацию среды выполнения, не подтверждает достижимость находки и не доказывает развёртывание исправленной версии.

Yandex Container Registry

Частичная поддержка. Сканер Container Registry статически проверяет Docker-образы и их программные компоненты, библиотеки и зависимости по базам известных уязвимостей.

Ограничение: Сканирование действует только для контейнерных образов в Container Registry и не охватывает модели, промпты и достижимость во время выполнения; устранение и проверка рабочей версии остаются отдельным процессом.

Yandex Security Deck

Частичная поддержка. Модуль обнаруживает уязвимости в контейнерных образах, хранящихся в Container Registry и Cloud Registry.

Ограничение: Покрывает известные уязвимости контейнерных образов, но не модели, промпты, конфигурационную логику и все программные компоненты; модуль находится на стадии предварительного доступа. Наличие сканирования не подтверждает устранение уязвимостей.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY4

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Связывает находку с точной рабочей зависимостью, достижимостью, владельцем и сроком исправления; исправление, компенсация или исключение проверяются на фактической версии.

Ограничение: Неизвестные уязвимости остаются остаточным риском.

Условие применения: Для программных зависимостей, загрузчиков и образов ИИ-приложения на протяжении управляемой эксплуатации.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Принимает уведомления и сигналы безопасности и запускает оценку достижимости, исправление и повторную проверку.

Ограничение: Изменение условий поставщика, отзыв и недоступность явно не заданы как классы триггеров.

Условие применения: При получении сигналов об уязвимостях управляемых AI-компонентов и программных зависимостей.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM04:2026 · Поддерживает часть требования

Учёт рабочей версии, проверка достижимости уязвимости и повторная проверка исправления уменьшают риск устаревших компонентов цепочки поставки.

Граница связи: Неизвестная закладка, непрозрачная модель и компонент без управляемого владельцем жизненного цикла остаются вне меры.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	647	227
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	695	294
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	971	44
AISAFE2-COMP-009 · Конвейер обучения и дообучения	972	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	973	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	974	58
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	975	67

Карточка	Связи	Стр.
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	976	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	977	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	978	82
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	979	84
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	1232, 1282	364
AISAFE2-MIT-017 · Изолированное исполнение инструментов	1247, 1283	379
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1272	405
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1284	415
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	1285	500
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1286	514
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1312	451
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1394	573

Мера защиты · AISAFE2-MIT-027

Закрепление ответственности и владельцев подтверждений

Краткое описание

Роли и владельцы закрепляются за системой, данными, рисками, эксплуатацией и каждым обязательным подтверждением.

Область применения

- Общая безопасность

Механизмы

- Шлюз допуска
- Управление и надзор

Защитный механизм

До выпуска или принятия сервиса реестр ответственности связывает роли и обязательные подтверждения с фактической архитектурой, границами данных и полномочиями сторон. В небольшой организации один участник может совмещать несколько ролей, если совмещение прозрачно, конфликт интересов оценён и для значимых решений предусмотрена независимая компенсирующая проверка.

Формы контроля

- Процессная
- Организационная

Ответственность

- Заказчик
- Третья сторона

Применимость

Глубина формализации определяется числом сторон, существенностью решений, чувствительностью данных и нормативными требованиями. В небольшой организации допустимо совмещение ролей с прозрачным обоснованием и компенсирующей проверкой там, где независимость действительно нужна. Мера не выполняет технический контроль, не переносит ответственность за управление риском целиком поставщику и не заменяет юридическую проверку договоров. Управление точным текстом, областью, сроком действия и отзывом утверждений относится к поддерживающей мере «Управление подтверждениями и остаточным риском утверждений».

Предпосылки

- Обязательная предпосылка — описанная фактическая архитектура и границы данных и управления.

Реализация

К обязанностям привязываются конфигурации, тесты, журналы, сроки и каналы инцидента. Договорные условия и внутренние назначения должны совпадать с реальными техническими полномочиями. Неразрешенный пробел блокирует высокорисковое решение, если именно эта роль или подтверждение признаны необходимыми.

Проверка

Для каждой функции, признанной необходимой в выбранной области, должны быть назначены конкретные роли и доступные подтверждения. Совмещение ролей фиксируется; для необратимых, разрушительных, платных или привилегированных действий, чувствительных данных и внешней коммуникации оно требует отдельного обоснования и доступной компенсирующей проверки. Проверяются штатный выпуск, смена поставщика, отсутствие владельца, недоступный контакт, конфликт договора и конфигурации, инцидент и пересмотр. Мера неэффективна при неразрешенном пробеле или ответственности, обозначенной только словом «совместно».

Негативный тест

Удаляют владельца одного обязательства, заменяют конкретное разделение словом «совместно» и моделируют недоступность поставщика: выпуск или принятие риска должны блокироваться до назначения конкретного владельца и результата.

Подтверждения

Подтверждения включают выбранное описание ролей, архитектурные границы, назначения и договорные обязательства, фактические технические полномочия, комплект конфигураций, тестов и журналов для каждой стороны, проверку контактов и записи пересмотра.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Выпуск, приемка или внешнее утверждение для затронутой области блокируются. В эксплуатации пробел немедленно эскалируется владельцу риска и временно закрывается назначенным ответственным с ограниченным сроком.

Исправление и повторная проверка

Назначают конкретного владельца действия и подтверждения, разделяют двусмысленную совместную ответственность и закрывают неразрешенный пробел. Повторяют сценарий отказа поставщика или смены ответственного: для каждого обязательства определены принимающий решение, срок и проверяемый результат; сохраняют актуальную матрицу и запись принятия.

Побочные эффекты

Процесс увеличивает административную нагрузку, а избыточное число участников может замедлить решение и размыть фактическое владение.

Остаточный риск

Формально заполненная матрица может не соответствовать реальным полномочиям, а договор не меняет техническое поведение. Остаточный сценарий — сторона не выполняет назначенную обязанность либо реагирует медленнее требуемого, хотя распределение документировано.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Managed Service for GitLab

Частичная поддержка. Code Ownership позволяет закрепить файлы и директории, в том числе выбранные по маске, за определенными пользователями.

Ограничение: Назначаются владельцы файлов в GitLab; роли владельцев данных, риска, эксплуатации и подтверждений вне репозитория остаются отдельными.

Yandex MetaData Hub

Косвенная поддержка. Data Catalog позволяет фиксировать ответственных за данные и использовать это назначение как часть карты владельцев подтверждений.

Ограничение: Владелец данных не заменяет владельцев модели, эксплуатации, безопасности, инцидентов и остаточного риска и не создаёт блокирующее решение при пробеле ответственности.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

> **GOVERN 2.1** · Поддерживает часть требования
>
> Реестр ответственности закрепляет владельцев AI-системы, данных, риска и подтверждений с доступными контактами.
>
> **Граница связи:** Коммуникации и роли по всем функциям управления AI-риском нужно определить на уровне организации.
>
> | | |
> |---|---|
> | **Субъект** | Организация, применяющая NIST AI RMF к AI-системе |
> | **Условие применения** | Для управления рисками AI-системы определены роли, обязанности и каналы взаимодействия в рамках добровольного NIST AI RMF. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	628	207
AISAFE2-WEAK-046 · Недостаточное управление соответствием требованиям	644	225
AISAFE2-COMP-008 · Данные для подсистемы обучения	980	52
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	981	67
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1287	415
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	1288, 1290	426

Мера защиты · AISAFE2-MIT-028

Управление подтверждениями и остаточным риском утверждений

Краткое описание

Реестр утверждений связывает каждое заявление о мере с областью, подтверждением, владельцем, сроком и принятым остаточным риском.

Область применения

- Общая безопасность

Механизмы

- Шлюз допуска
- Наблюдаемость
- Управление и надзор

Защитный механизм

Владелец утверждения регистрирует его точный текст, область, уровень подтверждения, ссылки на артефакты, дату и условия повторной проверки. Перед внутренним решением, аудитом или внешним использованием шлюз проверяет полноту и свежесть записи, её согласование и отсутствие расширения смысла относительно исходных подтверждений.

Формы контроля

- Процессная
- Организационная

Ответственность

- Заказчик

Применимость

Мера обычно применима к внешним, аудируемым и иным значимым утверждениям о мерах защиты, покрытии и эффективности. Для малозначимых внутренних рабочих записей организация может выбрать упрощенный способ при сохранении точной области и источника. Мера не подтверждает техническую работу без исходных тестов и эксплуатационных данных и не определяет продуктовую реализацию.

Предпосылки

- Обязательная предпосылка — прямые подтверждения исходной меры для точной версии, конфигурации, области и периода.

Реализация

Изменение системы автоматически делает затронутые утверждения требующими пересмотра. Мера не реализует исходный технический контроль. Записи о функциональной реализации, качестве на репрезентативной выборке, проверке конкретного развертывания, нефункциональных свойствах и вкладе в соблюдение требований ведутся раздельно. Для каждой фиксируются собственные область, версия, период, подтверждения и ограничения; успешное подтверждение одного типа не переносится на другой.

Проверка

Каждое используемое утверждение должно иметь владельца, область, дату, актуальные артефакты, ограничения и остаточный риск. «Не применимо» допускается только при подтвержденном отсутствии компонента или потока, а не при нехватке данных, прав или телеметрии. Проверяются подтвержденное утверждение в допустимой области, просроченное подтверждение, изменение конфигурации, более широкая формулировка, лабораторный результат, отзыв поставщика и отсутствие владельца. Мера неэффективна, если шлюз допускает утверждение без прямого подтверждения либо не отзывает его после значимого изменения.

Негативный тест

Запись со статусом «не проверено», истёкшим сроком, изменённой конфигурацией или отсутствующим подтверждением не должна считаться действующим утверждением.

Подтверждения

Подтверждения включают точный текст и область утверждения, версии и период, исходные конфигурации, тесты и эксплуатационные данные, запись независимой проверки, срок действия, журнал автоматического отзыва и результаты выборочной повторной валидации.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Утверждение блокируется или отзывается и не используется для выпуска, аудита, клиентского или публичного материала. Техническая функция может продолжать работу только независимо от этого утверждения и согласно собственной оценке риска.

Исправление и повторная проверка

Отзывают просроченное или не относящееся к конфигурации утверждение, добавляют отсутствующее подтверждение и пересматривают остаточный риск. Повторяют использование исходной записи после изменения конфигурации: она не дает статус подтвержденности, новая запись действует только для своей области и срока; сохраняют связь утверждения с подтверждением и решением.

Побочные эффекты

Процесс замедляет публикацию и аудит, а чрезмерно формальная детализация может затруднить использование актуальных подтверждений.

Остаточный риск

Качественная запись не усиливает слабый технический контроль и может устареть между проверками. Остаточный сценарий — актуальное подтверждение верно для выборки, но фактическая рабочая конфигурация отклоняется после проверки без обнаружения.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Object Storage

Косвенная поддержка. Object Lock может сохранить версию файла с подтверждениями и решением об остаточном риске от удаления или перезаписи на заданный срок.

Ограничение: Object Lock сохраняет выбранную версию, но не проверяет истинность, достаточность или актуальность подтверждения; режим и срок защиты задаются отдельно.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-046 · Недостаточное управление соответствием требованиям	645	225
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	982	67
AISAFE2-MIT-027 · Закрепление ответственности и владельцев подтверждений	1288, 1290	423
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1289	415
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1291	519
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1313	451

Карточка	Связи	Стр.
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1321	456

Мера защиты · AISAFE2-MIT-029

Независимая авторизация действий агента перед внешним эффектом

Краткое описание

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст непосредственно перед исполнением действия с внешним эффектом.

Механизмы

- Авторизация

Защитный механизм

Модель предлагает структурированное действие или транзакцию. Механизм политики получает доступные идентичности пользователя и агента, класс действия, объект, получателя, данные, стоимость, происхождение запроса и состояние сессии и выдает «разрешить», «запретить» или «передать человеку».

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера обычно применима, если действие агента меняет внешнее состояние, доступ, стоимость, передает данные или инициирует коммуникацию. Она не понимает весь человеческий смысл операции и не исправляет ошибочную классификацию действия.

Предпосылки

- Обязательными предпосылками являются достоверные идентичности, классификация эффектов и исполнитель, принимающий только решение для неизмененного действия.

Реализация

Исполнитель принимает только решение, связанное с неизмененным действием либо определенной группой операций. Контроль располагается перед каждым действием или транзакцией из заявленной области; прямые вызовы и повтор решения за пределами его области запрещаются.

Проверка

Каждое действие или транзакция с внешним эффектом в области меры должны иметь решение актуальной версии политики; неизвестные высокорисковые действия и измененные параметры должны отклоняться. Мера неэффективна при успешном внешнем эффекте без

требуемого решения либо при разрешении неизвестного высокорискового класса.

Негативный тест

Отправляют разрешенное контрольное действие и запрещенный вариант, затем после решения меняют получателя или аргумент и повторяют через прямой вызов исполнителя, повторное использование решения и отказ механизма политики. Неизмененное контрольное действие должно пройти, а запрещенный, измененный или непроверяемый вариант — не вызвать внешнего эффекта и получить явный отказ либо эскалацию.

Подтверждения

Подтверждения включают версию политики и классификатора действий, конфигурацию точки принудительного контроля, связанные с действием или транзакцией решения и журналы эффектов, результаты тестов неизвестного класса, изменения параметров, повтора и прямого обхода.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Высокорисковое и неизвестное действие блокируется. Для явно определенного низкорискового чтения может использоваться локальное правило минимальных прав; общий режим разрешения при недоступности запрещен.

Исправление и повторная проверка

Закрывают обход внешнего шлюза, исправляют сопоставление класса действия или связывание решения с субъектом и аргументами. Повторяют тот же запрещенный вызов и изменение параметров после разрешения: внешний эффект отсутствует; разрешенный вызов проходит только с отдельным действующим решением, которое сохраняют.

Побочные эффекты

Сложная политика создает конфликты, задержку и операционную нагрузку при разборе неоднозначных решений.

Остаточный риск

Политика видит только переданные ей идентичности, класс и параметры; пропущенный эффект, изменение между решением и исполнением или слишком грубая гранулярность могут разрешить нежелательное действие. Формально разрешённое действие также может использоваться с вредной целью, неразличимой доступными признаками.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex API Gateway

Частичная поддержка. Для авторизации HTTP-запроса API Gateway вызывает указанную в расширении функцию.

Ограничение: Авторизатор вызывается только на охваченных HTTP-маршрутах; политику действия, обходные пути и срок кэширования решения задаёт приложение.

Yandex Identity and Access Management

Частичная поддержка. Перед операцией над ресурсом Yandex Cloud IAM проверяет, есть ли у субъекта необходимые разрешения.

Ограничение: IAM не понимает семантику предложенного AI-действия, стоимость, получателя, данные и происхождение запроса и не покрывает внешние API вне Yandex Cloud.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT-GOAL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Авторизует каждый внешний эффект по доверенным идентичностям и контексту.

Ограничение: Авторизация действия не создаёт новую версию цели и не подтверждает изменение цели задания.

Условие применения: Перед внешним действием агента, которое должно соответствовать доверенной цели и разрешённым возможностям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CONTENT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Независимо авторизует субъект, объект, получателя и параметры до внешнего эффекта.

Ограничение: Авторизация внешнего действия не проверяет схему ответа, контекстное экранирование и обработку инертного текста у получателя.

Условие применения: Когда вывод модели приводит к действию с внешним эффектом: изменению состояния, передаче данных, расходу средств или коммуникации.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM-COMPROMISE1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает авторизацию с пользователем, агентом, объектом, получателем, данными, стоимостью и сессией непосредственно перед внешним эффектом.

Ограничение: Проверка перед эффектом не задаёт срок и аудиторию делегированного токена и не прекращает выданные права после отзыва сессии.

Условие применения: Если агент действует от имени пользователя, арендатора или сессии с отдельными правами.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM-LEASTPRIV3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Независимый механизм проверяет доверенную идентичность, контекст, целевой ресурс и действие перед внешним эффектом.

Ограничение: Карточка не назначает отдельную машинную идентичность каждому инструменту. Для Стандарта дополнительно проверяют исполнение пользовательской или делегированной политики самим бэкендом.

Условие применения: Перед действием инструмента над конкретным внешним ресурсом.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MCP1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Независимо авторизует ресурс, действие и доверенный контекст перед эффектом MCP-инструмента.

Ограничение: Права вызова и управления MCP Gateway задаются отдельно. Для выполнения требования нужно также подтвердить проверку целевого действия в бэкенде.

Условие применения: Перед действием MCP-инструмента в целевом сервисе.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Приказ ФСТЭК России от 11.04.2025 № 117 Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений

Редакция: Приказ № 117 от 11.04.2025 с изменениями по приказу № 137 от 08.05.2026

абзац после п. 61(г) · Поддерживает часть требования

Независимая авторизация действия агента ограничивает его прямое влияние на работу охваченной системы.

Граница связи: Мера не управляет изменением параметров модели и не охватывает AI-пути без действия агента; полномочия проверяются перед каждым внешним эффектом.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	В системе, подпадающей под приказ № 117, агент способен выполнять действия, влияющие на её функционирование.
Условие связи с карточкой	Каждое охваченное действие проверяется независимым механизмом авторизации непосредственно перед внешним эффектом.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 61 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

OWASP Artificial Intelligence Security Verification Standard

Редакция: AISVS 1.0, июнь 2026 года

C9.5.3 · Поддерживает часть требования

Независимый механизм политики разрешает или запрещает предложенное агентом действие перед внешним эффектом без решения модели о доступе.

Граница связи: Карта охватывает действия с внешним эффектом; доступ к данным и внутренним ресурсам проверяется отдельно.

Субъект	Владелец или проверяющий AI-приложения
Условие применения	Уровень AISVS 1.0 для приложения требует решения о доступе в логике приложения или независимом механизме политики.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	513	101
AISAFE2-WEAK-008 · Фундаментальный конфликт между возможностями и безопасностью	580	158
AISAFE2-WEAK-009 · Ненадёжное соблюдение системных инструкций	581	160
AISAFE2-WEAK-010 · Размытие обученных ограничений через контекст	582	162
AISAFE2-WEAK-011 · Поверхностная настройка модели на безопасное поведение	583	164
AISAFE2-WEAK-043 · Недостаточные ограничения в системной инструкции	637	220
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	651	232
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель	659	245
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	665	251
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	983	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	984	44
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	985	61
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	986	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	987	77
AISAFE2-MIT-016 · Ограничение доступных инструментов и проверка их аргументов	1244, 1295	375
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1256	389
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	1292	316

Карточка	Связи	Стр.
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1293	324
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1294	328
AISAFE2-MIT-017 · Изолированное исполнение инструментов	1296	379
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1297	383
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	1298	434
AISAFE2-MIT-031 · Изолированное выполнение браузерных агентов и агентов управления компьютером	1299	438
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	1300	459
AISAFE2-MIT-048 · Аутентификация, целостность и защита от повтора межагентских сообщений	1301, 1356	497
AISAFE2-MIT-066 · Версионирование доверенной цели агента и повторная авторизация	1384	561

Мера защиты · AISAFE2-MIT-030

Ограничение делегированных полномочий коннекторов

Краткое описание

Коннекторы получают только полномочия, необходимые для заданного внешнего вызова, без доступа к исходным пользовательским и внутренним токенам.

Механизмы

- Авторизация

Защитный механизм

Посредник выбирает или выдаёт для коннектора краткоживущее полномочие, связанное с конкретными получателем, заявленной аудиторией, операцией и ресурсом. Коннектор проверяет эти ограничения непосредственно перед вызовом и не принимает полномочие через прямой маршрут в обход посредника.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация выбирает срок, гранулярность и механизм делегирования по возможному эффекту, возможностям протокола и доступным компенсирующим мерам. Мера не исправляет уязвимость удаленного сервиса и не гарантирует безопасный смысл операции внутри выданного полномочия.

Предпосылки

- Обязательные предпосылки выбранного контроля — зарегистрированное происхождение коннектора и точная идентичность получателя; все заявленные делегированные вызовы принудительно проходят через механизм, а известный прямой путь к исполнителю в установленной области отсутствует.

Реализация

Механизм делегирования проверяет идентичность коннектора, требуемую операцию и получателя, после чего выдает или выбирает суженное полномочие для установленной области. Он может быть реализован посредником, шлюзом или эквивалентным механизмом, обеспечивающим принудительное прохождение всех заявленных вызовов. Разрешенные связи между машинной идентичностью и получателем фиксируются, а ответ и последующее перенаправление проверяются в соответствии с риском. Внешний AI API или компонент поставщика считается отдельным получателем: ему не передаются исходные пользовательские или внутренние токены, а маршрут ограничивается разрешенным API и не открывает прямой путь во внутренний контур. Для MCP, OAuth и сетевых интеграций дополнительно учитываются применимые угрозы запуска локального сервера, SSRF, подмены DNS, согласия пользователя и происхождения компонента; конкретный набор механизмов зависит от используемого протокола и архитектуры. Базовый контроль минимальных полномочий задает машинную идентичность агента или среды. Эта мера сужает делегированное полномочие конкретного коннектора для определенных получателя, операции, ресурса, срока и протокола.

Проверка

Исходный пользовательский или внутренний токен не должен попадать модели или коннектору, а полномочие одного получателя не должно приниматься другим. Мера неэффективна при неограниченной передаче исходного токена, вызове вне выбранного механизма или доступности внутреннего инструмента через неучтённый маршрут.

Негативный тест

Выполняют контрольный вызов с краткоживущим полномочием для точных получателя и операции, затем повторяют его с общим или широким полномочием, неверным получателем или неверной заявленной аудиторией. Отдельно проверяют перенаправление, SSRF, повторную отправку и прямой вызов исполнителя при отказе посредника. Только контрольный вызов должен пройти; остальные должны быть отклонены коннектором до обращения к целевому ресурсу.

Подтверждения

Подтверждения включают реестр коннекторов и получателей, конфигурации авторизации и разрешенных связей, фактические параметры срока, области и получателя полномочия, журналы вызовов и результаты применимых тестов прямого, межполучательского и сетевого обхода.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Полномочие не выдается и вызов блокируется. Переход на общий сервисный токен или прямое соединение запрещен; допустима отдельная ручная операция с собственной авторизацией.

Исправление и повторная проверка

Сужают делегированное полномочие до нужного получателя, операции, ресурса и срока либо закрывают прямой вызов коннектора. Повторяют подмену URI перенаправления, подмену получателя и ресурса, а также вызов после отзыва: каждый сценарий завершается отказом; сохраняют параметры выдачи и журналы проверки на стороне коннектора.

Побочные эффекты

Дополнительные обмены увеличивают задержку, а несовместимость версий протокола может временно нарушить работу интеграции.

Остаточный риск

Стандарт и реализации протокола меняются, а внешняя сторона может неверно проверять токен. Остаточный сценарий — скомпрометированный зарегистрированный коннектор злоупотребляет корректно суженным, но все еще достаточным полномочием.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Identity and Access Management

Частичная поддержка. Привязка связывает одну федерацию сервисных аккаунтов, один сервисный аккаунт Yandex Cloud и одного внешнего субъекта.

Ограничение: Федерация связывает внешнего субъекта с сервисным аккаунтом Yandex Cloud; права внешнего коннектора и минимальные роли на его стороне задаются отдельно.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MCP1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Использует узкие делегированные учётные данные для получателя, действия, ресурса, срока и протокола.

Ограничение: Делегирование полномочия коннектору не разделяет права вызова, изменения и привязки MCP Gateway и права целевого сервиса.

Условие применения: При делегировании полномочий коннектору, который вызывает инструмент через MCP Gateway.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SECRET1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ограничивает делегированные учётные данные коннектора, аудиторию, срок и протокол.

Ограничение: Не покрывает общую инвентаризацию и жизненный цикл секретов.

Условие применения: При выдаче или выборе делегированного полномочия для внешнего коннектора.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI03** · Поддерживает часть требования
>
> Сужение полномочий внешнего коннектора ограничивает один путь злоупотребления делегированной идентичностью агента.
>
> **Граница связи:** Мера действует на охваченные коннекторы; права других агентов и корректность проверки токена внешней стороной оцениваются отдельно.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент использует собственную, делегированную или унаследованную идентичность и связанные права.
Условие связи с карточкой	Агент обращается к внешнему коннектору с делегированным полномочием.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	541	118
AISAFE2-WEAK-058 · Вызов инструментов без учёта идентичности и полномочий инициатора	666	251
AISAFE2-WEAK-092 · Недостаточно ограниченные делегированные полномочия пользователя у агента	714	310
AISAFE2-WEAK-095 · Недостаточно ограниченные делегированные платёжные полномочия	717	314
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	988	44
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	989	61
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	990	69
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1250, 1302	383
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1273	405

Карточка	Связи	Стр.
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1298	429
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	1303, 1306	441

Мера защиты · AISAFE2-MIT-031

Изолированное выполнение браузерных агентов и агентов управления компьютером

Краткое описание

Браузерный агент и агент управления компьютером работают в отдельной среде или профиле с ограниченными данными, сетью и действиями интерфейса.

Механизмы

- Изоляция

Защитный механизм

Браузерный агент выполняется в отдельном профиле без произвольного доступа к пользовательским файлам cookie и секретам. Для агента управления компьютером эквивалентную границу задают отдельная среда выполнения и ограниченные права на файлы, окна и действия. Политика навигации, контроль загрузки и выгрузки, ограничения файлов и буфера обмена и отдельные учетные данные применяются внутри этой границы, а выбранные действия передаются во внешние шлюзы подтверждения и авторизации.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик

Применимость

Организация определяет глубину изоляции, допустимое состояние и внешние шлюзы по доступным данным, учётным записям и последствиям действий в пользовательском интерфейсе. Карточка намеренно ограничена браузерными агентами и агентами управления компьютером. Для инфраструктурных и иных агентов применяются общая изоляция исполнения, отдельная машинная идентичность с минимальными полномочиями, ограничения сетевых путей и независимая авторизация действий. Она не гарантирует правильное понимание интерфейса и не покрывает ручное действие пользователя вне изолированной сессии.

Предпосылки

- Обязательные предпосылки выбранного контроля — отделенная среда, ограниченные полномочия навигации, учетных данных и файлов и отсутствие пути в основной контур в обход этих границ.

Реализация

DOM, визуальный контекст и действия журналируются в объеме, необходимом для проверки и расследования. Использование основного пользовательского профиля недопустимо, если оно разрушает заявленную изоляцию.

Проверка

Сессия не должна видеть основной профиль, невыданные учётные данные, запрещённые сайты и файлы. Все действия в пользовательском интерфейсе, которые изменяют доступ, передают чувствительные данные либо создают платный, разрушительный или необратимый внешний эффект, должны проходить внешний шлюз. Мера неэффективна при доступе к запрещённому ресурсу или действии без шлюза.

Негативный тест

Открывают разрешённую контрольную страницу и страницу с визуальным внедрением инструкций или внедрением через DOM, требующим прочитать файл, данные cookie либо выполнить выгрузку, затем повторяют через новый профиль и при отказе браузера. Контрольная навигация должна завершиться в выделенной среде, а запрещённые данные и действие — остаться недоступными без отдельного шлюза; после отказа не должно сохраняться переносимое состояние или внешнее действие.

Подтверждения

Подтверждения включают конфигурацию профиля, сети, сайтов, файлов и буфера обмена, трассировки DOM, визуального контекста и действий, результаты сценарных атак, журналы решений внешнего шлюза и подтверждение очистки сессии и отзыва учетных данных.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Сессия завершается, учетные данные отзываются, файлы изолируются. Автоматический переход в основной браузер или продолжение без политики запрещены.

Исправление и повторная проверка

Закрывают доступ браузерной среды или среды управления компьютером к запрещённому ресурсу, очищают состояние либо проводят действие через требуемый шлюз. Повторяют исходную навигацию и попытку сохранить состояние между сессиями: запрещённый ресурс и прежние данные недоступны, внешнее действие не совершается без решения; сохраняют трассировку среды.

Побочные эффекты

Ограничения снижают совместимость и производительность, а эфемерный профиль усложняет работу с легитимными длительными сессиями.

Остаточный риск

Адаптивная страница может скрыть смысл действия или влиять на модель через визуальный контент. Остаточный сценарий — агент выполняет разрешённое действие в пользовательском интерфейсе на скомпрометированной странице, смысл которого не распознан.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Virtual Private Cloud

Частичная поддержка. Правила групп безопасности задаются отдельно для входящего и исходящего трафика.

Ограничение: Сетевая группа не видит домен, DOM и действие браузера; в группе по умолчанию исходящий трафик разрешён.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI05** · Поддерживает часть требования
>
> Изолированная среда браузерного агента или агента управления компьютером ограничивает часть последствий неожиданного исполнения кода в этой сессии.
>
> **Граница связи:** Мера не охватывает другие интерпретаторы, серверные инструменты и вызовы кода вне этой сессии; границу изоляции нужно испытать.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент способен получить или выполнить код через управляемую им среду либо инструмент.
Условие связи с карточкой	Агент управления браузером или компьютером способен выполнить код через свою среду.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-016 · Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)	542	118
AISAFE2-WEAK-090 · Недостаточная изоляция исполнения инструментов и сгенерированного кода	712	306
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	991	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	992	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	993	61
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1206	328
AISAFE2-MIT-017 · Изолированное исполнение инструментов	1248, 1304	379

Карточка	Связи	Стр.
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1257	389
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1299	429

Мера защиты · AISAFE2-MIT-032

Проверка исходящих данных перед передачей через коннекторы

Краткое описание

Исходящий шлюз проверяет назначение, получателя, класс и минимально необходимый объём данных до передачи через коннектор.

Механизмы

- Минимизация данных
- Авторизация

Защитный механизм

Контроль получает фактически передаваемые данные после всех преобразований, идентичность, получателя, действие и происхождение. Он минимизирует содержимое, блокирует запрещенные классы и контрольные секреты, направляет чувствительную передачу по выбранному маршруту подтверждения и формирует журнал решения.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера обычно применима при передаче данных внешней модели, инструменту, коннектору, получателю сообщения или иному недоверенному контуру. Она не распознает все смысловые, кодированные или распределенные способы утечки и не контролирует неучтенный канал.

Предпосылки

- Обязательные предпосылки — достоверный получатель, происхождение данных и классификация чувствительности в точке отправки.

Реализация

Проверка может быть реализована централизованным шлюзом или распределенными обертками коннекторов, если подтверждено покрытие всех заявленных исходящих путей и перенаправлений. Секреты и чувствительные значения не должны сохраняться в параметрах, диагностике или журналах обходного пути; проверка только исходной инструкции недостаточна.

Проверка

Каждая внешняя передача в установленной области должна иметь решение. Мера неэффективна при существовании неучтенного исходящего пути или успешной передаче контрольного секрета.

Негативный тест

Передают разрешенный контрольный объект и объект с уникальным синтетическим секретом, затем кодируют, суммаризируют и разбивают секрет между сообщениями и повторяют через перенаправление, прямой вызов и отказ классификатора. Контрольный объект должен дойти только до заданного получателя, а секрет — не появиться ни у одного получателя; непроверяемая передача должна блокироваться или уходить на утвержденный безопасный маршрут.

Подтверждения

Подтверждения включают инвентарь получателей и каналов, версию политики исходящих данных, конфигурации шлюза или оберток, журналы решений без сохранения секрета, результаты тестов с контрольными значениями и проверку прямых и перенаправленных путей.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Исходящая передача блокируется. Допускается локальное сохранение минимального результата для последующей ручной обработки; отправка с дополнительным мониторингом без проверки требует отдельного ограниченного решения владельца риска.

Исправление и повторная проверка

Исправляют классификацию, минимизацию или привязку получателя, из-за которых запрещенные данные ушли через коннектор. Повторяют отправку той же синтетической записи штатным и обходным путем: запрещенное поле не передается, а разрешенный минимальный набор достигает только выбранного получателя; сохраняют решение шлюза без исходного секрета.

Побочные эффекты

Проверка повышает задержку, может блокировать легитимную коммуникацию и требует осторожного журналирования, чтобы не создать новую копию чувствительных данных.

Остаточный риск

Семантически чувствительные данные могут быть выражены без известных признаков или распределены по нескольким разрешенным сообщениям. Остаточный сценарий — утечка через разрешенный канал в форме, не распознанной классификатором.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет получателя, конечную точку, назначение, класс, поля и минимальный состав данных до внешней передачи.

Ограничение: Исходящий шлюз проверяет получателя и состав, но не устанавливает правовое основание, роли сторон, договор и условия последующей обработки.

Условие применения: Перед передачей персональных или иных конфиденциальных данных внешней модели, API, SaaS-инструменту или другому получателю.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY4

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет получателя, конечную точку и состав исходящей передачи.

Ограничение: Не определяет страну, правовое основание, уведомление или договор.

Условие применения: На исходящем участке потока, который уполномоченная функция признала трансграничным или подпадающим под ограничения размещения.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Минимизирует исходящий набор данных и блокирует запрещённые поля и секреты перед внешним получателем.

Ограничение: Не аутентифицирует и не защищает канал и не управляет учётными данными поставщика.

Условие применения: Перед отправкой данных из Yandex Cloud внешнему модельному API через контролируемый коннектор.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM02:2026 · Поддерживает часть требования

Проверка получателя, класса и объёма фактически исходящих данных может остановить передачу закрытых полей через коннектор.

Граница связи: Ответы и журналы вне проверяемых исходящих маршрутов требуют отдельных мер; скрытые или распределённые чувствительные сведения могут пройти классификатор.

Субъект	Аналитик угроз AI-приложения
Условие применения	AI-приложение обрабатывает сведения, которые модель или приложение может раскрыть неуполномоченному получателю.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-009 · Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов	524	106
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	559	130
AISAFE2-WEAK-039 · Выдача инструментом чувствительных или избыточных данных	632	212
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	994	44
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	995	69
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	996	77
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	1224, 1305	349
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1274	405
AISAFE2-MIT-030 · Ограничение делегированных полномочий коннекторов	1303, 1306	434
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1333	462
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1362	514

Мера защиты · AISAFE2-MIT-033

Шлюз выпуска изменений RAG, векторных хранилищ и памяти

Краткое описание

Блокирующий шлюз разрешает выпуск изменений RAG, векторного хранилища или памяти только при выполнении утверждённых критериев и наличии свежих подтверждений для точного состава версии.

Механизмы

- Шлюз допуска
- Целостность

Защитный механизм

Шлюз собирает для точной версии RAG, векторного хранилища или памяти результаты всех применимых проверок и их сроки, проверяет обоснования неприменимости и связывает итоговое решение с выпускаемым артефактом. Отсутствующий, просроченный или несвязанный результат не создает разрешение на выпуск.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Целесообразность шлюза определяется значимостью изменений, архитектурой контура знаний или памяти и последствиями ошибочного выпуска. После выбора шлюза его блокирующее свойство сохраняется: он не заменяет содержательную проверку связанных мер, а неприменимость каждого результата должна быть обоснована.

Предпосылки

- Обязательны актуальные результаты всех применимых проверок допуска источников, авторизации извлечения, версионирования и возврата, минимизации, изоляции и управления памятью; каждый результат связан с точной версией и имеет действующий срок.

Реализация

Шлюз получает состав изменения, результаты требуемых тестов, подтверждение телеметрии и возврата, список исключений и владельцев. Он определяет применимые связанные меры защиты, проверяет свежесть их подтверждений и выдает решение «выпустить» или «заблокировать». Сам шлюз не выполняет авторизацию, фильтрацию или изоляцию. Точка принудительного контроля находится в процессе продвижения рабочего выпуска и не должна обходиться прямым административным изменением.

Проверка

Метод проверки сопоставляет точную версию выпуска с решением специализированного шлюза и трассирует прямой, служебный и откатный пути до рабочей базы знаний или памяти.

Негативный тест

Подают контрольное изменение с полным набором актуальных результатов и варианты с просроченной авторизацией, отсутствующим проверенным возвратом или необоснованным исключением из области меры, затем повторяют через аварийный и прямой административный путь. Только точная контрольная версия должна пройти шлюз; остальные не должны изменить рабочие RAG, индекс или память до устранения причины и нового решения.

Подтверждения

Подтверждения включают точный состав изменения, обоснование применимости, ссылки на актуальные результаты связанных мер, запись решения шлюза, результаты теста административного обхода, журнал исключений и проверку возврата к предыдущему выпуску.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Выпуск блокируется, рабочей остается предыдущая проверенная версия. Аварийное изменение допускается только по ограниченному исключению с компенсирующим контролем, владельцем риска и обязательной последующей проверкой.

Исправление и повторная проверка

Возвращают отсутствующий результат проверки, исправляют ошибочно заданную область неприменимости либо закрывают прямое продвижение изменённого RAG, индекса или памяти. Повторяют выпуск той же версии через штатный и обходной маршруты: он блокируется до действующих решений, после исправления проходит только связанный с ними артефакт; сохраняют манифест и решение шлюза.

Побочные эффекты

Шлюз может стать формальным чек-листом или использовать устаревшие результаты. Он увеличивает время выпуска и зависит от качества связанных проверок.

Остаточный риск

Шлюз не проверяет качество источников, авторизацию, очистку памяти или изменения после продвижения без отдельного контроля целостности. Даже при выполнении всех известных условий новая атака или ошибка данных может не входить в тестовый набор.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Audit Trails

Косвенная поддержка. Аудитные события конфигурации и поддерживаемые события уровня сервисов могут служить внешними подтверждениями для пользовательского шлюза выпуска.

Ограничение: Audit Trails даёт выбранные события конфигурации, но не принимает решение о выпуске и не блокирует обходной административный путь.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-RAG-INGEST1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Блокирует публикацию содержимого RAG, векторного хранилища и памяти без требуемых проверок и точной версии.

Ограничение: Шлюз публикации не ограничивает чтение источника и запись загрузчика в другие области индекса или хранилища.

Условие применения: При выпуске изменений RAG-корпуса или поискового индекса.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI06** · Поддерживает часть требования
>
> Блокирующий шлюз выпуска проверяет утверждённые условия перед продвижением новой версии базы знаний или памяти агента.
>
> **Граница связи:** Шлюз действует только на охваченные пути и известные проверки; вредный смысл и запись после выпуска требуют отдельных контролей.

Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агент сохраняет память или контекст и может повторно использовать недоверенное содержимое.
Условие связи с карточкой	Изменения постоянного RAG-корпуса, векторного хранилища или памяти проходят шлюз выпуска.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-011 · Компрометация баз знаний	528	109
AISAFE2-COMP-007 · Долговременная память	997	50
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	998	67
AISAFE2-COMP-023 · Подсистема формирования базы знаний	999	86
AISAFE2-MIT-007 · Допуск проверенных источников в контур RAG	1216, 1307	339
AISAFE2-MIT-008 · Авторизация извлекаемых данных на минимальной значимой гранулярности	1217	342
AISAFE2-MIT-009 · Версионирование и откат базы знаний	1221	346
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	1225	349
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	1228	356
AISAFE2-MIT-021 · Управление доверием и сроком жизни долговременной памяти	1260, 1308	396
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1281	415
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1322	456

Мера защиты · AISAFE2-MIT-034

Контролируемое продвижение моделей через реестр

Краткое описание

Управляемый выпуск связывает рабочую среду с проверенной версией, позволяет воспроизводимо вернуть предыдущую и отзывает небезопасную версию на всех учтённых путях.

Механизмы

- Жизненный цикл
- Шлюз допуска

Защитный механизм

Управляемый процесс разрешает смену активной модели только как проверяемый переход между точными неизменяемыми версиями. Для каждой версии инвентаризируются псевдонимы, конечные точки, прямые ссылки, кэши и разрешённые пути загрузки. Решение об отзыве атомарно закрывает эти пути, переводит рабочие ссылки на заранее проверенную безопасную версию либо останавливает функцию и блокирует повторное продвижение отозванного артефакта.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Мера применяется при управляемом продвижении, замене, откате или отзыве версии модели. Она охватывает все способы, которыми рабочая среда может выбрать или загрузить модель; качество версии и полный вывод остальных данных, учётных данных и интеграций регулируются отдельно. Для неизменяемого сценария, где рабочая модель никогда не заменяется и отсутствуют стадии или среды продвижения, мера не применяется.

Предпосылки

- Требуются точная идентификация артефакта, управляемые роли, согласованный безопасный набор для замены и проверяемый способ смены и возврата рабочей версии.
- Для каждой версии должен существовать инвентарь псевдонимов, конечных точек, прямых ссылок, кэшей и иных путей, способных продолжить её загрузку.

Реализация

Запрос продвижения, отката или отзыва указывает точный дайджест или эквивалентный неизменяемый ID, текущую и целевую стадии, результаты проверок, инициатора и одобряющего. Отзыв закрывает маршрутизацию, псевдонимы, прямые ссылки, манифесты развёртывания и кэшированные разрешения на охваченных путях; один изменённый

псевдоним не считается полным отзывом. Если совместимая безопасная версия не подтверждена, затронутая функция останавливается. Прямое изменение рабочей ссылки или повторная загрузка отозванной версии в обход реестра запрещены.

Проверка

Для контрольной версии перечисляют все рабочие ссылки и сначала подтверждают разрешённый вызов каждой. После отзыва повторяют вызов через каждый псевдоним, конечную точку, прямой ID, кэш и резервный маршрут; версия не должна загрузиться или исполниться. Отдельно подтверждают безопасную замену или остановку и невозможность повторного продвижения без нового явного решения.

Негативный тест

После отзыва пытаются вызвать модель по прежнему псевдониму, точному ID, прямой ссылке и из кэша, затем повторно продвинуть тот же дайджест: все операции блокируются, безопасная версия либо остановленное состояние сохраняются.

Подтверждения

Сохраняются политика переходов и отзыва, точный ID артефакта, инвентарь ссылок, решение, фактические версии сред, результаты тестов каждого пути и отчёт замены или остановки.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Неполный или неатомарный отзыв запрещает дальнейшее использование затронутой функции; система не продолжает с неизвестной или смешанной версией и не возвращается к отозванной версии как к резервной.

Исправление и повторная проверка

Закрывают обнаруженный обход, очищают кэш или исправляют связь реестра со средой, затем повторяют вызов по каждому прежнему пути и повторное продвижение: отозванная версия остаётся недоступной, безопасная контрольная версия работает.

Побочные эффекты

Срочный отзыв может остановить функцию, нарушить совместимость клиентов и потребовать очистки распределённых кэшей и согласованной смены нескольких конечных точек.

Остаточный риск

Реестр не удаляет копии у внешнего получателя и не заменяет полный вывод из эксплуатации. Неучтённая копия, скрытый путь или скомпрометированная управляющая плоскость могут продолжить исполнение отозванного либо продвинуть вредоносный артефакт.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex DataSphere

Частичная поддержка. Псевдоним DataSphere предоставляет стабильную конечную точку и позволяет заменять связанные узлы при обновлении развёрнутого сервиса.

Ограничение: Псевдоним переключает опубликованный сервис, но не проверяет версию модели, испытания, полномочия одобряющего и допустимость перехода.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-REL2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Отзыв блокирует точную версию на всех учтённых псевдонимах, конечных точках, прямых ссылках, кэшах и обходных путях, после чего переводит нагрузку на проверенную замену либо безопасно останавливает функцию.

Ограничение: Отзыв версии модели не задаёт совместимость и проверенную цель отката модели, индекса и конфигурации как единого выпуска.

Условие применения: При продвижении, замене, откате или отзыве управляемой версии модели.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM04:2026 · Поддерживает часть требования

Привязка рабочих ссылок к точной версии позволяет отозвать скомпрометированный артефакт модели и вернуть проверенную версию.

Граница связи: Откат происходит после обнаружения; он не удостоверяет безопасность исходной версии и не удаляет внешние копии.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

LLM05:2026 · Поддерживает часть требования

Управляемый отзыв версии модели позволяет остановить использование выявленного отравленного выпуска и вернуться к проверенному предыдущему.

Граница связи: Мера не обнаруживает отравление и не очищает источник данных или неучтённые копии модели.

Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	535	116
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	569	141
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	640	222
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	699	296
AISAFE2-COMP-002 · Модель и интерфейс инференса	1000	37
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1001	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1002	67
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	1229	359
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1236	368
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1309	509
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1363	514
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1373	525

Мера защиты · AISAFE2-MIT-035

Ведение проверяемого состава и происхождения компонентов AI-системы

Краткое описание

Проверяемый состав связывает рабочую версию AI-системы с происхождением моделей, кода, зависимостей, данных и конфигурации.

Область применения

- Общая безопасность

Механизмы

- Наблюдаемость
- Целостность

Защитный механизм

Конвейер сборки и развертывания формирует неизменяемый манифест из идентификаторов и версий модели, данных, кода, инструментов и результатов оценки. После выпуска механизм сверки сопоставляет этот манифест с фактическим составом среды выполнения и фиксирует

расхождение.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Организация определяет минимальный и расширенный состав по архитектуре, последствиям неизвестного компонента и требованиям расследования. Автоматический сбор предпочтителен для быстро меняющихся систем, но допустим эквивалентный проверяемый процесс. Мера не доказывает безопасность компонентов и не заменяет их профильные проверки. Для одноразового локального исследования, которое не выпускает, не передаёт и не сохраняет составной артефакт AI-системы, мера не применяется.

Предпосылки

- Нужны точная идентификация развертывания и управляемые источники версий моделей, данных, кода, инструментов и результатов оценки.
- Минимальный состав должен позволять воспроизвести выпуск и определить область инцидента без ручного предположения.

Реализация

Состав формируется автоматически либо через иной проверяемый процесс при загрузке данных, обучении, сборке, оценке и развертывании и связывает фактические версии с рабочим артефактом. Для систем с повышенными требованиями состав может подписываться и сверяться с рабочей средой автоматически. По идентификатору значимого компонента должна определяться область затронутых развертываний. Формат и полнота инвентаря являются параметрами реализации, а не самой мерой защиты. Для наборов данных инвентарь дополнительно фиксирует владельца, источник, назначение, класс, версию, места хранения, потребителей, срок хранения и долю синтетических данных. Для синтетической части указываются генератор и его версия, правила отбора и несинтетическая контрольная выборка. Одни только метаданные хранилища не подтверждают происхождение данных.

Проверка

По тестовому уязвимому компоненту должны определяться затронутые развёртывания. Мера неэффективна при невозможности определить фактический состав.

Негативный тест

Развертывают контрольный состав, совпадающий с манифестом, затем добавляют незарегистрированный инструмент, меняют инструкции или внешний API и повторяют ручную замену при отказе сборщика. Контрольный состав должен свериться, а неизвестная или устаревшая версия — обнаружиться и блокировать выпуск либо получить статус «состав неизвестен» до исправления.

Подтверждения

Подтверждения включают актуальный проверяемый состав, записи происхождения и владельцев, связь обучения и сборки с артефактом, отчет сверки рабочей среды и результат запроса области воздействия; подпись и машинно-читаемый формат подтверждаются, если

они выбраны как часть реализации.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Для значимого выпуска отсутствие компонента из выбранного состава блокирует решение либо требует документированного исключения. В эксплуатации потеря актуальности эскалируется; до сверки затронутый сегмент считается имеющим неизвестный состав и работает только в режиме, определённом для этого состояния конфигурацией развёртывания.

Исправление и повторная проверка

Дополняют состав отсутствующим компонентом или исправляют связь развертывания с моделью, данными, кодом, инструментами и оценкой. Повторяют сбор состава для того же выпуска и запрос области инцидента: все фактические объекты разрешаются в точные версии без необъясненных расхождений; сохраняют манифест и отчет сверки.

Побочные эффекты

Подробное происхождение увеличивает объем чувствительных метаданных и требования к хранению, а автоматический сбор добавляет зависимость в конвейер выпуска.

Остаточный риск

Состав подтверждает прослеживаемость, но не безопасность, лицензионную допустимость или отсутствие скрытой зависимости. Даже полный и подписанный инвентарь может содержать неверные сведения или вредоносный компонент, не распознанный как опасный.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Container Registry

Частичная поддержка. Уникальная контрольная сумма позволяет фиксировать точную версию контейнерного образа с кодом, средой выполнения и библиотеками.

Ограничение: Контрольная сумма покрывает только контейнерный образ и не описывает модель, данные, параметры обучения, RAG или фактический состав развёртывания. Сама контрольная сумма не подтверждает происхождение или использование версии в продуктивной среде.

Yandex MetaData Hub

Частичная поддержка. В Data Catalog представлены происхождение данных и связи между таблицами и представлениями.

Ограничение: Data Catalog находится на стадии предварительного доступа и не формирует полный состав AI/ML-системы: он не связывает версии модели, кода, гиперпараметров и среды с фактическим развёртыванием.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Фиксирует владельца, источник, назначение, класс, версию, места, потребителей и срок хранения набора данных.

Ограничение: Не инвентаризирует каждую регулируемую копию вне наборов данных.

Условие применения: Для наборов регулируемых данных, используемых в AI-сценарии.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Фиксирует владельца, источник, назначение, класс, версию, места хранения, потребителей, срок хранения и связь с выпуском.

Ограничение: Не все преобразования, ограничения использования и получатели заданы одной записью.

Условие применения: До загрузки или использования набора данных, RAG-корпуса либо набора оценки и при изменении его версии.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Фиксирует долю синтетических данных, генератор и его версию, правила отбора и несинтетическую контрольную выборку.

Ограничение: Не фиксирует все исходные данные генерации, параметры генерации и метку записи или пакета.

Условие применения: Если синтетические данные используются для обучения, дообучения или оценки.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-MODEL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает рабочий выпуск с точными идентификаторами и версиями модели, данных, кода, зависимостей, инструментов и конфигурации.

Ограничение: Манифест состава не устанавливает лицензионные условия, безопасный импорт и хранение принятой внутренней копии модели.

Условие применения: Для самостоятельно размещённой или сторонней модели, а также для доступных идентификаторов управляемого модельного API.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Формирует неизменяемый манифест модели, данных, кода, зависимостей, инструментов и конфигурации и сверяет его с фактическим рабочим составом.

Ограничение: Стандарт требует явно учесть также адаптеры, токенизатор, промпты, Guardrails, экземпляр модели, MCP и внешнего поставщика, если они используются. В карточке такой полный перечень не является обязательным.

Условие применения: Для каждого выпуска ИИ-системы с программными зависимостями, моделью, данными или внешними интеграциями.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

По идентификатору компонента определяет точные затронутые выпуски и развёртывания.

Ограничение: Инвентарь определяет затронутые версии, но сам не получает сигнал поставщика и не запускает ограничение, замену и повторную оценку.

Условие применения: Когда нужно сопоставить сигнал о компоненте с точными выпусками и рабочими развёртываниями ИИ-системы.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке

Редакция: Методические рекомендации Банка России № 3-МР от 16.06.2026

п. 5.4 (сведения о ПО и компонентах) · Поддерживает часть требования

Проверяемый состав модели, данных, кода и зависимостей даёт часть сведений для оценки доверия к поставщику AI-компонента.

Граница связи: Инвентарь не создаёт методику оценки доверия и не подтверждает аудит, сертификацию или защищённость инфраструктуры поставщика.

Субъект	Участник финансового рынка из п. 1.1 рекомендаций
Условие применения	Участник финансового рынка из пункта 1.1 оценивает доверие к данным или моделям поставщика либо к компонентам технологий ИИ с открытым кодом.
Условие связи с карточкой	Инвентарь содержит сведения о ПО, агентах, плагинах, интерфейсах и происхождении данных, относящихся к оцениваемому компоненту.
Сроки и переходные условия	Методические рекомендации № 3-МР носят рекомендательный характер.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	586	171
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1003	44
AISAFE2-COMP-008 · Данные для подсистемы обучения	1004	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1005	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1006	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1007	67
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	1008	84
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	1233, 1310	364
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1237	368
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1311	415
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1312	419
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	1313	426
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1314	509
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1315	514

Мера защиты · AISAFE2-MIT-036

Обнаружение и обработка дрейфа, смещения данных и деградации

Краткое описание

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации.

Механизмы

- Обнаружение
- Жизненный цикл

Защитный механизм

Мониторинг или иной регулярный контроль сравнивает текущие входы, выходы и доступные метки с базовым профилем, зафиксированным для точной версии, и формирует количественный либо качественный сигнал. Разбор отличает нормальное изменение, ошибку данных, деградацию модели и возможную атаку; затем выбираются ограничение, возврат версии, переобучение, изменение порога или документированное исключение.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Целесообразность определяется изменчивостью данных и поведения, степенью контроля над моделью, последствиями деградации и доступностью достоверных сигналов. Организация выбирает метрики, качественные признаки, группы, окна времени и режимы; при слабой измеримости явно учитывается неопределенность. В стабильном сценарии с низким риском организация может принять остаточный риск и не внедрять постоянный мониторинг. Версия и исходный базовый профиль фиксируются отдельным подготовительным контролем. Мера не применяется к детерминированному компоненту без модели, обучающих или входных распределений и без эксплуатационного критерия качества, способного измениться во времени.

Предпосылки

- Обязательны воспроизводимый базовый профиль точной версии и конфигурации, утвержденные метрики, окна наблюдения, значимые группы и допустимая задержка меток.

Реализация

Для каждой точной версии фиксируют базовый профиль, метрики, окна, значимые группы, допустимую задержку меток и поведение при потере источника. Порог связывают с конкретным действием: наблюдать, ограничить, вернуть версию, исправить данные или переобучить. Изменение профиля, метрики, порога или маршрута требует нового решения и повторного теста; значимый сигнал закрывается только с зафиксированной причиной, действием и подтверждением результата.

Проверка

Все выбранные метрики и качественные признаки должны иметь критерий значимости и действующий маршрут; контролируемое отклонение, если его можно достоверно воспроизвести, должно обнаруживаться и приводить к решению в пределах окна реакции, зафиксированного в определении сигнала. Для количественных сигналов измеряются ложные предупреждения, пропуски и задержка; для качественных фиксируются методика, выборка и неопределенность. Мера неэффективна при молчаливом нарушении выбранного критерия или закрытии значимого сигнала без решения.

Негативный тест

Подают стабильный контрольный поток и поток с заранее измеренным постепенным, резким или групповым дрейфом, затем задерживают метки и отключают обязательный источник. Стабильный поток должен остаться ниже порога из определения сигнала, дрейф — сформировать решение в пределах зафиксированного там окна реакции, а потеря источника — обнаружиться и перевести оценку в указанный там режим неопределённости, а не считаться отсутствием дрейфа.

Подтверждения

Подтверждения включают определения сигналов, базовый профиль и критерии, назначенных владельцев, результаты контролируемых проверок, журналы разбора и корректирующих действий и доступные показатели качества обнаружения.

Поведение при отказе

Безопасная деградация

Сценарий отказа

При наличии используется независимый резервный сигнал; для значимых решений включается ограниченный режим, ручная проверка или последняя проверенная версия. Продолжение без видимости требует временного решения владельца риска и усиленного контроля.

Исправление и повторная проверка

Пересчитывают базу, исправляют признак или источник данных и восстанавливают маршрут решения для значимого сигнала дрейфа. На том же временном окне повторяют контрольный сдвиг и стабильный поток: сдвиг обнаруживается и приводит к реакции из определения сигнала, стабильный поток остаётся в зафиксированном там допуске; сохраняют версию критерия и сравнительный отчет.

Побочные эффекты

Чувствительные пороги создают шум и операционную усталость, а частые корректирующие действия могут снизить доступность или стабильность модели.

Остаточный риск

Плохой базовый набор показателей и редкие события могут скрыть деградацию. Остаточный сценарий — опасное поведение возникает в малой подгруппе и не меняет агрегированные метрики до значительного воздействия.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Monitoring

Частичная поддержка. После вычисления приложением метрики дрейфа Monitoring может сравнивать ее с порогами, хранить историю статусов и отправлять уведомления.

Ограничение: Monitoring сравнивает переданную метрику с порогом; расчёт дрейфа, выбор реакции и проверка качества метрики остаются за приложением.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	504	100
AISAFE2-WEAK-028 · Недостатки наблюдаемости и обнаружения злоупотреблений	623	199
AISAFE2-COMP-002 · Модель и интерфейс инференса	1009	37
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	1010	48
AISAFE2-COMP-008 · Данные для подсистемы обучения	1011	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1012	54
AISAFE2-COMP-013 · Телеметрия и обнаружение	1013	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1014	69
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1015	86

Карточка	Связи	Стр.
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1238, 1318	368
AISAFE2-MIT-009 · Версионирование и откат базы знаний	1316	346
AISAFE2-MIT-013 · Проверка компонентов цепочки поставки перед выпуском	1317	364
AISAFE2-MIT-021 · Управление доверием и сроком жизни долговременной памяти	1319	396
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1320	415
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	1321	426
AISAFE2-MIT-033 · Шлюз выпуска изменений RAG, векторных хранилищ и памяти	1322	444
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1323, 1337	466
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1324	473
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1325	494

Мера защиты · AISAFE2-MIT-037

Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений

Краткое описание

Шлюз распределяет запросы на подтверждение по классу риска, контролирует нагрузку проверяющих и качество решений и при перегрузке блокирует действие либо направляет его по заранее заданному маршруту эскалации.

Механизмы

- Человеческий контроль

Защитный механизм

Шлюз связывает класс действия с ролью проверяющего, обязательным объёмом предпросмотра, числом подтверждающих, сроком решения и маршрутом эскалации. Для низкорисковых однородных операций допускаются выборка или групповое подтверждение, если показанный состав и пределы не скрывают высокорисковый эффект; высокорисковые действия проходят индивидуальный маршрут.

Формы контроля

- Техническая
- Организационная
- Выполняемая человеком

Ответственность

- Заказчик

Применимость

Целесообразность шаблона определяется числом и частотой человеческих решений, последствиями ошибки, доступностью проверяющих и риском усталости. Организация масштабирует индивидуальную проверку, выборку и групповые операции по классу риска. Шлюз управляет процессом проверки, но не является самостоятельным техническим контролем действия и не заменяет подтверждение конкретного высокорискового действия. Если процесс не содержит повторяющихся человеческих подтверждений, очереди или нескольких проверяющих, мера не применяется; детерминированно запрещённые операции блокируются без этого шлюза.

Предпосылки

- Обязательна действующая техническая точка подтверждения с неизменяемым предпросмотром, достоверной идентичностью проверяющего и блокированием исполнения до решения.

Реализация

Система измеряет время, отмены, повторные решения, долю одобрений и результаты выборки. Сам шаблон не подтверждает действие и зависит от технического шлюза «Подтверждение человеком высокорисковых действий до исполнения».

Проверка

Каждое действие или группа операций из выбранного класса должны проходить по маршруту, зафиксированному для этого класса в матрице классов и ролей, и обрабатываться указанной там ролью. До эксплуатации в конфигурации очереди фиксируются пределы нагрузки и времени, признаки формального одобрения и объём независимой выборки для каждого класса риска. Мера неэффективна при выполнении действия после просроченного решения или без просмотра требуемого контекста.

Негативный тест

Подают обычное контрольное действие и высокорисковое действие при пиковой очереди, отсутствии проверяющего, неполном предпросмотре, массовом одобрении и истекшем решении, затем повторяют обход через другой инструмент. Контрольное действие должно пройти по назначенному маршруту, а высокорисковое — блокироваться или эскалироваться до внешнего эффекта; неизмененное действие исполняется только после отдельного действующего решения.

Подтверждения

Подтверждения включают матрицу классов и ролей, конфигурацию очереди и замещения, записи решений, метрики нагрузки, времени и доли одобрений, результаты стрессового теста и материалы независимой выборки.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Высокорисковое действие блокируется или направляется замещающему ответственному. Для иных классов может использоваться заранее определенный ограниченный либо ручной маршрут. Упрощение контекста или автоматическое одобрение, которое уничтожает существенное свойство выбранного шлюза, недопустимы.

Исправление и повторная проверка

Закрывают исполнение после просроченного или формального подтверждения, исправляют представление обязательного контекста и перераспределяют очередь. Повторяют исходное действие без просмотра контекста и с просроченным решением: оба блокируются; после новой независимой проверки исполняется только неизмененное действие, а запись решения сохраняется.

Побочные эффекты

Дополнительные роли замедляют работу и создают дефицит специалистов; жесткие пределы очереди могут снижать доступность процесса.

Остаточный риск

Проверяющий может ошибиться или быть введен в заблуждение даже при приемлемой нагрузке. Остаточный сценарий — скоординированное либо ошибочное подтверждение вредного действия, формально соответствующее процессу.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI09** · Поддерживает часть требования
>
> Маршрут подтверждения по риску, предпросмотр действия и контроль нагрузки проверяющих снижают вероятность формального одобрения опасного шага.
>
> **Граница связи:** Даже при такой проверке человека могут убедить ложным обоснованием или вводящим в заблуждение интерфейсом; действия вне маршрута остаются вне меры.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Человек принимает решение на основе представления, сформированного агентом.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-023 · Перегрузка контура человеческой проверки	555	129
AISAFE2-WEAK-065 · Избыточная маршрутизация задач на человеческую проверку	674	265
AISAFE2-WEAK-066 · Недостаточная мощность и резервирование человеческой проверки	675	266
AISAFE2-WEAK-067 · Недостаточная инструментальная поддержка оператора HITL	676	267
AISAFE2-WEAK-068 · Недостаточный контроль качества решений и утомления операторов HITL	677	268
AISAFE2-WEAK-069 · Неопределённый режим переполнения и недоступности человеческой проверки	678	269

Карточка	Связи	Стр.
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	1016	65
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1258, 1326	389
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1300	429
AISAFE2-MIT-045 · Независимая проверка решений с существенными последствиями	1327, 1353	488

Мера защиты · AISAFE2-MIT-038

Контроль злоупотребления AI-сервисом

Краткое описание

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Механизмы

- Валидация
- Контроль ресурсов
- Обнаружение

Защитный механизм

Коррелятор объединяет запросы по субъекту, ключу, арендатору и сессии через эквивалентные интерфейсы, сопоставляет последовательность с версионируемой политикой злоупотреблений и передает единое решение всем точкам применения ограничения. Операционный процесс разбирает ложные срабатывания, санкции и обжалования.

Формы контроля

- Техническая
- Процессная
- Организационная

Ответственность

- Заказчик

Применимость

Целесообразность определяется доступностью интерфейса недоверенным субъектам, возможным масштабом вреда, наблюдаемыми сигналами, установленной политикой и правовыми основаниями обработки. Она не заменяет защиту самой системы от внедрения вредоносных инструкций, авторизацию действий, юридическую квалификацию содержимого или расследование вреда вне наблюдаемой области. Для закрытого контура без недоверенных пользователей мера может не применяться, если оценка риска не выявила сопоставимого внутреннего злоупотребления.

Предпосылки

- Нужны политика недопустимого использования, устойчивая корреляция субъекта между эквивалентными интерфейсами, наблюдаемые признаки кампании и разрешенные меры воздействия.
- Пороги и маршрут пересмотра спорной блокировки задаются до применения санкции.

Реализация

Владелец сервиса определяет сценарии недопустимого использования и сигналы на уровне субъекта, ключа, арендатора, сессии, запроса и результата. Контроль сочетает лимиты частоты и стоимости, корреляцию поведения и признаки кампаний. Выбранная мера воздействия — предупреждение, задержка, ограничение возможностей, ручная проверка, приостановка или отзыв доступа — применяется ко всем эквивалентным интерфейсам. При наличии данных корреляция охватывает распределённые и малоскоростные кампании и не опирается только на текст одного запроса. Сигнал модерации может быть входом корреляции; правила категорий содержимого и реакция на отказ классификатора задаются мерой «Исполнение версионируемой политики допустимости содержимого».

Проверка

На зафиксированных сценариях проверяют легитимный трафик и кампанию злоупотребления через несколько учётных данных, сессий и эквивалентных интерфейсов, включая низкую частоту и потерю одного обязательного источника корреляции. Измеряют задержку решения и ложные блокировки, проверяют применение санкции на каждом охваченном пути. Мера неэффективна, если кампания продолжается через эквивалентный интерфейс без нового решения, потеря обязательного сигнала незаметна либо санкция не действует на всех охваченных путях.

Негативный тест

Воспроизводят легитимный контрольный трафик и одну злоупотребляющую кампанию, разделенную между учетными данными, интерфейсами и низкочастотными вариантами формулировок, затем отключают одну зависимость корреляции. Контрольный трафик должен остаться разрешенным, а кампания — объединиться в одно решение и получить согласованное ограничение; потеря зависимости должна быть наблюдаемой и не открывать необработанный обход.

Подтверждения

Подтверждениями служат версия политики допустимого использования, правила корреляции субъектов и кампаний, результаты сценарных тестов, журналы санкций, метрики ложных блокировок и материалы пересмотра спорных решений.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Для функций, способных массово создавать внешний вред, включаются сниженные лимиты, ограниченный набор возможностей или ручная проверка. Автоматическое разрешение ранее заблокированной кампании при недоступности детектора запрещается.

Исправление и повторная проверка

Исправляют корреляцию субъекта и эквивалентных интерфейсов либо применение санкции, через которые исходная кампания продолжилась. Повторяют ту же распределенную последовательность с другого пути и нормальную нагрузку: кампания обнаруживается и ограничивается как единое целое, нормальный субъект сохраняет доступ; сохраняют сигнал, решение и примененные ограничения.

Побочные эффекты

Жесткие правила способны ограничить законное использование, требуют процедуры пересмотра и создают дополнительную операционную нагрузку.

Остаточный риск

Содержание и намерение могут оставаться неоднозначными, а злоумышленник может распределять действия между сервисами и субъектами. После внедрения сохраняется риск вредоносного использования допустимых возможностей, которое не образует наблюдаемого шаблона.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Smart Web Security

Частичная поддержка. Yandex Smart Web Security защищает инфраструктуру от DDoS-атак и ботов на уровне приложений L7.

Ограничение: SWS ограничивает ботов и DDoS на входе L7; распределённые низкоскоростные кампании, смысл использования AI и стоимость операций требуют других данных и правил.

Yandex SmartCaptcha

Частичная поддержка. SmartCaptcha выявляет похожие на автоматические запросы и отсеивает ботов, уменьшая автоматизированную часть злоупотребления публичным AI-интерфейсом.

Ограничение: Сервис не определяет категории недопустимого AI-использования, не коррелирует субъектов и распределённые кампании и не управляет стоимостью или отзывом доступа.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM06:2026 · Поддерживает часть требования

Сопоставление массовых запросов по субъекту и маршрутам помогает ограничить распределённое злоупотребление вычислениями AI-сервиса.

Граница связи: Ненаблюдаемая или формально допустимая нагрузка может пройти политику; без жёсткого бюджета расходы одной операции не ограничены.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вызовы модели или агента могут расходовать ресурсы, время или бюджет без установленной границы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-003 · Вовлечение пользователей в злонамеренные схемы	495	95
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	617	194
AISAFE2-WEAK-027 · Отсутствие экономического ограничения массовых запросов	619	196
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1017	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	1018	37
AISAFE2-COMP-013 · Телеметрия и обнаружение	1019	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1020	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1021	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1022	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1023	79
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1275, 1332	405
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	1328	316
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	1329	333
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	1330	349
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	1331	356
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	1333	441
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1334, 1348	485

Мера защиты · AISAFE2-MIT-039

Проверка пригодности модели для целевого сценария

Краткое описание

Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения.

Механизмы

- Валидация
- Шлюз допуска

Защитный механизм

Владелец сценария до оценки кандидата задает необходимые свойства, запрещенные режимы, критерии и условия отказа. Шлюз допуска проверяет модель на репрезентативных, граничных и стрессовых наборах в объёме, зафиксированном в методике для класса значимости изменения, включая группы и условия с различными последствиями ошибки.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Целесообразность и глубина независимой проверки определяются новизной сценария, существенностью последствий, прозрачностью модели и доступностью надежных данных. Для высокорискового изменения предпочтителен независимый воспроизводимый повтор; для ограниченного низкорискового пилота может быть достаточна более узкая проверка с явными границами. Мера не заменяет безопасную обработку результата, фиксацию версии и базового профиля и эксплуатационный контроль дрейфа.

Предпосылки

- Нужны точная версия модели и конфигурации, описание целевого сценария, репрезентативный набор и утвержденные пороги для обязательных метрик и значимых подгрупп.
- Провальная обязательная проверка должна блокировать применение модели в этом сценарии.

Реализация

Результат связывается с точной версией, конфигурацией, системными инструкциями и ограничениями применения. Допуск фиксирует разрешенные сценарии и случаи, требующие другой модели, ручного процесса или отказа от AI.

Проверка

На точной версии и конфигурации выполняют заранее зарегистрированную методику на репрезентативном, граничном и стрессовом наборах, сравнивают все обязательные и групповые метрики с порогами, зафиксированными в этой методике, и при необходимости проводят независимый повтор. Мера неэффективна, если модель допущена без заранее определенных критериев либо результаты нельзя связать с фактической версией.

Негативный тест

На точной версии и конфигурации запускают заранее зарегистрированный набор типичных, редких, неполных и OOD-входов, включая ошибочно уверенный случай и значимые подгруппы, а затем повторяют при недоступности модели. Кандидат допускается только при выполнении всех обязательных порогов, зафиксированных в зарегистрированной методике, и корректном безопасном маршруте отказа; заведомо пригодный контрольный кандидат должен пройти в заявленной области.

Подтверждения

Подтверждения включают описание назначения и запрещенных режимов, заранее утвержденную методику и наборы данных, точные версии модели и конфигурации, отчеты по метрикам и группам, результаты воспроизведения в выбранном объеме и запись решения с ограничениями применения.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Выбор или расширение сценария блокируется. Допускается ограниченный пилот без внешних последствий с отдельной областью, усиленной проверкой и сроком завершения оценки.

Исправление и повторная проверка

Исправляют набор сценариев, метрику или порог, которые скрыли непригодность модели, либо заменяют модель и конфигурацию. Повторяют исходный провальный и контрольный сценарии на точной версии: все обязательные и групповые метрики достигают порогов, зафиксированных в методике; сохраняют набор, конфигурацию и отчет.

Побочные эффекты

Полная оценка увеличивает срок и стоимость выбора, а слишком узкие критерии могут отклонить пригодную модель или ограничить полезный сценарий.

Остаточный риск

Оценочные данные могут не представлять будущие условия, а поставщик может изменить непрозрачную модель. Успешная оценка не гарантирует корректность каждого ответа. После допуска сохраняется риск неизвестного режима поведения и изменения распределения данных.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Задаёт версии, метрики, пороги, проверку на репрезентативных, граничных и стрессовых наборах до допуска.

Ограничение: Карточка не требует независимого проверяющего для каждого случая и не задаёт проверки разбиения, актуальности и метаданных RAG-корпуса.

Условие применения: До выбора модели или расширения её сценария применения, в том числе при работе через управляемый API без собственного набора данных.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Приказ ФСТЭК России от 11.04.2025 № 117 Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений

Редакция: Приказ № 117 от 11.04.2025 с изменениями по приказу № 137 от 08.05.2026

п. 61(в) · Поддерживает часть требования

Оценка модели на репрезентативных сценариях может дать статистические критерии для выявления недостоверных ответов.

Граница связи: Предварительный тест не собирает и не анализирует такие ответы в работе; критерии и пороги оператор должен утвердить для точной версии и задачи.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	В охваченной приказом № 117 системе AI-сервис выдаёт ответы пользователям при выполнении ими своих обязанностей.
Условие связи с карточкой	Результаты оценки точной модели используются для разработки статистических критериев выявления недостоверных ответов.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 61 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM07:2026 · Поддерживает часть требования

Испытания модели на целевых и граничных сценариях позволяют выявить часть ложных или неподтверждённых ответов до допуска версии.

Граница связи: Конечная выборка не подтверждает каждое утверждение в работе и не защищает от устаревшего контекста или чрезмерного доверия к ответу.

Субъект	Аналитик угроз AI-приложения
Условие применения	Вывод модели используется как фактическое основание ответа, рекомендации или решения.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	492	93
AISAFE2-WEAK-015 · Переобучение модели	593	175
AISAFE2-WEAK-017 · Отсутствие проверяемых критериев выбора модели	597	179
AISAFE2-WEAK-042 · Недостаточная пригодность модели для целевого сценария	636	217
AISAFE2-COMP-002 · Модель и интерфейс инференса	1024	37
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1025	67
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1026	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1027	79
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1239, 1335	368
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1323, 1337	456
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	1336	371
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1338	469
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1339	473
AISAFE2-MIT-047 · Устойчивость к состязательным входным данным	1340	494
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	1341	507

Мера защиты · AISAFE2-MIT-040

Допуск и целостность обучающих данных, разметки и признаков

Краткое описание

До обучения проверяются происхождение, права, целостность, разметка и допустимость каждого набора данных и признака.

Механизмы

- Целостность

Защитный механизм

Шлюз допуска связывает набор с происхождением, составом, правами использования, преобразованиями и разметкой в объеме, необходимом для выбранного уровня доверия. Для недоверенных или доступных атакующему источников применяются карантин, контрольные суммы и углубленные проверки аномалий, неожиданных дубликатов, сдвига распределения, целевых изменений разметки и признаков скрытой закладки.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

При собственном обучении или дообучении мера применяется к используемым обучающим, валидационным и тестовым наборам, разметке, признакам и преобразованиям. Она не заменяет отдельные контроли допуска источников RAG, защиты конвейера обучения и проверки приватности модели.

Предпосылки

- Обязательные предпосылки выбранного контроля — проверяемая запись происхождения, однозначно идентифицированная версия набора и контролируемый путь передачи в обучение.
- Для синтетической части нужны идентификаторы исходных данных, метода или модели, версии и параметров генератора, назначения, доли и известных ограничений.

Реализация

Конкретная версия данных и преобразований связывается с запуском обучения и выпущенным артефактом. До выпуска подтверждается независимость обучающей, валидационной и тестовой выборок: проверяются пересечения, утечки признаков и повторное использование объектов после преобразований. Для синтетической части фиксируются метод или модель, версия и параметры генератора, назначение, исходные данные, доля и известные ограничения. Каждая запись либо атомарный пакет имеет признак `synthetic=true`, отдельное происхождение, идентификатор и хеш; необработанный сгенерированный материал не смешивается с наблюдаемыми данными и не допускается без явного решения. Синтетические данные проходят тот же шлюз и отдельно сравниваются с несинтетической контрольной выборкой.

Проверка

Для контрольного несинтетического набора проверяют происхождение, разрешённое использование, целостность версии, подмену разметки и прямой обход шлюза. Синтетическую версию дополнительно прослеживают до точных генератора, параметров и версии исходных данных, сверяют долю и ограничения и сравнивают с несинтетической контрольной выборкой. Пакет без `synthetic=true`, с изменёнными параметрами, неизвестным исходным набором, смешанным происхождением или без отдельного решения о допуске должен блокироваться до обучения или оценки.

Негативный тест

Пытаются загрузить синтетический пакет без `synthetic=true`, с иными параметрами генератора и напрямую из необработанного результата генерации: все варианты блокируются до обучения или оценки. Отдельно передают несинтетический набор с изменённой после допуска версией или подменённой разметкой и пытаются обойти шлюз; варианты, нарушающие утверждённую политику, блокируются.

Подтверждения

Сохраняются запись происхождения и прав, метод или модель, версия и параметры генератора, версия исходных данных, назначение, доля, ограничения, метка `synthetic=true`, идентификатор и хеш пакета, решение шлюза и отдельное сравнение качества.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Для высокорискового или недоверенного набора запуск обучения либо выпуска блокируется, а данные изолируются от рабочего пути. В иных сценариях возможны отдельная проверка или возврат к последней подтвержденной версии, совместимой с целью обучения.

Исправление и повторная проверка

Удаляют неразрешенный объект, восстанавливают запись допуска и точные версии преобразований либо блокируют обучение в обход шлюза данных. Повторяют запуск с исходным набором и измененной версией признаков: оба блокируются до нового допуска; разрешенный запуск воспроизводит манифест, который сохраняют с решением.

Побочные эффекты

Строгий допуск замедляет обновление данных, увеличивает стоимость хранения снимков и может задержать исправление смещения набора.

Остаточный риск

Разрешенный источник может быть содержательно ошибочным или уже отравленным, а сложная скрытая закладка может не проявиться в тестах. После проверки сохраняется риск невыявленного статистически правдоподобного загрязнения.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex MetaData Hub

Частичная поддержка. Профилирование Data Catalog предоставляет статистику по таблицам и колонкам.

Ограничение: Сервис находится на стадии предварительного доступа, не помещает набор в карантин, не проверяет права и разметку, не обнаруживает все виды отравления данных и не блокирует запуск обучения.

Yandex Object Storage

Частичная поддержка. Версионирование и Object Lock позволяют сохранить неизменяемую версию набора данных и восстановить ее после нежелательного изменения или удаления.

Ограничение: Механизм не проверяет происхождение, права использования, разметку, преобразования, аномалии или отравление данных. Включённый Object Lock не подтверждает, что обучение использовало именно заблокированную версию.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA-CHECK1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет происхождение, целостность, разметку, аномалии, неожиданные дубликаты, сдвиг распределения и признаки отравления или скрытой закладки.

Ограничение: Карточка не задаёт весь набор проверок AI-DATA-CHECK1: парсинг, форматы полей, скрытые символы, исполняемые объекты, внедрённые инструкции и чувствительные данные требуют дополнительных проверок; RAG-корпус оценивают отдельно.

Условие применения: До обучения, дообучения, валидации или тестирования на недоверенном наборе данных.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает обучающие данные с происхождением, преобразованиями, правами использования и записью допуска.

Ограничение: Связь ограничена наборами для обучения, валидации и тестирования; реестр RAG-корпусов и прочих эксплуатационных данных она не охватывает.

Условие применения: При собственном обучении или дообучении — для обучающих, валидационных и тестовых наборов.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет независимость обучающей, валидационной и тестовой выборок и утечки после преобразований.

Ограничение: Не покрывает всю оценку качества модели и RAG.

Условие применения: При обучении или дообучении на собственных обучающих, валидационных и тестовых выборках.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Допуск связывает синтетическую часть с генератором, его версией и параметрами, исходными данными, назначением, долей и ограничениями. Метка synthetic=true, идентификатор, хеш и отдельное решение сохраняют её происхождение; качество сравнивается с несинтетической выборкой.

Ограничение: Карточка описывает синтетические наборы собственного обучения и дообучения. Использование синтетики только для оценки внешней модели требует отдельно определить тот же порядок допуска.

Условие применения: Для синтетической части обучающих, валидационных и тестовых наборов при собственном обучении или дообучении.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Связывает обучающий набор с происхождением, правами использования и точной версией, проверяет целостность и не допускает недоверенные данные без решения.

Ограничение: Для сторонней поставки отдельно фиксируют контакт поставщика, дату получения, независимое подтверждение целостности при его наличии и изолированный импорт принятой внутренней копии.

Условие применения: При получении извне данных для собственного обучения, дообучения, валидации или тестирования.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	505	100
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	587	171
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	591	173
AISAFE2-COMP-008 · Данные для подсистемы обучения	1028	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1029	54
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1240	368
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1338	466
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1342	473
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1343	478
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных	1344	503
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	1345	541

Мера защиты · AISAFE2-MIT-041

Защита конвейера обучения и дообучения

Краткое описание

Конвейер обучения фиксирует входы, код, среду, полномочия и результат и отделяет обучение от выпуска рабочей модели.

Механизмы

- Изоляция
- Целостность

Защитный механизм

Для систем с повышенными требованиями задание запускается от отдельной машинной идентичности в изолированной среде с минимальными полномочиями и контролируемым сетевым доступом. В иных архитектурах допустим эквивалентный механизм, если он ограничивает изменение входов, полномочия и прямой выпуск результата.

Формы контроля

- Техническая
- Архитектурная
- Процессная

Ответственность

- Заказчик

Применимость

Мера обычно применима при собственном обучении или дообучении и охватывает подготовку задания, исполнение и передачу результата в управляемое хранилище или реестр. Она не подтверждает качество данных «Допуск и целостность обучающих данных, разметки и признаков», пригодность модели «Проверка пригодности модели для целевого сценария» или безопасную загрузку артефакта «Проверка целостности и безопасная загрузка артефактов модели».

Предпосылки

- Обязательны допущенные и однозначно идентифицированные входы и управляемая связь результата с конкретным запуском.
- Для усиленной реализации требуются отдельная машинная идентичность, изолированная среда, закрытые хранилища и подписанная аттестация.

Реализация

Код, среда, библиотеки, гиперпараметры, конфигурация распределенного обучения и входные данные фиксируются до запуска в выбранном объеме. Секреты не должны записываться в артефакт. Результат связывается с идентификаторами входов, среды и выполненных проверок; подпись или аттестация является усиленным способом такой связи. Контуры разработки, обучения и рабочего выполнения разделяются по идентичностям, хранилищам, секретам и исходящим соединениям в заявленной области. Экспериментальный процесс не должен записывать результат прямо в рабочее хранилище или менять рабочую конфигурацию в обход утвержденного продвижения.

Проверка

Для выпущенного артефакта восстанавливают проверяемую связь с входами, кодом, параметрами, зависимостями, средой и идентичностью запуска; подписанную аттестацию проверяют, если она выбрана. Затем пытаются изменить каждый элемент после фиксации и обратиться к рабочим данным или выпуску через прямой путь. Мера неэффективна, если артефакт нельзя однозначно связать с проверенными входами и средой либо доступ возможен в обход выбранных границ. Для 100% выпущенных артефактов должен быть доступен манифест с идентификаторами входов, кода, среды и запуска.

Негативный тест

Запускают контрольное обучение из зафиксированных входов, затем после фиксации меняют код, данные, библиотеку или гиперпараметр и пытаются выпустить результат через прямой маршрут с избыточным доступом либо при частичном отказе. Контрольный запуск должен завершиться с проверяемой связью результата с точными входами и средой, а изменённый или непроверяемый — не должен выпускать артефакт или читать лишние данные и секреты.

Подтверждения

Подтверждения включают идентичность и права задания, версии кода, среды, библиотек и конфигурации, применимые политики сети и хранилищ, журнал запуска, связь результата с входами и результаты тестов; подпись, аттестация и воспроизводимый повтор включаются, если они выбраны как усиленная реализация.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Задание останавливается либо переводится в безопасное состояние, а результат не получает требуемого подтверждения и не допускается в управляемое хранилище. Повторный запуск выполняется после устранения причины в среде, соответствующей выбранным гарантиям; частичный результат нельзя выдавать за полностью проверенный.

Исправление и повторная проверка

Закрывают обход границы обучения, исправляют изоляцию среды или связь артефакта с проверенными входами и кодом. Повторяют исходный неразрешённый доступ и запуск обучения: доступ к рабочим данным отклоняется, результат выпускается только из зафиксированной среды и однозначно связывается с манифестом запуска; сохраняют манифест и, если она выбрана, аттестацию запуска и результата.

Побочные эффекты

Изоляция и воспроизводимость увеличивают стоимость и длительность обучения, а жесткое ограничение сети может затруднить диагностику и загрузку зависимостей.

Остаточный риск

Авторизованный участник, скомпрометированная управляющая плоскость или аппаратная среда могут повлиять на обучение. После проверки сохраняется риск вредоносного, но формально разрешенного изменения.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex DataSphere

Частичная поддержка. В настройке проекта DataSphere можно указать группу безопасности, если группы безопасности используются в организации.

Ограничение: Эти настройки не фиксируют код, данные, библиотеки и гиперпараметры, не доказывают минимальность прав и не создают контролируемый шлюз выпуска результата.

Yandex Identity and Access Management

Частичная поддержка. Сервисные аккаунты позволяют гибко настраивать права доступа программ к ресурсам.

Ограничение: IAM задаёт доступ к облачным ресурсам; сеть, состав задания и запрет прямого выпуска результата требуют отдельных контролей.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Разделяет идентичность, входы, код, среду и выход обучающего конвейера и не допускает прямую запись в продуктивную среду.

Ограничение: Не задаёт общую машину состояний для всех стадий данных.

Условие применения: При собственном обучении или дообучении, когда данные и результаты переходят из подготовки в рабочий выпуск.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-TRAIN1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Разделяет идентичность, вход, код, среду и выход обучения и запрещает прямую запись результата в продуктивную среду.

Ограничение: Карточка охватывает конвейер обучения и дообучения; отдельные экспериментальные запуски, точный путь доставки секретов и исключения для выхода в интернет проверяются дополнительно.

Условие применения: При собственном обучении или дообучении модели в DataSphere, Kubernetes, Compute Cloud либо другой среде Yandex Cloud.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM04:2026 · Поддерживает часть требования

Фиксация входов, кода, среды и результата обучения с отдельным выпуском затрудняет незаметную подмену модели через конвейер сборки.

Граница связи: Мера охватывает свой конвейер, но не проверяет сторонний датасет, качество весов и безопасную загрузку артефакта.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

LLM05:2026 · Поддерживает часть требования

Изоляция задания обучения и ограничение прав на входы и результат затрудняют несанкционированное изменение модели в конвейере.

Граница связи: Формально разрешённый, но загрязнённый набор может пройти обучение; мера не подтверждает качество данных или отсутствие закладки.

Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	595	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	605	184
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели	667	253
AISAFE2-COMP-008 · Данные для подсистемы обучения	1030	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1031	54
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1032	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1033	61
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	1230	359
AISAFE2-MIT-014 · Фиксация версии и базового профиля поведения модели	1241	368
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1324	456

Карточка	Связи	Стр.
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1339	466
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1342	469
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1374	525
AISAFE2-MIT-060 · Контролируемое продвижение данных между стадиями	1381	541

Мера защиты · AISAFE2-MIT-042

Контроль приватности обучающих данных и утечек через модель

Краткое описание

Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Механизмы

- Минимизация данных
- Валидация

Защитный механизм

До обучения данные минимизируются, очищаются от ненужных идентификаторов и дубликатов, а доступ и срок хранения ограничиваются. По результатам анализа угроз применяются подходящие методы, например ограничение вклада отдельных записей, регуляризация или дифференциальная приватность.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Третья сторона

Применимость

Мера обычно применима при обучении на чувствительных или персональных данных либо при использовании внешней модели с обоснованным риском запоминания обучающих данных, определения участия записи в обучении, восстановления данных по поведению модели или определения свойств набора данных. Она не заменяет общую минимизацию перед передачей, допуск обучающих данных, защиту артефакта от несанкционированного получения и правовую оценку обработки.

Предпосылки

- Обязательные предпосылки — известный состав чувствительных данных, модель угроз приватности и точная версия оцениваемой модели и интерфейса.

Реализация

Готовая модель проверяется на извлечение или реконструкцию запомненных фрагментов, определение участия конкретной записи в обучении, восстановление данных или признаков отдельных объектов по ответам модели, вывод агрегированных свойств обучающего набора и иные релевантные атаки на приватность. Из интерфейса убираются ненужные вероятности, необработанные оценки модели и диагностические сигналы.

Проверка

Для атак, выбранных по модели угроз, сравниваются целевая и контрольная модели, применимые классы данных, редкие и повторяющиеся записи, доступные уровни выхода и возможности адаптивного противника. Проверяются также падение полезности и корректность заявленного бюджета приватности, если он используется. Мера неэффективна, если утверждение основано только на удалении явных идентификаторов либо оценка не учитывает базовую частоту и не показывает преимущество атаки относительно случайного угадывания. Оценка точной модели должна соответствовать ограничениям преимущества выбранных атак относительно базовой частоты и диапазону полезности, зафиксированным в методике.

Негативный тест

Для точной версии модели выполняют выбранные атаки определения участия записи в обучении, извлечения или восстановления обучающих данных с известной базовой частотой и отдельно проверяют падение полезности; преимущество атаки и качество сравнивают с ограничениями преимущества атаки и диапазоном полезности, зафиксированными в методике.

Подтверждения

Подтверждения включают состав и минимизацию данных, модель угроз, параметры защитного метода и, если используется дифференциальная приватность, её бюджет, точную конфигурацию интерфейса, отчеты выбранных атак, показатели полезности и результат проверки методики.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Обучение на затронутых данных или выпуск модели блокируются. Допускается повторное обучение на минимизированном наборе, ограничение интерфейса либо использование модели без этих данных.

Исправление и повторная проверка

Устраняют избыточные обучающие записи или усиливают метод защиты и исправляют оценку, не учитывавшую базовую частоту атаки. Повторяют ту же атаку определения участия записи в обучении, извлечения или восстановления обучающих данных на точной версии и измеряют обычную полезность: преимущество атаки и качество укладываются в ограничения преимущества атаки и диапазон полезности, зафиксированные в методике; сохраняют методику и отчёт.

Побочные эффекты

Усиление приватности может снижать качество, особенно для редких групп, и увеличивать стоимость обучения и оценки.

Остаточный риск

Известные тесты не охватывают все будущие атаки. Редкая запись может оставаться уязвимой при приемлемых агрегированных метриках. После выпуска сохраняется риск новой техники восстановления или утечки через объединение ответов.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Security Deck

Частичная поддержка. Модуль DSPM обнаруживает чувствительные данные в поддерживаемых хранилищах и может служить входом для последующей минимизации или обезличивания обучающих данных.

Ограничение: DSPM не выполняет минимизацию, дедупликацию, дифференциальную приватность и тесты определения участия записи в обучении или восстановления данных модели. Результат сканирования не подтверждает очистку набора данных или отсутствие утечки через модель.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке

Редакция: Методические рекомендации Банка России № 3-МР от 16.06.2026

п. 5.6 · Поддерживает часть требования

Подготовка обучающих данных без ненужных идентификаторов поддерживает рекомендованную очистку набора перед обучением у поставщика.

Граница связи: Приватность обучения не устанавливает условия передачи, роль поставщика и правовой статус набора; отсутствие прямых идентификаторов не доказывает обезличивание.

Субъект	Участник финансового рынка из п. 1.1 рекомендаций
Условие применения	Поставщик услуг обучает модель ИИ на наборах данных, принадлежащих участнику финансового рынка из пункта 1.1.
Условие связи с карточкой	Ненужные идентификаторы удаляются при подготовке набора до передачи для обучения; испытания готовой модели сами по себе этот этап не заменяют.
Сроки и переходные условия	Методические рекомендации № 3-МР носят рекомендательный характер.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-013 · Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели	532	113
AISAFE2-WEAK-033 · Раскрытие сведений об отдельных обучающих записях через модель	626, 627	206
AISAFE2-WEAK-037 · Непропорциональный вклад отдельных обучающих записей в результат обучения	631	211
AISAFE2-WEAK-059 · Недостаточная изоляция арендаторов при дообучении модели	668	253
AISAFE2-COMP-002 · Модель и интерфейс инференса	1034	37
AISAFE2-COMP-008 · Данные для подсистемы обучения	1035	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1036	54
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1037	67
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1038	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1039	79
AISAFE2-MIT-010 · Минимизация и фильтрация чувствительных данных до передачи в AI-контур	1226	349
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1343	469
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1349	485

Мера защиты · AISAFE2-MIT-043

Защита артефактов модели от несанкционированного получения

Краткое описание

Артефакты модели защищаются контролем доступа, учётом копий и ограничением экспорта на протяжении жизненного цикла; шифрование применяют по выбранной архитектуре.

Механизмы

- Авторизация
- Изоляция

Защитный механизм

Политика проверяет полномочия на исполнение, чтение и экспорт артефакта и разделяет их, если это требуется по риску и поддерживается платформой; хранилища и копии защищены границами доступа, а разделение ключей применяют при выбранной архитектуре. Инвентарь охватывает рабочие, резервные, временные и диагностические пути, а попытки чтения и экспорта журналируются и сопоставляются с владельцем решения.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Мера обычно применима только к артефактам, доступным организации для хранения, передачи, загрузки или исполнения. Организация выбирает закрытый контур, разделение прав, шифрование и контроль памяти по ценности артефакта, модели противника и возможностям платформы. Мера не заменяет проверку целостности и контролируемое продвижение модели. Для внешней модели, доступной исключительно через API, мера не применяется; риск извлечения через интерфейс ограничивается отдельным контролем.

Предпосылки

- Обязательные предпосылки выбранного контроля — инвентарь артефактов и копий, управляемый доступ и проверяемые границы хранения, передачи и экспорта.

Реализация

Контроль может сочетать ограничение доступа, шифрование, разделение ключей, раздельные полномочия на исполнение и экспорт, сетевую изоляцию и наблюдение за массовым получением. Закрытый контур является усиленным вариантом для наиболее чувствительных артефактов. Политика охватывает резервные, временные, диагностические и аварийные копии, а прямой публичный адрес и неконтролируемый экспорт несовместимы с назначением меры. Доступ к памяти процесса и ускорителя ограничивается там, где платформа предоставляет достоверный механизм; отсутствие такого механизма учитывается в остаточном риске.

Проверка

Проверяются применимые сценарии: нарушение разделения чтения и экспорта, запрещённая загрузка, прямой доступ к хранилищу, экспорт из резервной или диагностической копии, устаревшая роль, кража ключа, пользовательский путь выгрузки, доступ из среды исполнения и массовое чтение. Мера неэффективна, если артефакт можно получить через путь, не связанный с субъектом и решением, либо отзыв доступа не прекращает новые загрузки и экспорт.

Негативный тест

Попытка получить артефакт через непривязанный путь, после отзыва доступа или с чужой идентичностью должна завершиться отказом и наблюдаемым событием.

Подтверждения

Подтверждения включают инвентарь артефактов и копий, политики доступа, журналы чтения и экспорта, результаты негативных тестов и отчет учения по отзыву; конфигурации шифрования, разделенных ключей, сети и среды включаются в подтверждения, если эти механизмы выбраны.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Новая загрузка и экспорт блокируются, подозрительная среда изолируется, а применимые полномочия и управляемые организацией ключи отзываются. Продолжение рабочей функции допускается только по отдельному временному решению, если уже загруженный экземпляр

изолирован, штатный путь экспорта отсутствует и дополнительное выполнение не расширяет область возможного получения.

Исправление и повторная проверка

Закрывают непривязанный путь чтения или экспорта, отзывают сохранившиеся права и охватывают резервную либо диагностическую копию. Повторяют исходную загрузку после отзыва и от чужой идентичности: артефакт не выдается, событие наблюдаемо; сохраняют инвентарь копий, решение доступа и журнал отказа.

Побочные эффекты

Ограничения усложняют отладку и аварийное восстановление, а разделение прав и ключей увеличивает операционную нагрузку.

Остаточный риск

Исполняемую модель нельзя считать некопируемой. Привилегированный инсайдер, компрометация управляющей плоскости, доступ к памяти процесса или GPU, аппаратная атака и злоупотребление уже разрешенным исполнением могут привести к получению модели. После внедрения сохраняется риск полного или частичного копирования через привилегированный либо ранее разрешенный путь.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Key Management Service

Частичная поддержка. KMS-ключи обеспечивают шифрование объектов на стороне сервиса и позволяют разделить полномочия на шифрование и расшифрование.

Ограничение: KMS защищает хранимый объект, но не блокирует экспорт авторизованным субъектом и доступ к расшифрованной модели.

Yandex Object Storage

Частичная поддержка. Политика бакета позволяет ограничить операции чтения и изменения артефактов по субъекту, действию, объекту или префиксу.

Ограничение: Политика действует для объектов данного бакета; память процесса, GPU, временные и внешние копии остаются вне неё.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

При выбранном шифровании дополняет защиту артефакта контролем чтения, экспорта, ключей и учётом рабочих и резервных копий.

Ограничение: Карточка не требует шифрования во всех вариантах реализации и охватывает только артефакты модели; режимы остальных данных и транспортных каналов определяют отдельно.

Условие применения: Для доступных организации артефактов модели, если в их защите выбрано шифрование.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ограничивает хранение, копирование и экспорт артефактов модели.

Ограничение: Связь охватывает хранилища артефактов модели и их копии; хранилища обучающих данных, RAG-документов и журналов требуют собственных правил доступа.

Условие применения: Для модельных артефактов и их рабочих, резервных, временных и диагностических копий под управлением организации.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-012 · Кража модели	529	111
AISAFE2-WEAK-016 · Слабый контроль доступа к обучающим данным, конвейерам и артефактам	596	176
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	606	184
AISAFE2-WEAK-023 · Незашифрованное хранение	610	186
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	612	188
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	615	190
AISAFE2-WEAK-032 · Недостаточная защита физического доступа к модели и инфраструктуре	625	204
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1040	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1041	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1042	61
AISAFE2-MIT-012 · Проверка целостности и безопасная загрузка артефактов модели	1346	359
AISAFE2-MIT-044 · Ограничение извлечения модели через интерфейс	1347	485
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	1357	500
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1375	525

Мера защиты · AISAFE2-MIT-044

Ограничение извлечения модели через интерфейс

Краткое описание

Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели.

Механизмы

- Контроль ресурсов
- Обнаружение

Защитный механизм

Контроль связывает запросы с субъектом и арендатором, применяет сквозные бюджеты и коррелирует схожие, покрывающие пространство входов и распределенные последовательности. Интерфейс возвращает только необходимые результату поля и точность, не раскрывает необработанные оценки модели, градиенты, скрытые представления и диагностические сигналы без обоснования.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Мера обычно применима, если ценный или чувствительный интерфейс модели доступен недоверенным субъектам и позволяет многократные запросы. Она не защищает доступный артефакт «Защита артефактов модели от несанкционированного получения»; общий лимит «Ограничение потребления ресурсов и стоимости» и обнаружение «Выявление цепочек AI-атак и злоупотреблений» являются связанными мерами, но не полным механизмом. Для закрытого интерфейса мера может не применяться, если оценка риска не выявила внутреннего сценария извлечения.

Предпосылки

- Обязательные предпосылки — устойчивая идентификация субъектов и интерфейсов, учет сквозного бюджета и инвентарь информативных выходов.

Реализация

При признаках кампании по извлечению модели используются задержка, снижение доступного сигнала, дополнительная проверка субъекта, ручная проверка, приостановка или отзыв доступа. Водяной знак или контрольная метка могут помогать расследованию, но не заменяют ограничение интерфейса.

Проверка

Базовая оценка может опираться на бюджетные тесты и имитацию атак; обучение суррогатной модели является усиленным вариантом, если оно необходимо для оценки остаточной извлекаемости. Мера неэффективна, если ограничение легко сбрасывается сменой пути или не охватывает наиболее информативный выход.

Негативный тест

Подают обычную контрольную нагрузку и адаптивную кампанию извлечения модели, распределённую между идентичностями, версиями и диагностическим интерфейсом с низкой скоростью. Контрольная нагрузка должна остаться доступной, а совокупный бюджет кампании — сохраняться между маршрутами и привести к ограничению до превышения порога извлекаемого сигнала, зафиксированного в версии правила.

Подтверждения

Подтверждения включают перечень интерфейсов и полей ответа, конфигурацию бюджетов и корреляции субъектов, версию правил, результаты выбранных сравнительных испытаний, журналы кампаний и санкций и метрики ложных блокировок и потери полезности.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Включаются минимальный формат ответа, сниженные бюджеты и ограничение диагностических интерфейсов. Ранее заблокированная кампания по извлечению не должна автоматически разрешаться при потере корреляции.

Исправление и повторная проверка

Объединяют бюджеты эквивалентных путей и идентичностей, сокращают информативный ответ либо исправляют корреляцию кампании. Повторяют исходную адаптивную последовательность через смену интерфейса: ограничение совокупного бюджета срабатывает до достижения порога извлекаемого сигнала из версии правила, обычная нагрузка проходит; сохраняют версию правила и отчёт сравнения.

Побочные эффекты

Сильное ограничение снижает полезность сервиса, увеличивает задержку и может ошибочно блокировать исследовательскую или пакетную нагрузку.

Остаточный риск

Невозможно гарантировать предотвращение функционального копирования публично доступной модели. Атакующий может распределять запросы между субъектами и источниками. После внедрения сохраняется риск частичного извлечения и обучения замены на допустимых ответах.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Smart Web Security

Частичная поддержка. ARL может группировать HTTP-запросы по IP, заголовку, cookie и другим признакам, применять лимиты и блокировать превышение.

Ограничение: ARL считает HTTP-запросы по заданным признакам, но не связывает распределённые кампании и не сокращает информативность ответа модели.

Yandex SmartCaptcha

Частичная поддержка. SmartCaptcha повышает стоимость и может остановить автоматизированные массовые запросы к публичному интерфейсу модели.

Ограничение: Сервис не связывает запросы сквозным бюджетом, не обнаруживает семантическое покрытие пространства входов и не сокращает логиты или другие сигналы ответа.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

MITRE ATLAS data release v2026.09

Редакция: ATLAS release v2026.09

> **AML.T0024.002** · Поддерживает часть требования
>
> Сквозные лимиты запросов и уменьшение информативности ответа затрудняют сбор выборки для копирования модели через API.
>
> **Граница связи:** Мера не исключает извлечение через разрешённые ответы и не защищает прямой доступ к весам.
>
Субъект	Аналитик техник атак на AI-систему
Условие применения	Защищаемая модель доступна через API инференса, позволяющий многократно получать ответы для обучения её функциональной копии.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-012 · Кража модели	530	111
AISAFE2-WEAK-018 · Избыточная наблюдаемость и воспроизводимость поведения модели через интерфейс	598, 599	180
AISAFE2-WEAK-019 · Избыточное раскрытие внутренних оценок модели	600, 601, 602	182
AISAFE2-WEAK-026 · Отсутствие ограничений частоты запросов	618	194
AISAFE2-WEAK-027 · Отсутствие экономического ограничения массовых запросов	620	196
AISAFE2-WEAK-030 · Отсутствие проверяемой маркировки модели и её выходов	624	201
AISAFE2-COMP-002 · Модель и интерфейс инференса	1043	37
AISAFE2-COMP-013 · Телеметрия и обнаружение	1044	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1045	69
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1046	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1047	79

Карточка	Связи	Стр.
AISAFE2-MIT-006 · Ограничение потребления ресурсов и стоимости	1213	333
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1276	405
AISAFE2-MIT-038 · Контроль злоупотребления AI-сервисом	1334, 1348	462
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1347	481
AISAFE2-MIT-042 · Контроль приватности обучающих данных и утечек через модель	1349	478

Мера защиты · AISAFE2-MIT-045

Независимая проверка решений с существенными последствиями

Краткое описание

Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями.

Механизмы

- Человеческий контроль

Защитный механизм

Владелец процесса определяет по последствиям и нормативным требованиям классы решений, требующие независимой проверки, и факты, которые нельзя принимать только из AI-вывода. Для решений с этическими последствиями независимым основанием также служат этические критерии, зафиксированные для соответствующего класса решений; выявленное несоответствие блокирует завершение либо переводится на явную эскалацию.

Формы контроля

- Техническая
- Процессная
- Организационная
- Выполняемая человеком

Ответственность

- Заказчик

Применимость

Организация определяет применимость по существенности последствий и обязательным нормативным требованиям. Она не заменяет безопасное отображение «Проверка и безопасная обработка AI-вывода», информирование «Отображение происхождения и статуса подтверждения AI-вывода», подтверждение исполняемого действия агента «Подтверждение человеком высокорисковых действий до исполнения» и правовые процедуры конкретной области. Для решений без существенных последствий мера может не применяться при документированном принятии остаточного риска.

Предпосылки

- Нужны перечень решений с существенными последствиями, независимые от AI-вывода источники проверки, компетентный проверяющий и техническая блокировка завершения без его решения.
- Обязательный контекст, срок решения и маршрут обжалования задаются до эксплуатации.

Реализация

Интерфейс показывает исходные данные, происхождение, ограничения и расхождения, но не подменяет проверку предупреждением. Проверяющий фиксирует использованный независимый источник, основание решения и отклонение от рекомендации. Для выбранных классов система блокирует завершение при отсутствии проверки и поддерживает эскалацию, повторное рассмотрение и выборочный аудит против формального одобрения.

Проверка

Подают уверенный ошибочный AI-вывод, проверяют блокировку закрытия решения, происхождение независимого источника, компетентность и идентичность проверяющего и наличие записи об обжаловании; отдельно проводят корректный контрольный случай. Мера неэффективна, если тот же AI-вывод или производный от него текст используется как независимое подтверждение либо интерфейс позволяет завершить решение без проверки.

Негативный тест

Подают уверенный ошибочный AI-вывод и корректный контрольный вывод, затем для ошибочного используют производный от него «независимый» источник, неполный контекст, высокую очередь и замещающего проверяющего. Ошибочный вариант не должен закрыть решение без компетентной проверки по действительно независимым данным и доступного обжалования; корректный вариант проходит только с полным действующим подтверждением.

Подтверждения

Подтверждения включают перечень решений, требующих независимой проверки, правила блокирования, записи использованных независимых источников и оснований, результаты тестов обхода, метрики очереди и формальных одобрений, а также материалы выборочного аудита и обжалований.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Решение не принимается на основании AI-вывода. Процесс переходит к независимому ручному или детерминированному пути либо приостанавливается до доступности компетентной проверки.

Исправление и повторная проверка

Убирают производный AI-текст из независимого подтверждения, закрывают обход проверки и возвращают компетентному проверяющему исходные данные. Повторяют тот же уверенный ошибочный вывод: решение не завершается без независимого основания, обжалование доступно; сохраняют использованные источники и запись решения.

Побочные эффекты

Дополнительная проверка увеличивает срок и стоимость решения, а дефицит компетентных участников способен создать очередь.

Остаточный риск

Проверяющий подвержен эффекту чрезмерного доверия автоматической рекомендации и может использовать источник с общей причиной ошибки. После внедрения сохраняется риск согласованной ошибки AI и независимого канала либо формального человеческого участия.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке

Редакция: Методические рекомендации Банка России № 3-МР от 16.06.2026

п. 2.5 · Поддерживает часть требования

Независимая проверка человеком помогает оценить результат автоматической AI-операции до его окончательного принятия.

Граница связи: Пункт 2.5 также рекомендует возможность изменить результат автоматической операции. Проверка до принятия решения сама по себе не даёт полномочия или технического способа изменить уже выполненную операцию.

Субъект	Участник финансового рынка из п. 1.1 рекомендаций
Условие применения	Участник финансового рынка из пункта 1.1 использует ИИ для автоматических операций в критически важном процессе и оценил риски информационной безопасности ИИ как высокие.
Условие связи с карточкой	Независимая проверка выполняется человеком, способным оценить результат по данным, не производным от того же AI-вывода.
Сроки и переходные условия	Методические рекомендации № 3-МР носят рекомендательный характер.

Открыть положение в источнике

Приказ ФСТЭК России от 11.04.2025 № 117 Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных

информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений

Редакция: Приказ № 117 от 11.04.2025 с изменениями по приказу № 137 от 08.05.2026

п. 61(г) · Поддерживает часть требования

Независимая проверка до решения с существенными последствиями может ограничить использование недостоверного AI-ответа.

Граница связи: Мера действует только для выбранных решений; остальные функции системы и автоматическая реакция на недостоверный ответ требуют отдельных правил.

Субъект	Оператор системы, подпадающей под приказ ФСТЭК № 117
Условие применения	Решения или функции охваченной приказом № 117 системы зависят от ответов AI-сервиса, которые могут оказаться недостоверными.
Условие связи с карточкой	Выбранные решения с существенными последствиями нельзя завершить без независимой проверки исходных данных и вывода.
Начало действия положения	2026-03-01
Сроки и переходные условия	Пункт 61 действует с 1 марта 2026 года; приказ № 137 его не изменяет.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-002 · Последствия принятия решений на основе результатов AI-системы	493	93
AISAFE2-IMP-005 · Нарушение этических норм по отношению к пользователю AI-системы	499	98
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	514	101
AISAFE2-WEAK-001 · Избыточное доверие пользователя к AI-выводу	571	143
AISAFE2-WEAK-054 · Стохастическая недостоверность вывода генеративной модели	658	242
AISAFE2-WEAK-088 · Отсутствие проверяемой оценки этического воздействия AI-системы	709	301
AISAFE2-COMP-014 · Контур человеческой проверки и подтверждения	1048	65
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1049	67
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1050	72
AISAFE2-MIT-002 · Проверка и безопасная обработка AI-вывода	1197, 1350	320
AISAFE2-MIT-005 · Отображение происхождения и статуса подтверждения AI-вывода	1212, 1351	331
AISAFE2-MIT-037 · Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений	1327, 1353	459
AISAFE2-MIT-019 · Подтверждение человеком высокорисковых действий до исполнения	1352	389

Мера защиты · AISAFE2-MIT-046

Безопасная деградация при недоступности AI-зависимости

Краткое описание

При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим, не допуская частичного выполнения операции или опасных повторных попыток.

Область применения

- Общая безопасность

Механизмы

- Восстановление
- Жизненный цикл

Защитный механизм

Конечный автомат различает недоступность, задержку, частичный ответ и восстановление AI-зависимости. Для каждого состояния он выбирает режим, зафиксированный в матрице состояний и безопасных режимов, а прекращение повторных вызовов после серии ошибок и ключ идемпотентности не допускают повторного или противоречивого внешнего действия при переходе и возврате.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Третья сторона

Применимость

Мера обычно применима, если доступность или безопасное состояние функции зависит от модели, RAG, инструмента или внешнего AI API. Она не заменяет аварийную остановку автономного агента «Аварийная остановка и возврат агента в безопасное состояние», общий план инцидента «Локализация и восстановление после AI-инцидентов» и обязательства поставщика по доступности.

Предпосылки

- Нужны инвентарь AI-зависимостей, матрица их состояний отказа и заранее выбранный безопасный режим для каждой затронутой функции.
- Внешние действия должны иметь идемпотентную границу или иной способ исключить повторный эффект при тайм-ауте и возврате.

Реализация

На границе зависимости применяются подходящие механизмы безопасной обработки отказа: тайм-аут, прекращение повторных вызовов после серии ошибок, ограничение параллелизма, повторные вызовы с контролем повторного эффекта, классификация ошибок или их сочетание.

Для функции выбирается допустимый режим: не-AI алгоритм, последняя проверенная локальная версия, только чтение, очередь с идемпотентным повтором, ручной процесс, ограниченная функциональность или отказ операции. Конкретный набор зависит от архитектуры; переход и возврат не должны повторно выполнять платное, разрушительное или внешнее действие.

Проверка

Для каждого состояния зависимости вводят недоступность, задержку, частичный ответ и восстановление, измеряют переход, допустимый результат, время возврата и число внешних эффектов, включая повтор после тайм-аута. Мера неэффективна, если режим не определен для фактического состояния, нарушена идемпотентность либо переключение создает более тяжелое воздействие, чем отказ.

Негативный тест

Для контрольного запроса последовательно вводят тайм-аут, частичный ответ и восстановление AI-зависимости после того, как внешний приемник уже применил действие, затем повторяют запрос и отказ нескольких зависимостей. Функция должна перейти в безопасный режим из матрицы состояний и безопасных режимов, не повторить внешний эффект и вернуться в согласованное состояние; при штатной зависимости тот же запрос выполняется один раз обычным маршрутом.

Подтверждения

Подтверждения включают матрицу состояний и безопасных режимов, конфигурации тайм-аутов, повторов и ограничения параллелизма, результаты отказных испытаний, трассировки транзакций и идемпотентности, метрики переключения и подтверждение возврата без дублирования эффекта.

Поведение при отказе

Безопасная деградация

Сценарий отказа

Система остается в заранее выбранном безопасном режиме или блокирует функцию. Самопроизвольное продолжение через неподтвержденный ответ, неизвестную модель или неконтролируемую резервную зависимость запрещается.

Исправление и повторная проверка

Исправляют переход в безопасный режим, идемпотентность или классификацию частичного отказа, создавшие дублирование либо более тяжелое воздействие. Повторяют тот же тайм-аут или частичный ответ и восстановление под нагрузкой: система переходит в выбранный режим и возвращается без повторного внешнего действия; сохраняют трассировку состояния.

Побочные эффекты

Резервный режим может быть менее точным, свежим или функциональным и сам может отказать. Дополнительная логика усложняет систему и способна создавать ошибки переключения.

Остаточный риск

После внедрения сохраняется риск непредусмотренного сочетания отказов и решения на устаревших данных.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Workflows

Частичная поддержка. Тайм-ауты, управляемые переходы и ограниченные повторные попытки могут быть частью обработки отказа внешней AI-зависимости.

Ограничение: Тайм-аут и число повторов ограничивают шаг, но не дают готового безопасного режима без AI и не гарантируют идемпотентность внешнего действия.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-022 · Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала	554	127
AISAFE2-WEAK-064 · Неизолированные отказы внешних зависимостей	673	262
AISAFE2-WEAK-079 · Зависимость от единственного поставщика сервисов или площадки размещения	689	284
AISAFE2-COMP-002 · Модель и интерфейс инференса	1051	37
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1052	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	1053	48
AISAFE2-COMP-007 · Долговременная память	1054	50
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1055	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1056	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1057	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1058	75
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1059	82
AISAFE2-MIT-009 · Версионирование и откат базы знаний	1222	346
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1280	410
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1364	514

Мера защиты · AISAFE2-MIT-047

Устойчивость к состязательным входным данным

Краткое описание

Подготовка, ограничения входа и испытания на состязательных примерах повышают устойчивость классической ML- или мультимодальной модели.

Механизмы

- Валидация

Защитный механизм

Для модели задаются возможности атакующего, допустимые изменения входа и физические или цифровые каналы. По результатам применяются независимые подходы: нормализация и проверка согласованности сенсоров, обучение устойчивости, обнаружение аномалии, ансамбль, ограничение доверия, отказ модели от ответа и переход к безопасному или человеческому решению.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Режимы белого, серого и чёрного ящика выбираются по фактическому доступу атакующего и проверяющей стороны, а не образуют обязательную полную матрицу. Мера не заменяет проверку формата «Проверка входных данных по схемам и правилам», защиту от внедрения вредоносных инструкций «Разделение доверенных инструкций и недоверенного контента при сборке контекста» и «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» и общую проверку пригодности «Проверка пригодности модели для целевого сценария».

Предпосылки

- Обязательные предпосылки — модель угроз с реальными возможностями атакующего, точная версия модели и воспроизводимая незащищенная базовая конфигурация.

Реализация

Для точной версии и конфигурации фиксируют канал и знания атакующего, бюджет изменения и правила генерации и адаптации атак. Каждый защитный механизм связывают с решением модели, детектора, отказа или независимого маршрута и проверяют одновременно на состязательном и чистом контрольном наборе. Результаты, пороги качества и допустимые режимы отказа привязывают к выпускаемой версии; изменение модели, интерфейса или бюджета требует повторной оценки.

Проверка

Результат сравнивается с незащищенной базовой версией и воспроизводится на точной конфигурации. Мера неэффективна, если тест не соответствует реальному каналу либо защита исчезает после известной адаптации противника. В заявленной модели угроз доля успешных адаптивных атак, качество и ложные отказы на обычных данных должны оставаться в пределах, зафиксированных в методике оценки.

Негативный тест

Для точной версии запускают чистый контрольный набор и адаптивную состязательную атаку в выбранных по модели угроз режимах доступа, каналах и бюджетах, зафиксированных в методике оценки, включая применимые физические или цифровые преобразования. Защищённое решение должно выдержать порог из этой методики либо перейти к отказу или независимому маршруту, а качество и ложные отказы на чистом наборе — остаться в установленных пределах; неизвестная надёжность не считается успехом.

Подтверждения

Подтверждения включают модель угроз и диапазоны изменений, точную конфигурацию, версию набора атак, сравнительные результаты защищенной и базовой моделей, показатели обычного качества и ложных отказов и независимый воспроизводимый повтор.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Модель отказывается от решения, снижает автоматизацию или передает случай безопасному независимому процессу. Принудительное продолжение при неизвестной надежности входа запрещается для решений с существенными последствиями.

Исправление и повторная проверка

Обновляют модель угроз, тестовый канал или защиту для адаптации, на которой исходный состязательный вход сохранил эффект. Повторяют ту же атаку и контрольные чистые примеры на точной конфигурации: атака не приводит к защищаемому решению, а обычное качество остаётся в допуске из методики оценки; сохраняют бюджет атаки и сравнительный отчет.

Побочные эффекты

Устойчивость может снижать обычное качество и повышать число отказов; сложные испытания увеличивают стоимость выпуска и вычислений.

Остаточный риск

Универсальной устойчивости к состязательным входным данным не существует, а защита против одного класса атак может не работать против другого. После внедрения сохраняется риск адаптивной или физической атаки вне проверенного диапазона.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	506	100
AISAFE2-COMP-002 · Модель и интерфейс инференса	1060	37
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1061	67
AISAFE2-MIT-001 · Проверка входных данных по схемам и правилам	1196	316
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1201	324
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1207	328
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1277	405
AISAFE2-MIT-036 · Обнаружение и обработка дрейфа, смещения данных и деградации	1325	456
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1340	466

Мера защиты · AISAFE2-MIT-048

Аутентификация, целостность и защита от повтора межагентских сообщений

Краткое описание

Межагентское сообщение принимается только после проверки отправителя, получателя, целостности, связи с задачей, свежести и защиты от повтора.

Механизмы

- Целостность

Защитный механизм

Отправитель криптографически связывает с сообщением свою идентичность, получателя, идентификатор задачи, неизменяемые параметры действия, время выпуска, срок действия и уникальный идентификатор или одноразовое значение. Получатель до обработки проверяет подлинность отправителя, целостность канонического представления сообщения, соответствие получателя и задачи, актуальность срока и отсутствие уникального идентификатора в защищенном от гонок реестре повторов.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Мера применима к межагентским границам, где сообщение влияет на планирование, выбор инструмента, делегирование задачи или внешний эффект. Гранулярность связи с задачей и срок хранения идентификаторов повторов выбираются по длительности операций и возможным последствиям. Мера не определяет, является ли смысл сообщения безопасным, и не защищает от вредоносного агента, который корректно аутентифицирован и действует в пределах выданной ему роли.

Предпосылки

- Требуются достоверные идентичности сторон, защищенное управление ключами, согласованные часы или эквивалентный механизм срока действия и атомарная регистрация использованных идентификаторов.

Реализация

Сообщение с неверной подписью или MAC, измененным содержимым, чужим получателем или задачей, истекшим сроком либо уже использованным идентификатором отклоняется. Успешная проверка сообщения подтверждает его происхождение и свежесть, но не дает полномочия на внешний эффект: неизмененное действие независимо авторизуется непосредственно перед исполнением.

Проверка

В области меры каждое принятое сообщение должно иметь проверяемого отправителя, точного получателя и задачу, действующий срок и ранее не использованный идентификатор. Мера неэффективна, если измененное, чужое, просроченное или повторное сообщение влияет на обработку.

Негативный тест

Отправляют одно корректно подписанное контрольное сообщение, затем меняют получателя, задачу или параметр действия и воспроизводят исходное сообщение параллельно до и после истечения окна при отказе реестра повторов. Контрольное сообщение должно быть принято ровно один раз, а измененное, просроченное или повторное — отклонено до обработки; отказ проверяющей зависимости не должен разрешать сообщение.

Подтверждения

Подтверждения включают схему A2A-потоков, формат подписываемого сообщения, политику доверия и ротации ключей, параметры свежести, конфигурацию реестра повторов, журналы решений без секретного материала и результаты негативных и конкурентных тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Сообщение не обрабатывается, задача останавливается или переводится на отдельный подтвержденный канал. При недоступности проверки, неопределенной идентичности либо невозможности атомарно проверить повтор общий режим разрешения запрещен.

Исправление и повторная проверка

Исправляют проверку подписи, получателя, свежести или атомарный реестр повторов, позволившие сообщению повлиять на обработку. Повторяют исходное измененное сообщение и параллельную повторную отправку: оба отклоняются до планирования или действия, штатное сообщение принимается один раз; сохраняют решение проверки без секретного материала.

Побочные эффекты

Криптографическая обработка и реестр повторов добавляют задержку и состояние, а короткое окно может отклонить легитимную долгую операцию. Ошибка канонизации способна создавать несовместимость между реализациями.

Остаточный риск

Гарантии действуют только для охваченных протокола, идентичности, ключа, часов и реестра повторов. Скомпрометированный ключ, аутентифицированный вредоносный агент или ошибка классификации задачи всё ещё могут создать формально корректное сообщение с нежелательным эффектом.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-APP-INJECT2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Аутентифицирует отправителя и получателя и связывает задачу, свежесть, целостность и защиту от повтора для сообщений A2A.

Ограничение: Карточка не задаёт обязательную версию схемы, метку доверия и разрешённые типы сообщений для каждого отправителя. Права пользователя на передаваемые RAG-документы проверяются отдельно.

Условие применения: При передаче межагентского сообщения, влияющего на планирование, делегирование, выбор инструмента или внешнее действие.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Agentic Applications 2026

Редакция: OWASP Top 10 for Agentic Applications 2026, декабрь 2025 года

> **ASI07** · Поддерживает часть требования
>
> Проверка отправителя, получателя, связи с задачей, целостности и свежести сообщения отсекает подмену и повтор части межагентских команд.
>
> **Граница связи:** Корректно подписанное сообщение с опасным смыслом, скомпрометированный отправитель и утечки через маршрутизацию требуют иных мер.
>
Субъект	Аналитик угроз агентного AI-приложения
Условие применения	Агенты обмениваются сообщениями, для которых значимы отправитель, получатель, задача и защита от повтора.
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	515	101
AISAFE2-IMP-008 · Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)	521	104
AISAFE2-WEAK-049 · Слабые или отсутствующие механизмы установления доверия в A2A	652	232
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1062	42
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1063	61
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1064	82
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1251	383
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1301, 1356	429
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1354	324

Карточка	Связи	Стр.
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1355	400

Мера защиты · AISAFE2-MIT-049

Очистка и изоляция ресурсов общей среды инференса

Краткое описание

Изоляция и очистка общей среды инференса предотвращают перенос состояния между арендаторами, пользователями, задачами и последовательными заданиями.

Область применения

- Общая безопасность

Механизмы

- Изоляция
- Жизненный цикл

Защитный механизм

Каждое задание получает связанную с арендатором и задачей аренду ресурсов. До передачи ресурса другой границе доверия среда выполнения завершает работу с предыдущим контекстом, удаляет доступные ссылки, очищает либо безопасно переинициализирует буферы памяти GPU и хоста, записи кэша, временные файлы и состояние повторно используемого процесса.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применима при совместном или повторном использовании ресурсов, способных сохранять данные после завершения задания. Конкретные операции зависят от оборудования, драйвера, среды выполнения и оркестратора и должны подтверждаться для фактической версии. Наличие управляемого или облачного сервиса само по себе не доказывает очистку, изоляцию или эффективность контроля. Мера не применяется к полностью одноразовой изолированной среде без повторного назначения среды выполнения, кэша, пространства имён или временного объекта другому контексту.

Предпосылки

- Нужны инвентарь сохраняющих состояние ресурсов, границы арендатора и задачи, принудительная точка возврата ресурса в пул и поддерживаемый платформой способ очистки либо изоляции.

- Общая изоляция сети или доступа не подтверждает очистку памяти и временного состояния.

Реализация

Ресурс возвращается в общий пул только после подтвержденного завершения всех применимых шагов. Если очистка не поддерживается, не завершилась или ее результат неопределен, ресурс изолируется, пересоздается либо помещается в карантин; для существенного риска используется отдельный рабочий процесс, контейнер, виртуальная машина или эквивалентная граница. Та же граница применяется между независимыми запросами, сессиями и заданиями, даже если они принадлежат одному арендатору: кэш, пространство имен, временный объект и повторно используемая среда выполнения связываются с конкретным владельцем и сроком. Очистка и запрет повторного назначения действуют после успешного завершения, ошибки, тайм-аута, отмены, сбоя рабочего процесса и сверки состояния после перезапуска.

Проверка

Выполняется последовательный тест с контрольным значением A/B: задание арендатора A размещает уникальные контрольные значения во всех заявленных типах памяти, кэша, буфера и временного хранения, после завершения и очистки задание арендатора B пытается получить их через штатные и доступные низкоуровневые пути. Ни одно контрольное значение не должно быть доступно B. Дополнительно проверяются параллельное использование, аварийное завершение A, тайм-аут очистки, возврат рабочего процесса в пул, перезапуск процесса и обновление среды выполнения или драйвера. Мера неэффективна при обнаружении контрольного значения, повторном использовании до подтверждения очистки или наличии неучтенного сохраняющего состояние пути.

Негативный тест

После помещения уникального контрольного значения в задачу A аварийно завершают ее, возвращают рабочий процесс в пул и запускают задачу B другой границы доверия, включая параллельное назначение и тайм-аут очистки: B не должна увидеть контрольное значение.

Подтверждения

Подтверждения включают инвентарь ресурсов, конфигурацию аренды и карантина, документацию поддерживаемых операций, версии платформы, результаты A/B-тестов и журналы возврата ресурсов без содержимого пользовательских данных.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Ресурс не возвращается в общий пул и помещается в карантин либо уничтожается и пересоздается. Новое задание другой границы доверия на нем запрещается до подтвержденного восстановления контроля.

Исправление и повторная проверка

Устраняют сохраняющий состояние путь, исправляют карантин или запрещают возврат ресурса в пул до подтвержденной очистки. Повторяют исходный тест с контрольным значением после аварийного завершения и параллельной аренды: новая задача не видит контрольное значение, неподтвержденный ресурс остается в карантине; сохраняют журнал переходов ресурса.

Побочные эффекты

Переинициализация, отдельные рабочие процессы и карантин снижают утилизацию, увеличивают задержку и стоимость. Платформа может не предоставлять наблюдаемость для всех уровней аппаратного и системного состояния.

Остаточный риск

Неучтенный буфер, микропрограмма, побочный канал или привилегированный оператор могут сохранить доступ к данным, несмотря на корректную очистку известных ресурсов. Тесты подтверждают только охваченные пути, версии и режимы совместного использования.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA-ISOL2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Покрывает память, кэши, буферы, временные файлы, состояние сессии и процесса и независимые сессии одного арендатора после успеха, ошибки, тайм-аута, отмены и перезапуска.

Ограничение: Очистка повторно используемой среды не задаёт правила разрешённого долговременного состояния во внешнем хранилище: его авторизацию, срок и удаление определяют отдельно.

Условие применения: Когда среда последовательно или параллельно обслуживает независимые задания, сессии, пользователей либо арендаторов.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Изолирует и очищает память, кэш, буферы, временные файлы и состояние повторно используемого процесса между арендаторами и задачами.

Ограничение: Не покрывает идентичности, сеть и постоянные хранилища.

Условие применения: При совместном или последовательном использовании памяти, кэша, буферов, временных файлов либо процессов разными задачами и арендаторами.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	536	116
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	560	130
AISAFE2-IMP-032 · Несанкционированное смешение артефактов между изолированными контекстами	567	139

Карточка	Связи	Стр.
AISAFE2-WEAK-050 · Остаточные данные в общей инфраструктуре исполнения моделей	653	234
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1065	58
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1285	419
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1357	481
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1376	525

Мера защиты · AISAFE2-MIT-050

Неизменяемое версионирование и проверяемое восстановление наборов данных

Краткое описание

Неизменяемые версии и независимые копии позволяют проверить и восстановить точный набор данных для обучения, тестирования или валидации.

Область применения

- Общая безопасность

Механизмы

- Восстановление
- Целостность

Защитный механизм

Реестр связывает каждый значимый набор данных с неизменяемой версией, манифестом и независимыми копиями в отдельной границе отказа. Процедура восстановления выбирает требуемую версию, проверяет её состав и пригодность и только затем возвращает в рабочий контур.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Глубина защиты определяется критичностью набора, стоимостью воспроизведения, обязательствами хранения и допустимым временем простоя. Легко воспроизводимый низкорисковый набор исключают из области меры только по документированному решению

владельца после успешного восстановления требуемой версии из неизменяемых входов по проверяемому рецепту сборки. Резервная копия без проверки содержимого и сценария использования не считается подтвержденным восстановлением.

Предпосылки

- Требуются стабильная идентификация версий, полный манифест, доступные контрольные суммы и независимые полномочия на удаление и восстановление.
- Контроль доступа и подтверждение происхождения дополняют меру, но не заменяют независимую копию и учение по восстановлению.

Реализация

Каждая допущенная версия набора данных получает неизменяемый идентификатор, манифест состава и контрольные суммы. Неизменяемая или защищённая от удаления копия и как минимум одна независимая копия хранятся так, чтобы одна учётная запись, ошибка автоматизации или отказ домена не могли удалить все экземпляры. Политика удержания связана с владельцем и жизненным циклом модели. Восстановление выбирает именно требуемую версию, проверяет манифест и контрольные суммы и подтверждает пригодность данных для предусмотренного обучения, тестирования или валидации до их использования.

Проверка

Для каждого набора из области меры выполняется восстановление в изолированную среду в пределах заявленных RPO и RTO. Мера неэффективна, если удаление одной границей уничтожает все копии, нужная версия отсутствует, целостность не подтверждается или RPO/RTO не достигаются.

Негативный тест

Удаляют или повреждают основную копию контрольного набора данных и блокируют основной домен и учётную запись, затем запрашивают восстановление конкретной рабочей, но не последней версии. Восстановленная версия должна совпасть с манифестом и контрольными суммами и уложиться в заявленные RPO/RTO; поврежденная либо неполная копия не должна возвращаться в рабочий контур.

Подтверждения

Подтверждения включают реестр версий и владельцев, политики неизменяемости и хранения, схему независимых копий и полномочий, манифесты и контрольные суммы, отчеты учений и измеренные RPO/RTO.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Повреждённая или неподтверждённая версия не используется для обучения, тестирования, валидации или выпуска. Затронутый процесс блокируется либо переходит на предыдущую версию, которая уже прошла проверку восстановления для этого набора; отсутствие проверенной копии немедленно эскалируется владельцу риска.

Исправление и повторная проверка

Восстанавливают независимую копию, исправляют манифест либо полномочия удаления, из-за которых нужная версия не восстановилась. Повторяют удаление основной копии и восстановление именно исходной рабочей версии: контрольные суммы и проверка

пригодности совпадают, измеренные RPO/RTO укладываются в решение; сохраняют отчет учения.

Побочные эффекты

Независимые неизменяемые копии увеличивают стоимость хранения и усложняют выполнение законных требований об удалении; слишком долгий срок хранения может сам увеличить риск раскрытия данных.

Остаточный риск

Копия, независимая только по имени, может разделять с основной домен отказа, учётную запись или ошибку удаления; повреждение или нежелательное содержимое может попасть во все копии. Проверенное восстановление подтверждает только испытанные версию, объём и момент времени и не доказывает корректность исходного набора.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Создаёт неизменяемые версии набора данных и проверяемое восстановление.

Ограничение: Карточка охватывает наборы обучения и оценки. Версионирование и восстановление не авторизуют переходы между стадиями и не запрещают обратную запись.

Условие применения: Когда наборы данных для обучения, тестирования или валидации проходят несколько стадий и требуется сохранить или восстановить точную версию.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Использует неизменяемые идентификаторы, управляемые версии и проверяемое восстановление наборов данных.

Ограничение: Мера охватывает восстановление наборов для обучения и оценки; она не проверяет доверие к загрузке, автора изменения и разрешение на публикацию. Жизненный цикл RAG-корпуса требует отдельного контроля.

Условие применения: При изменении наборов данных для обучения, тестирования или валидации, для которых сохраняют предыдущую безопасную версию.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 7 · Поддерживает часть требования

Независимая проверенная версия позволяет восстановить персональные данные в наборе для обучения или оценки, если они были изменены либо уничтожены вследствие несанкционированного доступа.

Граница связи: Связь ограничена такими датасетами; восстановление остальных персональных данных и законность срока хранения копии требуют отдельной проверки.

Субъект	Оператор персональных данных
Условие применения	Набор для обучения, тестирования или валидации содержит персональные данные, изменённые или уничтоженные вследствие несанкционированного доступа.
Условие связи с карточкой	Сохранилась независимая копия нужной версии, а восстановление её состава и пригодности проверено.

Открыть положение в источнике

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM05:2026 · Поддерживает часть требования

Неизменяемые версии и проверенное восстановление набора позволяют сравнить изменения и вернуть пригодную версию после выявления отравления.

Граница связи: Карточка охватывает наборы обучения и оценки. Копия не выявляет вредную запись и может содержать то же загрязнение; пригодность выбранной версии проверяется отдельно. RAG-корпус требует отдельного контроля восстановления.

Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.
Условие связи с карточкой	Отравление затронуло набор данных для обучения, тестирования или валидации; для восстановления выбирают проверенную пригодную версию.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-031 · Утрата датасета в результате атаки	563	137
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	700	296
AISAFE2-COMP-008 · Данные для подсистемы обучения	1066	52
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1344	469
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1365	514

Карточка	Связи	Стр.
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1377	525

Мера защиты · AISAFE2-MIT-051

Контроль полноты обязательного анализа и безопасный выпуск результата

Краткое описание

Неизменяемый план обязательных проверок и блокирующий финальный контроль не выпускают штатный или разрешающий результат, пока все обязательные стадии объявленного плана не завершены.

Область применения

- Общая безопасность

Механизмы

- Шлюз допуска

Защитный механизм

Для принятого объекта среда выполнения формирует ограниченный неизменяемый план анализа и принимает от каждой обязательной стадии ровно одно конечное состояние. Выпуск результата разрешается только после проверки полноты и непротиворечивости всего плана.

Формы контроля

- Техническая

Ответственность

- Заказчик

Применимость

Организация определяет обязательные стадии и допустимую неприменимость отдельно для фактических маршрутов и профилей. Мера подтверждает полноту только объявленного плана: она не доказывает качество детектора, достаточность политики, полноту выборки и безопасность содержимого, которое успешно прошло все стадии.

Предпосылки

- Требуются неизменяемая версия профиля, однозначные идентификаторы стадий, закрытый набор конечных состояний и принудительный путь через финализатор.

Реализация

Отсутствующий результат, повторная или переписанная запись, ошибка, истечение срока, отмена либо необоснованная неприменимость делают анализ неполным. Финализатор проверяет полноту до выдачи и в этом случае возвращает фиксированный результат, который не разрешает действие и не выдает защищаемое содержимое. Транспорт, телеметрия, резервная ветвь и пакетная обработка не должны превращать его в обычный успех, частичный успешный ответ или пустой разрешающий результат.

Проверка

Для проверки запускают каждый поддерживаемый транспорт с инструментированным планом обязательных стадий и сопоставляют решение финализатора с зафиксированными конечными состояниями.

Негативный тест

Через каждый транспорт подают полный контрольный план и варианты без обязательной стадии, с повторной или переписанной записью, тайм-аутом и ошибкой преобразования, затем повторяют через прямой обход и смешанный пакет. Полный согласованный план должен выдать обычный результат, а любой неполный или противоречивый вариант — только фиксированный неразрешающий результат без защищаемого внешнего эффекта.

Подтверждения

Сохраняются версия профиля, план без пользовательского содержимого, конечные состояния, фиксированный код отказа и результаты негативных тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Защищаемый результат не выпускается, а ошибка регистрации события не меняет решение.

Исправление и повторная проверка

Возвращают отсутствующую обязательную стадию в план, исправляют финализатор или закрывают транспортный обход, выпустивший неполный результат. Повторяют тот же тайм-аут, исключение и прямой вызов: защищаемый результат не выпускается до полного согласованного набора состояний; сохраняют версию профиля и конечные состояния.

Побочные эффекты

Ошибка в составе обязательных стадий может оставить проверку вне плана или необоснованно блокировать запросы. Дополнительная фиксация состояний увеличивает задержку и требования к надёжности среды выполнения.

Остаточный риск

Все объявленные стадии могут завершиться штатно, но использовать слабую политику, недостаточный детектор или неполную выборку.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-006 · Неожидаемые (для разработчика AI-системы) ответы моделей	507	100
AISAFE2-IMP-007 · Обход ограничений или предусмотренной логики работы Agent Core	516	101
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	561	130
AISAFE2-WEAK-051 · Выдача результата без завершения обязательных стадий вычислений	654	235
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1067	72

Карточка	Связи	Стр.
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1068	77
AISAFE2-MIT-039 · Проверка пригодности модели для целевого сценария	1341	466
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1366	514

Мера защиты · AISAFE2-MIT-052

Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации

Краткое описание

Управляемый вывод из эксплуатации отключает маршруты и полномочия компонента и доводит до решения вопросы его данных, копий, интеграций и внешних обязательств.

Область применения

- Общая безопасность

Механизмы

- Жизненный цикл

Защитный механизм

Инвентарь охватывает исходные данные, производные артефакты, временные и реплицированные копии, кэши, журналы, резервные копии, идентичности, учетные данные, разрешения, доступные действия, конечные точки, маршруты, интеграции и копии, которые внешняя сторона сохраняет в пределах управляемого жизненного цикла. Для каждого объекта фиксируются владелец, назначение, срок хранения и решение: удалить, передать, сохранить до установленного срока либо удерживать по действующему обязательному основанию.

Формы контроля

- Техническая
- Процессная
- Организационная
- Договорная

Ответственность

- Заказчик
- Третья сторона

Применимость

Глубина инвентаря и способ удаления определяются архитектурой, классами данных, сроками хранения, моделью поставки и контролем над внешними сторонами. Мера применяется как к плановой замене, так и к аварийному выводу. Она не обещает мгновенное физическое уничтожение резервной копии и не подменяет отдельное решение о законном сроке хранения. Мера не применяется к компоненту, который не вводился в эксплуатацию и не создал управляемых данных, полномочий, маршрутов, интеграций или внешних обязательств.

Предпосылки

- Нужны актуальный инвентарь компонентов и потоков, владельцы объектов и полномочий, правила хранения, способ подтвердить удаление или истечение срока и маршрут работы с внешними копиями.

Реализация

Сначала прекращаются новые вызовы и отключаются возможности, маршруты и интеграции компонента, затем отзываются относящиеся к нему учетные данные и разрешения и удаляются активные, производные и временные объекты. Реплики, резервные и внешние копии остаются открытыми обязательствами до подтвержденного удаления или окончания срока. Общий ресурс не уничтожается, если он обслуживает другие системы; из него удаляются пространство имен, доступ и возможность действия выводимого компонента.

Проверка

После вывода проверяются вызов старой конечной точки, использование отозванных учетных данных, выполнение прежнего действия, чтение активного и производного объекта, наличие временной копии, восстановление из реплики или резервной копии и статус внешней сохраненной копии.

Негативный тест

После вывода проверяются старая конечная точка, отозванные учётные данные, прежнее полномочие, активная копия и восстановление из реплики; ни один закрытый объект не должен оставаться доступным.

Подтверждения

Сохраняются акт по каждому объекту, результаты негативных проверок, подтверждения отзыва и удаления, сроки исчезновения отложенных копий, владельцы и даты следующего действия.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Вывод не закрывается. Затронутый маршрут, идентичность, пространство имен или интеграция блокируются от повторного использования, открытая зависимость получает владельца и срок, а после исправления повторяется исходная проверка.

Исправление и повторная проверка

Закрывают оставшуюся конечную точку, отзывают действующее полномочие либо добавляют неучтённую копию в обязательства вывода. Повторяют исходный доступ и восстановление из реплики: объект недоступен, полномочие не действует, незавершённая внешняя зависимость остаётся открытым обязательством; сохраняют акт и результаты проверок.

Побочные эффекты

Удаление может быть отложено резервным хранением, обязательным удержанием или возможностями внешней стороны. Ошибка при работе с общим ресурсом способна нарушить другие системы.

Остаточный риск

Неучтенная копия, общий административный контур, физический носитель или внешняя сторона могут сохранить данные или доступ после логического вывода. Подтверждение относится только к перечисленным объектам, полномочиям и срокам.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

Yandex Cloud AI Studio

Частичная поддержка. Responses API позволяет удалить сохранённый объект Response по его идентификатору. Для поискового индекса Vector Store можно задать expires_after для автоматического удаления после периода неактивности или выполнить удаление вручную.

Ограничение: Удаляется названный Response или индекс Vector Store; связанные файлы, Conversations, журналы и резервные копии живут по отдельным правилам. Для конкретного развёртывания выполнение всех операций надо проверить.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Закрывает маршруты, учётные данные, интеграции и решения по копиям данных при выводе из эксплуатации.

Ограничение: Связь относится к завершению жизненного цикла; актуальную карту получения, обработки, резервирования, поддержки и передачи данных до вывода ведут отдельно.

Условие применения: При выводе из эксплуатации AI-компонента, который обрабатывал регулируемые данные.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DECOM1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Отключает маршруты, учётные данные и интеграции и учитывает сохраняемые и удаляемые копии данных и артефактов.

Ограничение: Для полного выполнения требования дополнительно нужны проверка готовности процедуры для каждого продуктивного компонента, экспорт конфигурации перед разрушающими операциями и повторная инвентаризация остаточных объектов.

Условие применения: При подготовке к выводу любого продуктивного AI-компонента и при самом выводе модели, индекса, агента или приложения.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-REL2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Прекращает вызовы и доступ выводимого компонента, закрывает его интеграции и учитывает удаляемые и сохраняемые копии.

Ограничение: Вывод из эксплуатации не задаёт совместимый комплект модели, индекса и конфигурации для отката и не равнозначен немедленному отзыву точной версии со всех путей.

Условие применения: При плановой замене или аварийном выводе модели, индекса либо другого AI-компонента из эксплуатации.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Федеральный закон от 26.07.2026 № 243-ФЗ О поддержке развития технологий искусственного интеллекта в Российской Федерации

Редакция: Оригинальный текст № 243-ФЗ от 26.07.2026

ст. 8, п. 2 · Поддерживает часть требования

Управляемый вывод AI-компонента из эксплуатации поддерживает часть правил обращения с моделью на завершающем этапе её жизненного цикла.

Граница связи: Связь относится к разработчику суверенной или национальной большой фундаментальной модели в смысле № 243-ФЗ с 1 марта 2027 года; правила её применения и обновления задаются отдельно.

Субъект	Разработчик суверенной или национальной большой фундаментальной модели
Условие применения	Модель и её разработчик соответствуют определениям и условиям № 243-ФЗ.
Условие связи с карточкой	Разработчик устанавливает правила вывода этой модели из эксплуатации.
Начало действия положения	2027-03-01
Сроки и переходные условия	Статья 8 вступает в силу 1 марта 2027 года.

Открыть положение в источнике

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Редакция: NIST AI 100-1, AI RMF 1.0

GOVERN 1.7 · Поддерживает часть требования

Инвентарь активов, отзыв полномочий и проверка старых маршрутов дают проверяемую процедуру вывода AI-компонента из эксплуатации.

Граница связи: План замены, коммуникации и оценка риска после вывода остаются организационными решениями.

Субъект	Организация, применяющая NIST AI RMF к AI-системе
Условие применения	В организации выбран процесс поэтапного вывода AI-системы из эксплуатации в рамках добровольного NIST AI RMF.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-015 · Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)	537	116
AISAFE2-WEAK-035 · Недостатки управления жизненным циклом данных	629	207
AISAFE2-WEAK-053 · Небезопасный вывод AI-компонента из эксплуатации	656	240
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1069	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	1070	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1071	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1072	44
AISAFE2-COMP-007 · Долговременная память	1073	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	1074	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1075	54
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1076	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1077	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1078	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	1079	64
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1080	67
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1081	69
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1082	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1083	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1084	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1085	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1086	82
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	1087	84
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1088	86
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1309	448
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1314	451
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1367	514
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1378	525

Мера защиты · AISAFE2-MIT-053

Прослеживаемая архитектура, модель угроз и пересмотр риска

Краткое описание

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске.

Область применения

- Общая безопасность

Механизмы

- Шлюз допуска
- Наблюдаемость
- Управление и надзор

Защитный механизм

Описание включает компоненты, границы доверия, субъекты, классы данных и пути к модели, RAG-контуру, памяти, инструментам, внешним получателям, журналам и хранилищам. Для каждой значимой связи фиксируются источник, назначение, идентичность, класс данных, протокол, аутентификация, сетевой путь, источник события и владелец; отдельно — разрешённые и запрещённые сценарии. Достоверный сигнал о компрометации, критической уязвимости, отзыве, существенном изменении условий или недоступности связывается с точными зависимостями, версиями, конечными точками и выпусками и запускает решение об ограничении, замене и повторных проверках.

Формы контроля

- Архитектурная
- Процессная
- Организационная

Ответственность

- Заказчик

Применимость

Детализация определяется последствиями ошибки и сложностью системы. Для малого сценария допустима компактная схема, если она перечисляет все значимые границы и проверяемые потоки. Для внешней или поставляемой зависимости должны быть определены применимые каналы сигналов и способ сопоставления с точной рабочей версией. Мера не заменяет технические контроли, исправление уязвимости, безопасный режим отказа и договорную оценку поставщика и не доказывает полноту модели угроз или уведомлений.

Предпосылки

- Требуются доступный инвентарь фактических ресурсов и зависимостей, стабильная версия описания и назначенные владельцы значимых связей.
- Для каждой поставляемой зависимости нужны точная версия или иной неизменяемый идентификатор, применимые официальные либо договорные каналы сигналов и владелец

их обработки.

Реализация

Редакция сопоставляется с фактическим инвентарём. Изменение модели, корпуса, инструмента, поставщика, конечной точки, идентичности, класса данных или политики, а также сигнал о компрометации, критической уязвимости, отзыве, изменении условий обработки или недоступности запускают пересмотр до выпуска или продолжения затронутого использования. Для сигнала фиксируют источник, время, область и уверенность; его сопоставляют с точными версиями и выпусками, а неизвестную область не считают незатронутой. Результат определяет затронутые системы, владельца, решение сохранить, ограничить или заменить зависимость, обязательные повторные проверки, срок и согласующего.

Проверка

Сопоставляют версионируемую схему с фактическим инвентарём и трассируют разрешённые и запрещённые потоки через значимые границы. Затем подают тестовый официальный сигнал об отзыве точной версии: процедура должна найти все затронутые выпуски, не пометить неподтверждённую область безопасной, создать повторную оценку, решение об ограничении или замене и перечень повторных проверок. Контрольная незатронутая версия должна остаться вне области решения с зафиксированным обоснованием. Неизвестный рабочий компонент или поток считается блокирующим расхождением.

Негативный тест

Добавляют неизвестный компонент и проводят поток в обход описанной границы, затем подают сигнал об отзыве рабочей версии и одновременно отключают один заявленный канал уведомлений. Сверка должна обнаружить расхождение, определить затронутые выпуски, перевести неизвестную область в блокирующее состояние и потребовать решение до продолжения; незатронутая контрольная версия не должна ограничиваться без основания.

Подтверждения

Подтверждения включают точную редакцию архитектуры, фактический состав и версии, реестр каналов сигналов, исходное уведомление и его время, результаты сопоставления, затронутые выпуски, решение об ограничении или замене и результаты повторных проверок.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Неучтённый компонент, поток или затронутая версия не допускаются к выпуску либо изолируются от продолжения использования. Потеря обязательного канала или невозможность определить область сигнала переводят состояние в неизвестное и требуют эскалации; они не трактуются как отсутствие воздействия.

Исправление и повторная проверка

Восстанавливают канал, уточняют идентификаторы зависимости и затронутые выпуски, обновляют модель угроз и решение о применимых мерах. Повторяют исходный сигнал: каждая затронутая версия получает ограничение, замену либо принятое решение и обязательные проверки, а контрольная незатронутая версия остаётся доступной с обоснованием.

Побочные эффекты

Ложный или неоднозначный сигнал может временно ограничить допустимую зависимость и задержать выпуск; дополнительные каналы и переоценки увеличивают операционную нагрузку.

Остаточный риск

Неизвестное событие, ошибочное сопоставление версии или неполный инвентарь могут оставить затронутый выпуск в работе. Согласованная, но ошибочная или неполная модель угроз может пропустить реальный путь воздействия; формально завершённая переоценка также может принять недостаточное ограничение или проверку.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Инвентаризирует топологию, зоны доверия, сетевые пути и разрешённые или запрещённые потоки.

Ограничение: Описание архитектуры не обеспечивает принудительный контроль соединений.

Условие применения: Для входящих, исходящих и межкомпонентных соединений ИИ-системы в Yandex Cloud.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SDLC1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Версионирует архитектуру, границы доверия, субъекты, классы данных, пути, риски, меры и пересмотр до выпуска.

Ограничение: Для полного выполнения требования каждый существенный риск дополнительно связывают с конкретной границей, мерой и способом проверки. Неизвестный поток остаётся блокирующим расхождением; полнота модели угроз не гарантируется.

Условие применения: Перед первым продуктивным выпуском ИИ-системы в Yandex Cloud и после изменения архитектуры, данных, модели, инструмента или внешней связи.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SUPPLY6

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Официальный сигнал о компрометации, критической уязвимости, отзыве, существенном изменении условий или недоступности связывается с точными зависимостями, версиями, конечными точками и выпусками и запускает решение об ограничении, замене и повторных проверках.

Ограничение: Охват зависит от выбранных каналов и точных идентификаторов зависимости. Сопоставление с выпуском не доказывает полноту уведомлений или устранение причины.

Условие применения: Для поставляемых моделей, данных, библиотек, образов и внешних API с определёнными каналами уведомлений.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке

Редакция: Методические рекомендации Банка России № 3-МР от 16.06.2026

п. 3.1 · Поддерживает часть требования

Мера создаёт и пересматривает модель угроз безопасности информации AI-системы как вход для выбора мер защиты.

Граница связи: Состав угроз и методику организация определяет для точной финансовой системы; наличие архитектурной схемы не доказывает полноту модели угроз.

Субъект	Участник финансового рынка из п. 1.1 рекомендаций
Условие применения	Участник финансового рынка из пункта 1.1 разрабатывает или применяет AI-систему и определяет актуальные угрозы безопасности её информации.
Условие связи с карточкой	Версионируемая модель угроз описывает фактические AI-компоненты, потоки и границы системы.
Сроки и переходные условия	Методические рекомендации № 3-МР носят рекомендательный характер.

Открыть положение в источнике

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 1 · Поддерживает часть требования

Версионируемая модель угроз помогает определить опасные для персональных данных границы AI-компонентов в информационной системе.

Граница связи: Статус ИСПДн, все её потоки, актуальные угрозы и предписанный уровень защищённости оператор устанавливает отдельно; наличие AI-модели угроз этого не подтверждает.

Субъект	Оператор персональных данных
Условие применения	Персональные данные обрабатываются в ИСПДн, в состав которой входят описываемые AI-компоненты.
Условие связи с карточкой	Модель угроз охватывает потоки персональных данных и границы AI-компонентов этой ИСПДн.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	641	222
AISAFE2-WEAK-071 · Отсутствие явной модели доверенных зон и потоков данных	680	271
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	696	294
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1089	67
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1090	79
AISAFE2-COMP-022 · Программные артефакты и цепочка поставки	1091	84
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1264, 1360	400
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1278	405
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1286	419
AISAFE2-MIT-035 · Ведение проверяемого состава и происхождения компонентов AI-системы	1315	451
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1358	383
AISAFE2-MIT-020 · Аварийная остановка и возврат агента в безопасное состояние	1359	392
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1361	415
AISAFE2-MIT-032 · Проверка исходящих данных перед передачей через коннекторы	1362	441
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1363	448
AISAFE2-MIT-046 · Безопасная деградация при недоступности AI-зависимости	1364	492
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных	1365	503
AISAFE2-MIT-051 · Контроль полноты обязательного анализа и безопасный выпуск результата	1366	507
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1367	509
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1368	519
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	1369	531
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1370	573
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1371	577

Мера защиты · AISAFE2-MIT-054

Принудительное соблюдение утвержденных ограничений обработки и передачи данных

Краткое описание

Контроль потоков данных проверяет и применяет утверждённые условия обработки, хранения и внешней передачи непосредственно перед выполнением операции.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Управление конфигурацией

Защитный механизм

Версионируемое решение задает назначение, классы данных, разрешенные места, получателей, конечные точки, последующую передачу, сроки хранения и применимые исключения. Перед обработкой, записью и внешней передачей технический шлюз сопоставляет фактический поток с этими атрибутами; неизвестный или изменившийся атрибут блокирует операцию до нового решения.

Формы контроля

- Техническая
- Процессная
- Организационная
- Договорная

Ответственность

- Заказчик

Применимость

Мера применяется только там, где существует действующее решение уполномоченного владельца и технически наблюдаемые атрибуты потока. Для каждого исключения фиксируются владелец, срок, компенсирующая мера и повторный пересмотр. Неизвестное место обработки или получатель не считается разрешением по умолчанию. Мера не применяется только к потоку, на который после подтверждённой инвентаризации не распространяется ни одно утверждённое ограничение обработки, хранения или передачи.

Предпосылки

- Требуются актуальное версионируемое решение, классификация данных, полный инвентарь существенных копий и принудительная точка до обработки, хранения или передачи.

Реализация

Инвентарь охватывает исходный объект, контекст модели, RAG, память, кэш, журнал, резервную копию, экспорт и внешнего обработчика. Шифрование, TLS, регион ресурса, договор без точной области и наличие функции поставщика не заменяют это решение.

Проверка

Сначала выполняется разрешенный синтетический поток, затем по одному изменяются класс данных, получатель, конечная точка, место, назначение и срок хранения.

Негативный тест

В разрешённом синтетическом потоке по одному заменяют класс данных, назначение, место, получателя, конечную точку и срок на заведомо запрещённое или неизвестное политике значение. Каждый вариант должен получить отказ до обработки, записи или передачи. Исходный поток и вариант с другим разрешённым значением должны проходить после отдельного решения шлюза.

Подтверждения

Сохраняются версия решения, фактические атрибуты потока, решение шлюза без чувствительного содержимого, результаты негативных тестов, исключение и срок пересмотра.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Обработка, запись или отправка при неизвестном атрибуте блокируется; допускается только заранее определённый режим без передачи и создания новой копии.

Исправление и повторная проверка

Исправляют классификацию атрибута, версию утвержденного решения или точку принудительной проверки, позволившие создать неразрешенную копию. Повторяют тот же синтетический поток с измененным назначением или получателем: обработка блокируется до записи или передачи, разрешенный поток проходит; сохраняют версию решения и журнал шлюза без чувствительного содержимого.

Побочные эффекты

Контроль зависит от полноты классификации, инвентаря и внешнего решения и может блокировать допустимый поток при их рассинхронизации. Он не управляет поведением получателя после передачи за пределами охваченной границы.

Остаточный риск

Ошибочное или устаревшее решение, скрытый путь, неверная классификация или неучтенный внешний обработчик могут привести к запрещенной обработке несмотря на корректную работу шлюза.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Принудительно исполняет утверждённое решение по наблюдаемым атрибутам цели, класса, получателя, места и срока хранения.

Ограничение: Карточка не определяет правовое основание, роли сторон и корректность утверждённого решения; эти вопросы решает уполномоченная сторона организации.

Условие применения: При продуктивной обработке персональных или иных регулируемых данных по решению уполномоченной стороны организации.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Принудительно исполняет утверждённые место, получателя, назначение и срок для модели, RAG, памяти, кэша, журнала, резервной копии и экспорта и блокирует неизвестные копии.

Ограничение: Инвентарь и проверка атрибутов потока не заменяют полную карту жизненного цикла с владельцами ресурсов, мерами защиты и способом удаления каждой регулируемой копии.

Условие применения: Для регулируемых данных во всех управляемых копиях: модельном контексте, RAG, памяти, кэше, журнале, резервной копии и экспорте.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Блокирует неизвестного или изменённого получателя, место, конечную точку, назначение, класс или срок до нового решения.

Ограничение: Техническое исполнение внешнего решения не определяет допустимость и договорные условия каждого получателя.

Условие применения: Для внешней передачи данных по действующему решению о получателе, назначении, месте обработки и сроках.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY4

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Технически исполняет внешне утверждённое решение о стране, месте, получателе, конечной точке, классе и назначении с запретом по умолчанию.

Ограничение: Исполнение списка стран и получателей не квалифицирует поток как трансграничный и не принимает правовое решение о передаче.

Условие применения: Если поток данных, запрос, результат, журнал или копия может пересечь границу, определённую применимым правом или политикой размещения.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Исполняет утверждённые цель, получателя, место и срок для фактического потока.

Ограничение: Контроль использует утверждённые сведения о наборе, но не создаёт полную запись его происхождения, преобразований и стадий жизненного цикла.

Условие применения: При обработке набора данных по уже утверждённым ограничениям назначения, места, получателя и срока.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет утверждённые атрибуты потока данных до передачи или сохранения.

Ограничение: Не покрывает все сетевые соединения и соединения контура управления.

Условие применения: Для соединений, по которым передаются или сохраняются данные с утверждёнными ограничениями обработки.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Проверяет назначение, класс данных, получателя, конечную точку, место и последующую передачу перед внешней операцией.

Ограничение: Не контролирует весь сетевой путь, TLS и жизненный цикл секрета.

Условие применения: При вызове модельного API вне Yandex Cloud, если для передаваемых данных установлены ограничения назначения, получателя, места или срока.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 18, ч. 5 · Поддерживает часть требования

Технический контроль может заблокировать запись и иные охваченные операции с данными граждан РФ в неразрешённой базе при их сборе.

Граница связи: Нужно установить факт сбора данных граждан РФ, место каждой базы и применимость исключений пунктов 2, 3, 4, 8 части 1 статьи 6; карта исполняет только предварительно утверждённое решение.

Субъект	Оператор персональных данных
Условие применения	При сборе персональных данных граждан РФ, в том числе через Интернет, выполняются запись, систематизация, накопление, хранение, уточнение или извлечение с использованием базы за пределами РФ; исключения пунктов 2, 3, 4, 8 части 1 статьи 6 не действуют.
Условие связи с карточкой	Шлюз различает фактическую операцию и местонахождение базы и применяет утверждённый запрет до её выполнения.
Начало действия положения	2025-07-01
Сроки и переходные условия	Редакция части 5 статьи 18 по № 23-ФЗ применяется с 1 июля 2025 года.

Открыть положение в источнике

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 25(1)-(2) · Поддерживает часть требования

Шлюз применяет утверждённые ограничения обработки, хранения и передачи данных при фактической операции и блокирует неизвестный маршрут по умолчанию.

Граница связи: Техническое правило зависит от правового решения и не охватывает весь объём защиты данных по замыслу и по умолчанию.

Субъект	Контролёр персональных данных по GDPR
Условие применения	Установлены фактическая операция обработки и соразмерность меры риску.
Начало действия положения	2018-05-25

Открыть положение в источнике

Article 44 · Поддерживает часть требования

Шлюз может заблокировать передачу персональных данных получателю или в место, не разрешённое утверждённым решением.

Граница связи: Правовое основание и гарантии передачи по главе V устанавливаются отдельно; другие пути и последующие передачи требуют проверки.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Установлена фактическая передача персональных данных в третью страну и территориальная применимость GDPR.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-024 · Утечка чувствительных данных в недоверенный контур	562	130
AISAFE2-WEAK-046 · Недостаточное управление соответствием требованиям	646	225
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель	660	245
AISAFE2-WEAK-061 · Избыточно открытая внутренняя телеметрия	670	255
AISAFE2-WEAK-070 · Отсутствие управляемой политики допустимого использования AI-сервисов	679	270
AISAFE2-WEAK-072 · Отсутствие обязательного технического контроля межзонных потоков данных	681	273
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	704	299
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1092	35
AISAFE2-COMP-002 · Модель и интерфейс инференса	1093	37
AISAFE2-COMP-003 · Управление контекстом общения с моделью	1094	40
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1095	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1096	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	1097	48
AISAFE2-COMP-007 · Долговременная память	1098	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	1099	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1100	54
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1101	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1102	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	1103	64
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1104	69

Карточка	Связи	Стр.
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1105	72
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1106	75
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1107	77
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1108	79
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1109	82
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1110	86
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1265	400
AISAFE2-MIT-028 · Управление подтверждениями и остаточным риском утверждений	1291	426
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1368	514
AISAFE2-MIT-057 · Принудительное ограничение сетевых путей AI-компонентов	1380	531
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1387	571

Мера защиты · AISAFE2-MIT-055

Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps

Краткое описание

Ролевой процесс выдаёт, пересматривает и отзывает права операторов, разработчиков и администраторов MLOps при приёме, смене обязанностей и завершении работ.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Управление и надзор

Защитный механизм

Для каждой человеческой и административной учётной записи владелец процесса связывает роль MLOps с определённым для неё набором полномочий, основанием, сроком и владельцем доступа. Выдача проходит согласование, плановый пересмотр подтверждает сохраняющуюся необходимость, а смена обязанностей или завершение работ запускают отзыв прежних прав до назначения новых.

Формы контроля

- Процессная
- Организационная

Ответственность

- Заказчик

Применимость

Мера применяется к операторам, разработчикам, администраторам и иным людям, которые имеют доступ к конвейерам обучения, реестрам моделей, средам разработки, выпуска и эксплуатации MLOps. Владелец процесса определяет периодичность пересмотра и срок отзыва по критичности полномочий и темпу кадровых изменений. Машинные идентичности заданий и авторизация отдельных действий регулируются профильными техническими мерами.

Предпосылки

- Требуются актуальный перечень человеческих и административных учётных записей MLOps, ролевой каталог с владельцами и связь кадровых событий с заявками на изменение доступа.
- Для долговечных и общих учётных записей должны быть определены запрет или ограниченные исключения с владельцем, сроком и способом атрибуции действий.

Реализация

При приёме специалист получает только права утверждённой роли на необходимые среды и ресурсы. При смене обязанностей прежний набор сначала пересматривается и отзывается, затем отдельно назначается новый; накопление прав между ролями запрещается. При завершении работ отзыв охватывает группы, прямые назначения, административные роли, локальные учётные записи, токены и аварийный доступ. Плановый пересмотр формирует список владельцу доступа и руководителю, требует явного подтверждения каждого исключения и автоматически эскалирует просроченный ответ. Долговечные и общие учётные записи заменяются персональными или краткоживущими; временное исключение имеет владельца, срок и журнал использования.

Проверка

Сверяют кадровый реестр, каталог идентичностей и фактические назначения во всех охваченных MLOps-системах. Для выборки принятых, переведённых и завершивших работу специалистов проверяют основание, соответствие роли, дату последнего пересмотра и время полного отзыва. Отдельно ищут прямые назначения, просроченные временные права, долговечные и общие учётные записи. Мера неэффективна, если прежнее полномочие действует после срока отзыва, определённого владельцем процесса для этого класса полномочий, либо действие нельзя связать с конкретным человеком.

Негативный тест

Создают контрольного пользователя, переводят его на роль с меньшими полномочиями и затем завершают работы; прежний доступ через группу, прямое назначение, токен и общую административную учётную запись должен прекратиться в срок, определённый владельцем процесса для этого класса полномочий.

Подтверждения

Сохраняются ролевой каталог, перечень учётных записей и назначений, заявки и согласования выдачи, результаты периодических пересмотров, кадровые события, журналы отзыва, исключения для временных и общих учётных записей и отчёты негативных тестов.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Новая выдача и расширение прав блокируются при отсутствии владельца, ролевого основания или согласования. Просроченное подтверждение переводит назначение в очередь обязательного отзыва; для критичного аварийного доступа допускается только ограниченное по сроку исключение с отдельной атрибуцией и последующим пересмотром.

Исправление и повторная проверка

Отзывают устаревшие прямые и групповые назначения, закрывают долговечную или общую учётную запись либо оформляют ограниченное исключение с владельцем и сроком. Затем повторяют вход и привилегированную операцию прежними полномочиями после смены обязанностей или завершения работ: доступ отклоняется во всех охваченных системах, действие атрибутируется, а актуальная минимальная роль сохраняет работоспособность.

Побочные эффекты

Ошибочный или преждевременный отзыв может остановить выпуск, обучение или реагирование, а частые пересмотры создают нагрузку на владельцев доступа.

Остаточный риск

Эффективность процесса зависит от полноты кадровых событий, каталогов и интеграций; неучтённая локальная или внешняя учётная запись может сохранить доступ, а подтверждение руководителя не доказывает минимальность ролевой модели. Действующая учётная запись всё ещё может быть скомпрометирована, а сговор владельца и пользователя — сохранить избыточный доступ.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-COMPLY-PII3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Управляет жизненным циклом и отзывом доступа людей и администраторов.

Ограничение: Не разделяет редактора правил, пользователя ключа или таблицы, оператора детокенизации и читателя свидетельств.

Условие применения: Если правила преобразования, ключи, таблицы соответствия или свидетельства доступны людям через управляемые MLOps-среды.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Выдаёт, пересматривает и отзывает доступ людей в MLOps при найме, смене обязанностей и завершении работ, включая прямые, групповые, токенные и аварийные пути.

Ограничение: Пересмотр доступа людей не охватывает продуктивные машинные идентичности, их роли и области прав.

Условие применения: Для людей и административных учётных записей, имеющих доступ к продуктивным MLOps-средам и ресурсам.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(b) · Поддерживает часть требования

Выдача, пересмотр и отзыв прав людей в MLOps ограничивают один из путей доступа к системам обработки персональных данных.

Граница связи: Мера не охватывает машинные идентичности и остальные технические границы; достаточность защиты по статье 32 зависит от риска всей обработки.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Пользователи MLOps имеют доступ к обработке персональных данных или администрируют её.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-045 · Недостатки управления учётными записями и правами в контуре MLOps	643	223
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1111	61
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1112	67
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1266	400
AISAFE2-MIT-011 · Изоляция и ограничение экспорта векторных хранилищ	1372	356
AISAFE2-MIT-034 · Контролируемое продвижение моделей через реестр	1373	448
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1374	473
AISAFE2-MIT-043 · Защита артефактов модели от несанкционированного получения	1375	481
AISAFE2-MIT-049 · Очистка и изоляция ресурсов общей среды инференса	1376	500
AISAFE2-MIT-050 · Неизменяемое версионирование и проверяемое восстановление наборов данных	1377	503
AISAFE2-MIT-052 · Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации	1378	509
AISAFE2-MIT-059 · Усиленная аутентификация привилегированных человеческих учётных записей	1379	538

Мера защиты · AISAFE2-MIT-056

Исполнение версионируемой политики допустимости содержимого

Краткое описание

Контроль применяет к входам и выходам версионируемые правила категорий содержимого и заданную безопасную реакцию до продолжения обработки.

Механизмы

- Валидация
- Управление конфигурацией
- Обнаружение

Защитный механизм

Версионируемая политика связывает маршрут, направление, категорию и действие: разрешить, ограничить, заблокировать или передать на ручное рассмотрение. Оркестратор вызывает выбранный детектор или модератор до следующей опасной границы, проверяет его ответ и применяет заранее заданное поведение при ошибке, тайм-ауте или недоступности; там, где политика требует блокирования, неопределённость не превращается в разрешение.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик

Применимость

Мера применяется, если утверждённая продуктовая, корпоративная, договорная или иная политика требует модерации конкретного входного либо выходного маршрута. Категории, пороги и действия задаются владельцем политики для конкретной области; один детектор не переносится автоматически на другой язык, формат или получателя. Мера не применяется только к маршруту, для которого политика явно не требует категориальной модерации.

Предпосылки

- Требуются утверждённые категории, охватываемые входные и выходные маршруты, действие для каждой категории и версия политики.
- Для ошибки, тайм-аута и недоступности детектора заранее задаётся безопасное поведение отдельно для каждого маршрута.

Реализация

Политика и конфигурация детектора выпускаются как управляемые версии. Вход проверяется до вызова модели или инструмента, выход — до отображения, сохранения или исполнения. Ответ детектора разбирается по строгому контракту; неизвестный статус считается ошибкой. Если правило требует блокирования, допускаются только блокировка, безопасно

ограниченный ответ или ручное рассмотрение. Журнал хранит идентификатор запроса, версию правила, категорию, решение и техническую причину без избыточного исходного содержимого.

Проверка

Для каждой категории и каждого охваченного маршрута выполняют разрешённый и запрещённый тест входа и выхода. Отдельно моделируют тайм-аут, ошибочный ответ, недоступность детектора и обход через альтернативный интерфейс. Проверяют, что блокировка входа исключает последующий вызов, блокировка выхода исключает доставку, а решение разрешённого примера связано с точной версией политики.

Негативный тест

Отключают детектор и отправляют запрещённый пример через основной и обходной интерфейсы: оба пути должны дать утверждённую безопасную реакцию без вызова или выдачи результата.

Подтверждения

Сохраняются версия политики и конфигурации, матрица категория—маршрут—действие, журналы решений без исходного содержимого, результаты позитивных, негативных и отказных тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При ошибке контроля маршрут блокируется, ограничивается или передаётся на ручное рассмотрение согласно точной версии политики; если политика требует блокирования, продолжение запрещено до восстановления или отдельного решения владельца.

Исправление и повторная проверка

Исправляют правило, контракт ответа или обходной маршрут и повторяют тот же запрещённый пример вместе с разрешённым контролем; первый блокируется до следующей границы, второй проходит по новой зафиксированной версии.

Побочные эффекты

Ложные срабатывания могут блокировать допустимое содержание и увеличивать задержку либо нагрузку на ручное рассмотрение.

Остаточный риск

Вероятностный или конечный детектор не распознаёт все смысловые эквиваленты и не заменяет авторизацию, проверку схемы, минимизацию или независимую проверку действия. Новая формулировка, язык, формат или ошибочно заданная категория могут пройти контроль, а разрешённое содержимое — оказаться опасным в неизвестном бизнес-контексте.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CONTENT1

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Версионируемая политика связывает маршрут, направление и категорию содержимого с разрешением, ограничением, блокировкой или ручным рассмотрением и задаёт реакцию на ошибку модератора.

Ограничение: Категории, пороги и действия утверждает владелец политики. Вероятностный модератор может пропустить новую формулировку или язык, отсутствующие в проверенной области.

Условие применения: На входных и выходных маршрутах, для которых утверждённая продуктовая, корпоративная, договорная или правовая политика требует модерации.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-003 · Недостаточная проверка AI-вывода на границе доверия	574	146
AISAFE2-WEAK-080 · Ограниченная полнота внешнего семантического контроля вывода	691	286
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1113	35
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1114	77
AISAFE2-MIT-004 · Выявление и обработка внедрения вредоносных инструкций и обхода ограничений	1208	328

Мера защиты · AISAFE2-MIT-057

Принудительное ограничение сетевых путей AI-компонентов

Краткое описание

Разрешающий список и запрет по умолчанию ограничивают входящие, исходящие и межкомпонентные сетевые соединения утверждёнными путями.

Область применения

- Общая безопасность

Механизмы

- Изоляция
- Авторизация
- Управление конфигурацией

Защитный механизм

Для каждого соединения задаются источник, цель, направление, назначение, идентичность нагрузки и применимые домен, адрес, протокол и порт. Сетевой, прокси- или прикладной рубеж разрешает только утверждённое сочетание и блокирует неизвестный маршрут; все доступные прямые пути к внутреннему получателю подчиняются той же либо более строгой политике.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к сетево доступным входам, исходящим вызовам и межкомпонентным соединениям AI-системы. Глубина проверки определяется наблюдаемыми атрибутами и возможностями среды; отсутствие частного канала у сервиса не доказывает невозможность прикладного разрешающего списка или закрытия иных путей.

Предпосылки

- Требуются актуальная схема фактических конечных точек и потоков, владельцы связей и стабильные атрибуты, по которым можно принудительно отличить разрешённый маршрут.
- Должны быть перечислены прямые, сервисные, административные и резервные пути к каждому защищаемому получателю.

Реализация

Правила строятся от запрета по умолчанию и минимально необходимого направления. Для внешнего API проверяются фактические DNS- и TLS-назначение и прокси-маршрут; для внутренних связей — идентичность нагрузки и поддерживаемые сетевые атрибуты. Публичный шлюз не должен оставлять прямой путь к внутреннему сервису доступнее защищённого пути. Изменение конечной точки, поставщика, сети, идентичности или протокола требует нового решения до применения. Одна только привязка сетевого объекта без проверки эффективных правил не считается результатом.

Проверка

Для каждой связи выполняют разрешённое соединение и попытки из запрещённого источника, к соседней цели, по лишнему порту или протоколу и напрямую к внутреннему сервису. Для исходящего пути проверяют неутверждённый домен или адрес и фактического TLS-получателя. Фиксируют решение рубежа и отсутствие события у запрещённой цели.

Негативный тест

Удаляют обязательный атрибут из разрешённого соединения и повторяют вызов через основной и прямой путь: оба должны блокироваться, тогда как неизменённый разрешённый маршрут остаётся работоспособным.

Подтверждения

Сохраняются версия карты соединений, эффективные правила, идентификаторы конечных точек и нагрузок, результаты разрешённых и запрещённых соединений и журнал решения рубежа.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Неизвестное соединение блокируется. При недоступности точки принуждения новый маршрут не открывается автоматически; допускается только заранее проверенный ограниченный канал.

Исправление и повторная проверка

Закрывают обходной маршрут или сужают эффективное правило, затем повторяют исходное запрещённое и разрешённое соединения; первое не достигает цели, второе проходит по единственному утверждённому пути.

Побочные эффекты

Строгий разрешающий список усложняет динамическое обнаружение сервисов и может вызвать отказ при смене адреса или внешней зависимости без синхронного обновления политики.

Остаточный риск

Сетевой контроль не подтверждает допустимость содержимого, полномочия пользователя, криптографическую защиту или безопасность разрешённого получателя. Скомпрометированная цель, общий прокси, туннелирование или неверная идентичность нагрузки могут сохранить путь злоупотребления.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Сетевая политика связывает источник и получателя с разрешёнными протоколами и портами и по умолчанию блокирует неутверждённые соединения между арендаторами и обходные пути.

Ограничение: Сетевая изоляция не разделяет идентичности, память процесса, файлы, секреты и права вычисления арендаторов.

Условие применения: Для сетевых путей арендаторов, систем или классов данных в общей среде.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Разрешённые межкомпонентные пути связывают точные источник, получателя, протокол и порт с фактической топологией; остальные соединения и неутверждённые публичные конечные точки блокируются по умолчанию.

Ограничение: Сетевое правило не реализует прикладную аутентификацию сторон. Поддерживаемые способы подключения и эффективные правила проверяют для каждого выбранного сервиса.

Условие применения: Для сетевых путей между приложением, модельным API, RAG, агентом, инструментами и журналами.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Сетевая политика разрешает вызов внешнего модельного API только от утверждённого источника к точной конечной точке получателя и по умолчанию блокирует прямой или альтернативный обход управляемой границы.

Ограничение: Сетевой контроль не определяет допустимый состав данных, тайм-ауты модельного вызова и жизненный цикл учётных данных внешнего поставщика.

Условие применения: На исходящем сетевом пути из ИИ-системы в Yandex Cloud к модельному API внешнего поставщика.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Сетевая политика с запретом по умолчанию оставляет только утверждённые пути к AI-компоненту и охватывает прямое обращение к внутреннему сервису.

Ограничение: Сетевая политика не задаёт прикладную аутентификацию клиента, максимальный размер HTTP-запроса, частотные лимиты и состав журнала публичного маршрута.

Условие применения: Для публичного входа ИИ-приложения и прямых или резервных сетевых путей к его внутреннему сервису.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 32(1)(b)** · Поддерживает часть требования
>
> Разрешающий список соединений ограничивает неутверждённые сетевые пути к AI-компонентам, обрабатывающим персональные данные.
>
> **Граница связи:** Сетевой рубеж не обеспечивает сам по себе целостность, доступность и устойчивость всех систем обработки.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-072 · Отсутствие обязательного технического контроля межзонных потоков данных	682	273
AISAFE2-WEAK-083 · Достижимость служебных интерфейсов AI-компонентов из недоверенных сетей	694	292
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1115	35
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1116	44
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1117	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1118	69
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1369	514
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1380	519

Мера защиты · AISAFE2-MIT-058

Аутентификация на внешнем входе AI-приложения

Краткое описание

Каждый неанонимный внешний маршрут проверяет учётные данные клиента до начала AI-обработки и не допускает обхода через прямое обращение к внутренней конечной точке.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Валидация

Защитный механизм

Первый доверенный вход проверяет учётные данные утверждённым аутентификатором, связывает результат с субъектом и передаёт только проверенный контекст. Все прямые и альтернативные маршруты к тому же внутреннему сервису закрыты либо выполняют эквивалентную проверку; анонимный режим существует как отдельная явно объявленная политика, а не как результат ошибки.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к публичным, партнёрским и иным внешним маршрутам, где сценарий не утверждён как анонимный. Конкретный протокол и издатель выбираются по архитектуре; аутентификация шлюза не заменяет прикладную авторизацию пользователя и ресурса. Мера не применяется к явно утверждённой анонимной функции без пользовательской идентичности; её данные, действия, лимиты и последствия должны быть ограничены отдельной политикой.

Предпосылки

- Требуются инвентарь внешних и прямых внутренних конечных точек, утверждённая схема аутентификации, доверенные издатель и получатель и ключевой материал.
- Анонимные маршруты должны быть перечислены отдельно с владельцем и ограниченной областью.

Реализация

Учётные данные проверяются до маршрутизации к модели, RAG, инструменту или бизнес-сервису: подпись или эквивалент, издатель, получатель, срок, отзыв и обязательные атрибуты. Внутренний сервис принимает идентичность только из защищённого канала входного рубежа и не доверяет клиентскому заголовку с именем. Прямая конечная точка недоступна извне либо выполняет ту же проверку. Ошибка аутентификатора, получения ключей или проверки времени не переводит маршрут в анонимный режим.

Проверка

Для каждого маршрута выполняют запросы без учётных данных, с неверной подписью, издателем или получателем, истёкшими и отозванными учётными данными, поддельным заголовком идентичности и прямой запрос к внутреннему сервису. Только действующие учётные данные проходят аутентификацию; журнал и последующая трасса содержат один проверенный идентификатор субъекта без значения секрета.

Негативный тест

Отключают аутентификатор и повторяют запрос без учётных данных через шлюз и напрямую к внутреннему сервису: оба пути блокируются, а событие модельного или инструментального вызова отсутствует.

Подтверждения

Сохраняются версия конфигурации входа, список конечных точек, параметры издателя и получателя без секретов, результаты отрицательных тестов и коррелированные записи шлюза и внутреннего сервиса.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При ошибке аутентификатора или неизвестных учётных данных запрос блокируется; автоматический переход в анонимный режим и доверие клиентскому заголовку идентичности запрещены.

Исправление и повторная проверка

Закрывают обход, исправляют доверенные параметры и повторяют исходные запросы без учётных данных и к внутреннему сервису; они отклоняются, а действующие контрольные учётные данные проходят и сохраняют идентичность.

Побочные эффекты

Недоступность издателя или ошибка синхронизации ключей может блокировать законных клиентов; кэширование ключей и времени требует безопасного отказа.

Остаточный риск

Аутентификация не подтверждает право на ресурс и сама по себе не ограничивает стоимость. Похищенные действующие учётные данные или ошибочно доверенный издатель позволяют пройти проверку; дальнейшая область ограничивается авторизацией, сроками и мониторингом.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

На маршруте с обязательной аутентификацией учётные данные проверяются до AI-обработки; прямой путь к внутреннему сервису закрыт либо выполняет ту же проверку.

Ограничение: Входная аутентификация и запрет обхода не задают лимиты, максимальный размер запроса и журнал событий публичного маршрута.

Условие применения: На публичном или партнёрском маршруте, для которого утверждена обязательная аутентификация клиента.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(b) · Поддерживает часть требования

Проверка клиента на внешнем входе AI-приложения ограничивает несанкционированный путь к обработке персональных данных.

Граница связи: Аутентификация не заменяет право на конкретный ресурс и другие меры защиты; анонимные и обходные маршруты оцениваются отдельно.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Неанонимный внешний маршрут достигает персональных данных; прямой доступ к backend закрыт.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-081 · Неаутентифицированный внешний вход в AI-приложение	692	289
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1119	35
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1120	61
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1121	72

Мера защиты · AISAFE2-MIT-059

Усиленная аутентификация привилегированных человеческих учётных записей

Краткое описание

Привилегированный вход и повышение прав требуют стойкой многофакторной проверки, соразмерной последствиям компрометации учётной записи.

Область применения

- Общая безопасность

Механизмы

- Авторизация

Защитный механизм

Политика связывает привилегированные человеческие роли и операции с обязательной многофакторной или более стойкой проверкой, контролирует доверенные способы восстановления и не позволяет обходной группе либо резервному маршруту молча ослабить требование. Для особо значимых операций применяется повторная проверка независимым фактором.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к людям с административными, управляющими, выпускными, IAM- и иными полномочиями, которые владелец IAM относит к привилегированным. Конкретные факторы и стойкость выбираются по возможностям поставщика идентификации и модели угроз; сервисные аккаунты в эту меру не входят.

Предпосылки

- Требуются полный перечень привилегированных человеческих идентичностей и путей входа, включая федеративные, локальные и аварийные учётные записи.
- Должны быть утверждены допустимые факторы, восстановление, срок сессии, операции повторной проверки и правила ограниченного исключения.

Реализация

Политика применяется в поставщике идентификации или эквивалентной доверенной точке и охватывает прямые назначения и членство в группах. Для высоких последствий предпочтителен устойчивый к фишингу фактор, если он поддерживается; это не превращается в универсальное требование конкретной технологии. Восстановление не должно быть слабее обычного входа. Аварийная учётная запись не используется ежедневно, имеет отдельного владельца, строгий фактор, мониторинг и последующий пересмотр. Повышение привилегий и чувствительные изменения требуют свежего подтверждения.

Проверка

Из инвентаря выбирают пользователя с прямой привилегией, члена группы, федеративного пользователя и аварийную учётную запись. Проверяют вход только с паролем, неверным или отсутствующим вторым фактором, через восстановление и после истечения срока свежести, зафиксированного в версии политики MFA и условного доступа. Разрешённый вход и повторная проверка должны пройти, запрещённые варианты — не получить привилегированную сессию.

Негативный тест

Для тестового привилегированного пользователя отключают второй фактор или помещают его в обходную группу и пытаются выполнить административную операцию: операция блокируется и создаёт проверяемое событие.

Подтверждения

Сохраняются список привилегированных идентичностей, версия политики MFA и условного доступа, утверждённые исключения, результаты тестов входа и повторной проверки и журналы решений без значений факторов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При недоступности требуемой проверки привилегированный вход и повышение прав блокируются; аварийный доступ возможен только по отдельному ограниченному сценарию с атрибуцией и пересмотром.

Исправление и повторная проверка

Удаляют обход, усиливают восстановление или подключают идентичность к политике, затем повторяют вход только с паролем и привилегированную операцию: они отклоняются, а штатный многофакторный путь проходит.

Побочные эффекты

Потеря фактора или отказ поставщика идентификации может задержать администрирование и реагирование; аварийный путь требует отдельной операционной готовности.

Остаточный риск

Сильная аутентификация не обеспечивает минимальность роли и не предотвращает злоупотребление легитимного администратора. Захваченная сессия, скомпрометированное устройство, принуждение или ошибочно доверенный поставщик могут обойти защиту до повторной проверки.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-082 · Недостаточная стойкость аутентификации привилегированных человеческих учётных записей	693	290
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1122	61
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1123	79
AISAFE2-MIT-055 · Управление жизненным циклом прав доступа операторов, разработчиков и администраторов MLOps	1379	525

Мера защиты · AISAFE2-MIT-060

Контролируемое продвижение данных между стадиями

Краткое описание

Версионируемый шлюз переводит данные между стадиями только по разрешённым переходам после проверки; обратная запись, прямая запись в утверждённый объект и обход шлюза запрещены.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Жизненный цикл
- Шлюз допуска
- Целостность

Защитный механизм

Реестр стадий связывает каждый объект или неизменяемую версию с состоянием, владельцем, допустимыми субъектами записи и следующим переходом. Шлюз проверяет условия перехода и создаёт новую версию в целевой стадии; идентичности предыдущей, параллельной и рабочей среды не получают запись назад, в обход проверки или в уже утверждённый объект.

Формы контроля

- Техническая
- Архитектурная
- Процессная

Ответственность

- Заказчик

Применимость

Мера применяется, если данные проходят две и более стадии подготовки, разметки, обучения, оценки, индексирования или публикации. Физическое разделение может быть объектным, версионным или инфраструктурным, если права и переходы остаются проверяемыми.

Предпосылки

- Требуются перечисленные стадии, неизменяемая идентичность объекта или версии, субъекты записи и чтения и критерии каждого допустимого перехода.
- Рабочая среда и параллельные стадии должны иметь отдельные идентичности либо эквивалентно проверяемые ограничения записи.

Реализация

Исходные, очищенные, размеченные, обучающие, оценочные и опубликованные состояния представлены отдельными объектами или неизменяемыми версиями. Каждое продвижение ссылается на исходную версию, проверки, решение и целевую версию. Изменение создаёт новую исходную версию и повторяет допуск. Рабочая среда может читать только опубликованный объект, необходимый её функции, и не может записывать в исходный, оценочный или утверждённый корпус. Обратная и параллельная запись запрещается фактической политикой IAM или приложения, а не только соглашением об именах.

Проверка

Прослеживают одну запись от источника до опубликованной версии и на каждом переходе проверяют субъекта, проверки и неизменяемые вход и выход. Пытаются записать из рабочей среды в исходную и оценочную стадии, из одной параллельной стадии в другую, опубликовать объект напрямую без решения и изменить утверждённую версию. Только настроенный переход к следующей стадии может создать новую целевую версию.

Негативный тест

От имени рабочей среды пытаются изменить утверждённый корпус и напрямую опубликовать производный объект: обе операции отклоняются, а то же изменение через предусмотренный шлюз создаёт новую прослеживаемую версию.

Подтверждения

Сохраняются модель стадий, фактические права, точные исходные и целевые версии, решения шлюза, журналы переходов и результаты разрешённых и запрещённых тестов записи.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Неизвестная стадия, отсутствующее решение или неатомарный переход блокируют продвижение; прежняя утверждённая версия остаётся активной, а частичный результат помещается в карантин.

Исправление и повторная проверка

Удаляют обходные права записи, восстанавливают последнюю утверждённую версию, признают недействительными производные результаты нарушенного пути и повторяют те же тесты обратной и прямой записи и штатное продвижение.

Побочные эффекты

Неизменяемые версии и отдельные субъекты записи увеличивают объём хранения, задержку конвейера и операционную сложность.

Остаточный риск

Контролируемый переход не подтверждает качество, правовую допустимость или отсутствие отравления данных. Скомпрометированный владелец шлюза или ошибочные критерии могут продвинуть вредную, но формально утверждённую версию, а неучтённая внешняя копия — остаться вне модели стадий.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-DATA3

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Типизированные неизменяемые стадии разрешают только утверждённое продвижение вперёд; обратные, боковые и прямые записи в продуктивную стадию или изменение объекта на месте блокируются.

Ограничение: Мера не подтверждает качество, правомерность или отсутствие отравления исходных данных; эти свойства проверяются отдельным шлюзом допуска.

Условие применения: Когда данные проходят подготовку, разметку, обучение, оценку, индексирование или публикацию с переходом между стадиями.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM04:2026 · Поддерживает часть требования

Версионируемый переход между стадиями мешает подменить уже утверждённый набор данных через обратную запись или обход шлюза.

Граница связи: Вредный набор может пройти ошибочно настроенный шлюз; управление переходами не подтверждает происхождение и качество самих данных.

Субъект	Аналитик угроз AI-приложения
Условие применения	Выпуск AI-системы зависит от модели, пакета, данных или иной поставки, которую можно скомпрометировать.

Открыть положение в источнике

LLM05:2026 · Поддерживает часть требования

Проверяемый допуск новой версии данных между стадиями сокращает пути скрытой записи в набор для обучения или извлечения.

Граница связи: Допустимый переход не доказывает отсутствие отравления; ошибочный критерий может пропустить вредную версию.

Субъект	Аналитик угроз AI-приложения
Условие применения	В обучение или рабочую обработку поступают данные, подмена которых способна изменить поведение AI-системы.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-013 · Отсутствие прослеживаемости происхождения и преобразований данных	588	171

Карточка	Связи	Стр.
AISAFE2-WEAK-014 · Неконтролируемый допуск материала в источники знаний и контуры обратной связи	592	173
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	607	184
AISAFE2-COMP-008 · Данные для подсистемы обучения	1124	52
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1125	54
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1126	86
AISAFE2-MIT-040 · Допуск и целостность обучающих данных, разметки и признаков	1345	469
AISAFE2-MIT-041 · Защита конвейера обучения и дообучения	1381	473

Мера защиты · AISAFE2-MIT-061

Непубличность хранилищ AI-системы по умолчанию

Краткое описание

Хранилища данных и артефактов по умолчанию закрыты для публичного и иного неутверждённого доступа; исключение допускается только по заранее принятому решению, ограниченному объектом, назначением, владельцем и сроком.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Изоляция
- Управление конфигурацией

Защитный механизм

Инвентарь связывает каждое хранилище и значимый объект с операциями чтения и изменения, субъектами, прикладной авторизацией, сетевым путём и решением о публичности. IAM, ACL, политика ресурса, прикладная политика и сетевая доступность проверяются совместно; анонимный или иной неизвестный субъект блокируется. Исключение ограничивается точным объектом, назначением, владельцем и сроком.

Формы контроля

- Техническая
- Архитектурная
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к базам, бакетам, индексам, реестрам, дискам, журналам, резервным копиям и иным постоянным хранилищам AI-системы. Публичное распространение допустимо только для явно выбранного объекта и цели; общая публичность родительского ресурса не наследуется молча. Мера не применяется к объекту, который по утверждённой функции является публичным и имеет точное ограниченное решение.

Предпосылки

- Требуются точные идентификаторы хранилищ и объектов, владельцы, операции, субъекты и фактические прикладные и сетевые пути.
- Должны быть известны все слои эффективного доступа: IAM, группы, ACL, политика ресурса, прикладной RBAC и публичная сетевая доступность.

Реализация

Новое хранилище создаётся непубличным и без анонимного чтения. Доступ выдаётся минимальному субъекту на минимальный объект и операцию; роли управляющей плоскости не считаются прикладной авторизацией автоматически. Проверка учитывает прямые и унаследованные назначения, ACL объекта, политику бакета или базы данных, публичную конечную точку, подписанную ссылку и обход через приложение. Публичное исключение не расширяется на всё хранилище, каталог или организацию и пересматривается при изменении данных, владельца, маршрута или назначения.

Проверка

Для каждого хранилища выполняют разрешённое чтение и изменение, затем анонимное чтение, чтение субъектом другого арендатора, операцию соседнего типа и прямой вызов конечной точки хранилища в обход приложения. Эффективную конфигурацию сопоставляют с объектом и исключением. Запрещённые операции должны завершиться отказом на фактическом пути.

Негативный тест

Добавляют тестовое анонимное или широкое разрешение на охваченный объект и проверяют обнаружение и отказ чтения; после удаления разрешения штатный субъект сохраняет доступ.

Подтверждения

Сохраняются инвентарь, эффективные IAM, ACL и политики, сетевая конфигурация, точные исключения и результаты разрешённого, анонимного, межарендаторского и обходного тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Объект с неизвестными или непроверенными фактическими эффективными правами доступа не публикуется и не подключается к рабочему пути; широкое разрешение удаляется либо ресурс изолируется до решения.

Исправление и повторная проверка

Удаляют анонимный или широкий доступ, сужают политику и закрывают обходной путь, затем повторяют те же запрещённые чтения и разрешённую контрольную операцию.

Побочные эффекты

Сужение доступа может нарушить интеграции, аналитику и аварийное восстановление; зависимости должны переходить на отдельные идентичности с минимальными правами.

Остаточный риск

После законного экспорта контроль больше не ограничивает копию. Скомпрометированный разрешённый субъект, ошибочная прикладная авторизация или неизвестная копия могут раскрыть данные даже при корректной непубличной конфигурации.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Эффективная политика доступа ограничивает хранилища и журналы разрешёнными субъектами; проверки охватывают анонимное чтение, другого арендатора и прямой обход приложения.

Ограничение: Политика хранилищ и журналов не изолирует сеть, исполняемую память и временное состояние арендаторов.

Условие применения: Для хранилищ и журналов, обслуживающих несколько систем, арендаторов или классов данных.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET5

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Эффективная политика делает каждое хранилище моделей, данных и артефактов непубличным по умолчанию, ограничивает исключение точным объектом и сроком и отрицательным тестом подтверждает отказ неаутентифицированного пути.

Ограничение: Непубличность не определяет режим шифрования хранилища и не фиксирует его связь с AI-CRYPT1. Разрешённый экспорт создаёт копию за пределами этой границы.

Условие применения: Для бакетов, баз, индексов, дисков, реестров, журналов и резервных копий ИИ-системы в Yandex Cloud.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(b) · Поддерживает часть требования

Закрытие хранилища по умолчанию снижает риск несанкционированного чтения персональных данных в AI-системе.

Граница связи: Непубличность не исключает неправомерный разрешённый доступ, экспорт и сбои доступности; статья 32 требует оценки всех существенных путей.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Персональные данные находятся в бакете, индексе, базе, резервной копии или журнале; для каждого такого объекта определены и проверены фактические права доступа.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-022 · Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации	608	184
AISAFE2-COMP-007 · Долговременная память	1127	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	1128	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1129	56
AISAFE2-COMP-013 · Телеметрия и обнаружение	1130	64
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1131	75
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1132	86
AISAFE2-COMP-024 · Артефакты работы AI-системы	1133	88

Мера защиты · AISAFE2-MIT-062

Шифрование данных и артефактов при хранении

Краткое описание

Каждая управляемая постоянная, резервная, временная и экспортная копия использует подтверждённый режим шифрования выбранного хранилища.

Область применения

- Общая безопасность

Механизмы

- Управление конфигурацией
- Авторизация
- Жизненный цикл

Защитный механизм

Инвентарь связывает копию с поддерживаемым режимом шифрования, ключом при его использовании, субъектами применения ключа и планом восстановления. Создание или подключение объекта проверяет фактическую конфигурацию. Если требуется пользовательский ключ, доступ, ротация, блокирование и удаление управляются так, чтобы незашифрованная или необратимо потерянная копия не стала резервным путём.

Формы контроля

- Техническая
- Архитектурная
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к данным, индексам, артефактам, дискам, снимкам, резервным копиям, временным и экспортным копиям в пределах управляемой области. Конкретный алгоритм и модель ключей выбираются по документированным возможностям сервиса и утверждённой криптографической политике.

Предпосылки

- Требуются инвентарь всех копий, точные ресурсы хранения, поддерживаемые режимы и утверждённая политика выбора ключа.
- Для пользовательского ключа нужны отдельные владельцы, минимальные права использования и администрирования, план ротации и восстановления и известные последствия блокирования или удаления.

Реализация

Шифрование проверяется при создании, копировании, резервировании и экспорте, а не только у основного объекта. Системный ключ поставщика допустим, если он соответствует утверждённому решению; пользовательский ключ не заявляется для неподдерживаемого сервиса. Значение ключа не включается в конфигурацию, журнал или свидетельства. Если используется пользовательский ключ, его изменение выполняется с проверяемым восстановлением; последний пригодный ключ не блокируется и не удаляется без отдельного безопасного плана.

Проверка

Для каждой категории копии сопоставляют эффективную конфигурацию шифрования и модель ключа. Штатная нагрузка читает контрольный объект, субъект без права на объект получает отказ; при использовании пользовательского ключа отдельно проверяется отказ

субъекту без права на этот ключ. При пользовательском ключе на изолированном тестовом ресурсе репетируют его смену и восстановление; при системном ключе фиксируют доступную документацию поставщика о его жизненном цикле и остаточный риск. Незашифрованная копия, снимок или экспорт должны быть обнаружены и заблокированы.

Негативный тест

Создают тестовую копию без требуемого режима и пытаются подключить её к рабочему пути: подключение блокируется. Если используется пользовательский ключ, отдельно проверяют отказ в чтении зашифрованного объекта субъекту без права на этот ключ.

Подтверждения

Сохраняются инвентарь копий, конфигурация режима, результаты разрешённого чтения и отказа; для пользовательского ключа — идентификатор без секретного материала, эффективные права и результаты репетиции смены и восстановления, для системного ключа — доступная документация поставщика о модели управления.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Копия с неизвестным или нарушенным режимом не подключается и помещается в карантин. При недоступности ключа система не создаёт незашифрованный резервный вариант и следует проверенному пути восстановления.

Исправление и повторная проверка

Шифруют либо безопасно заменяют незащищённую копию; при пользовательском ключе исправляют его права или восстановление. Повторяют исходное подключение и чтение от разрешённого и запрещённого субъектов.

Побочные эффекты

Управление собственным ключом повышает риск недоступности и необратимой потери, усложняет резервное копирование и восстановление и может ограничить переносимость сервиса.

Остаточный риск

Шифрование при хранении не ограничивает законное чтение и не защищает открытые данные в памяти или на выходе. Скомпрометированная разрешённая нагрузка, оператор с правом расшифрования или неучтённая копия могут раскрыть данные.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Каждая постоянная, временная, резервная и экспортная копия шифруется при хранении, а неизвестная или непроверяемая копия изолируется до восстановления её защиты.

Ограничение: Шифрование при хранении не защищает передачу и открытые данные после разрешённого расшифрования; транспорт проверяют отдельно.

Условие применения: Для постоянных, временных, резервных и экспортных копий данных или артефактов; пользовательский ключ — только при поддержке выбранным сервисом.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-023 · Незашифрованное хранение	611	186
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	705	299
AISAFE2-COMP-007 · Долговременная память	1134	50
AISAFE2-COMP-008 · Данные для подсистемы обучения	1135	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1136	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1137	58
AISAFE2-COMP-013 · Телеметрия и обнаружение	1138	64
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1139	75
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1140	86
AISAFE2-COMP-024 · Артефакты работы AI-системы	1141	88
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1388	571

Мера защиты · AISAFE2-MIT-063

Защищённая передача данных и артефактов между компонентами

Краткое описание

Все охваченные внешние и межкомпонентные каналы используют утверждённый защищённый протокол с проверкой подлинности получателя и запретом перехода на менее защищённый режим связи.

Область применения

- Общая безопасность

Механизмы

- Целостность
- Авторизация
- Управление конфигурацией

Защитный механизм

Инвентарь каналов связывает источник, получателя, протокол, способ установления доверия и допустимые параметры. Клиент и сервер требуют защищённый режим, проверяют идентичность удалённой стороны или конечной точки и не переходят на открытый либо

неподтверждённый транспорт при ошибке. Прокси, очереди, служебные и резервные пути включаются в ту же область.

Формы контроля

- Техническая
- Архитектурная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к сетевым, брокерским и интеграционным каналам, по которым AI-система передаёт данные, промпты, ответы, метаданные учётных данных или артефакты. Конкретные протоколы и параметры задаются актуальной утверждённой политикой для среды. Мера может не применяться к передаче, которая доказуемо не покидает одну защищённую границу и не использует отдельный канал; предположение о внутреннем доверии недостаточно.

Предпосылки

- Требуются инвентарь внешних, внутренних, служебных и резервных каналов и утверждённые способы проверки удалённой стороны для каждого.
- Должны быть управляемые якоря доверия, сертификаты или эквивалентные учётные данные и процесс их безопасной смены и отзыва.

Реализация

Защищённый транспорт обязателен на каждом участке; завершение защиты на прокси не доказывает защиту следующего участка. Клиент проверяет имя узла, сервисную идентичность или иной утверждённый атрибут получателя и не отключает проверку сертификата или удалённой стороны. Перенаправление, резервный режим, путь проверки работоспособности, соединение с очередью и внутренний административный API не остаются открытыми. Секреты и чувствительное содержимое не помещаются в диагностику установления соединения.

Проверка

Для каждого канала выполняют штатную передачу и тесты открытого протокола, неверной идентичности получателя, недоверенного или истёкшего сертификата, понижения защиты, прямого обходного участка и отказа сервиса доверия. Запрещённый вариант не должен создать событие получения содержимого у цели.

Негативный тест

Подменяют сертификат или идентичность получателя и предлагают открытый резервный режим: соединение прекращается до отправки, тогда как неизменённый защищённый канал проходит.

Подтверждения

Сохраняются карта каналов, эффективная конфигурация протокола, метаданные доверия без закрытых ключей, результаты положительного и отрицательных тестов соединения и подтверждение отсутствия доставки по запрещённому пути.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При невозможности подтвердить защищённый канал передача блокируется; автоматический открытый резервный режим и отключение проверки удалённой стороны запрещены.

Исправление и повторная проверка

Исправляют цепочку доверия, идентичность конечной точки или открытый участок и повторяют тесты защищённого канала, неверной удалённой стороны и понижения защиты; только утверждённый канал передаёт объект.

Побочные эффекты

Ротация и ошибки доверия могут остановить интеграцию; управление сертификатами и сервисными идентичностями повышает эксплуатационную сложность.

Остаточный риск

Защищённый транспорт не авторизует бизнес-операцию и не минимизирует передаваемые данные. Скомпрометированная доверенная конечная точка, издатель или открытые данные после завершения защиты могут раскрыть сведения.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Аутентифицированный транспорт защищает конфиденциальность и целостность данных между точными сторонами и блокирует незашифрованный, пониженный или направленный другому получателю маршрут.

Ограничение: Защищённая передача не покрывает хранимые копии и управление ключами их шифрования.

Условие применения: Для внешних и межкомпонентных каналов передачи данных, индексов или модельных артефактов.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Защищённый протокол аутентифицирует удалённую конечную точку, сохраняет конфиденциальность и целостность вызова внешнего модельного API и блокирует незашифрованный либо пониженный маршрут.

Ограничение: Версию TLS, доверенные сертификаты и другие параметры задают для выбранного API. Защищённый канал не разрешает передачу данных и не авторизует операцию.

Условие применения: При передаче запросов и ответов между ИИ-системой в Yandex Cloud и внешним модельным API.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET7

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Защищённый внешний транспорт аутентифицирует удалённую сторону, предотвращает раскрытие и изменение сообщения на пути и отклоняет незашифрованный или пониженный вариант соединения.

Ограничение: TLS не проверяет право клиента на действие, не задаёт лимиты и не закрывает прямой путь к внутреннему сервису.

Условие применения: На внешнем транспортном канале публичного или партнёрского ИИ-приложения.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

> **Article 32(1)(b)** · Поддерживает часть требования
>
> Защищённый канал между AI-компонентами поддерживает конфиденциальность и целостность персональных данных при передаче.
>
> **Граница связи:** Шифрование канала не защищает хранимые копии и права конечного получателя и не подтверждает устойчивость всей обработки.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Канал передаёт персональные данные; утверждённый протокол шифрует их и обеспечивает целостность на каждом участке, а подлинность получателя проверяется до передачи.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-024 · Незашифрованные каналы передачи	613	188
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	706	299
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1142	35
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1143	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	1144	48
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1145	58
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1146	69
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1147	75
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1148	82

Карточка	Связи	Стр.
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1149	86
AISAFE2-MIT-069 · Наследование класса чувствительности векторными представлениями	1389	571

Мера защиты · AISAFE2-MIT-064

Управляемый жизненный цикл секретов AI-системы

Краткое описание

Каждый секрет хранится, доставляется, пересматривается, заменяется и отзывается по проверяемой записи владельца, назначения и разрешённых потребителей.

Область применения

- Общая безопасность

Механизмы

- Авторизация
- Жизненный цикл
- Управление конфигурацией
- Минимизация данных

Защитный механизм

Реестр связывает секрет и его версии с владельцем, назначением, потребителями, минимальной ролью и областью, способом доставки, сроком или периодом пересмотра, заменой и отзывом. Значение находится в специализированном хранилище или эквивалентной защищённой системе и не попадает в исходный код, историю сборки, образ, набор данных, системный промпт, открытую конфигурацию, журнал или свидетельства. Доставка раскрывает значение только утверждённой нагрузке на требуемое время.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к API-ключам, токенам поставщика, паролям, материалу подписи и другим секретным значениям, которые использует AI-система. Если архитектура применяет только краткоживущие токены идентичности без хранимого секрета, соответствующая ветвь отмечается неприменимой со ссылкой на контроль идентичности.

Предпосылки

- Требуются инвентарь секретов и версий, владельцы, назначения, потребители, точные роли и области действия, а также пути хранения и доставки.
- Для каждого секрета нужна процедура замены и отзыва, учитывающая кэширование у потребителей и непрерывность сервиса.

Реализация

Долгоживущий секрет по возможности заменяют ограниченными краткоживущими учётными данными. Иначе среда выполнения получает ссылку или извлекает значение из утверждённого хранилища под своей машинной идентичностью; значение не встраивается в код, параметры сборки, образ, пользовательские данные, набор данных, системный промпт или обычную конфигурацию среды. Журналы и свидетельства сохраняют только ID и версию секрета и решение. При ротации создают новую версию, обновляют и проверяют потребителей, затем отзывают старую; аварийный отзыв не восстанавливает молча общий резервный ключ.

Проверка

Каждую активную версию сопоставляют с потребителями, доставкой и эффективными правами. В исходном коде, истории сборки, артефактах, образах, параметрах развёртывания, наборах данных, промптах, конфигурации, журналах и свидетельствах ищут уникальный синтетический маркер. Тестовую версию отзывают, после документированного истечения кэша проверяют отказ, затем заменяют её без раскрытия значения. Неавторизованная нагрузка не должна получить секрет.

Негативный тест

Синтетический секрет помещают в неутверждённую конфигурацию, затем пытаются получить его из соседней нагрузки и использовать после отзыва: выпуск или сканирование обнаруживает копию, получение отклоняется, а старая версия перестаёт работать.

Подтверждения

Сохраняются ID и версии секретов без значений, владельцы и потребители, эффективные права IAM и ключей, конфигурация доставки, результаты сканирования, события ротации и отзыва и тесты замены.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Неизвестный, раскрытый или просроченный секрет изолируется и отзывается; затронутое действие блокируется до проверенной замены. Система не переходит на общие открытые учётные данные.

Исправление и повторная проверка

Удаляют запрещённые копии, выпускают новую ограниченную версию, обновляют проверенных потребителей и отзывают старую. Повторяют тот же поиск, неавторизованное получение и вызов после отзыва: они завершаются отказом, а утверждённый потребитель работает.

Побочные эффекты

Ротация и отзыв могут прервать сервис; недоступность хранилища и задержка доставки требуют проверенного ограниченного поведения без резервного открытого значения.

Остаточный риск

Хранилище секретов не заменяет минимальные права машинной идентичности и не предотвращает злоупотребление авторизованного потребителя; внешний поставщик может продолжить принимать закэшированное значение после отзыва. Секрет также может раскрыться в памяти, сторонней системе или непросканированной копии.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-CRYPT1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Ключевой материал и связанные секреты имеют владельца, назначение, потребителей, контролируемые хранение и доставку, срок действия или период пересмотра, замену и проверяемый отзыв.

Ограничение: Жизненный цикл ключей и секретов не выбирает и не проверяет режим шифрования данных при хранении и передаче.

Условие применения: Для ключевого материала и секретов, используемых выбранной схемой шифрования ИИ-системы.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-IAM4

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

API-ключ связан с владельцем и потребителями, хранится и доставляется контролируемо, имеет явную область действия, срок или период пересмотра и процедуру замены и отзыва.

Ограничение: Жизненный цикл API-ключа не выбирает документированную схему аутентификации и нужную роль конкретного AI Studio API.

Условие применения: Если клиент AI Studio использует API-ключ сервисного аккаунта или другой хранимый секрет.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-INFRA6

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Секрет связан с владельцем, назначением и разрешёнными потребителями; чужая нагрузка не должна получить его, а замена и отзыв проверяются на фактических потребителях.

Ограничение: Управление секретами не изолирует сеть, память, файлы и остальные данные совместной среды.

Условие применения: Для секретов нагрузок, работающих в общей среде нескольких систем или арендаторов.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-NET3

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Секрет внешнего поставщика связан с владельцем, назначением и потребителями, доставляется из утверждённого хранилища и имеет ограниченную область, срок действия или период пересмотра, порядок замены и отзыва.

Ограничение: Область и отзыв проверяются в системе внешнего поставщика. Удаление локального значения не доказывает, что поставщик перестал принимать прежний ключ.

Условие применения: Если внешний модельный API использует хранимый ключ, токен или другой секрет организации.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-SECRET1

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Реестр связывает секрет и версию с владельцем, назначением, потребителями, ролью и областью, доставкой, сроком или периодом пересмотра, заменой и отзывом. Значение исключается из кода, истории сборки, образов, наборов данных, промптов, открытой конфигурации, журналов и свидетельств.

Ограничение: Разрешённый потребитель может злоупотребить полученным секретом. Место хранения, способ доставки и проверку отзыва выбирают для фактической среды.

Условие применения: Для хранимых API-ключей, токенов поставщиков, паролей, материала подписи и иных секретных значений ИИ-системы.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(b) · Поддерживает часть требования

Управляемая выдача и отзыв секретов ограничивают доступ к AI-компонентам, обрабатывающим персональные данные.

Граница связи: Мера не защищает от злоупотребления уже разрешённым потребителем и не подтверждает доступность и устойчивость всех систем обработки.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Секрет даёт доступ к системам обработки персональных данных, их расшифрованию или передаче.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-025 · Недостаточное управление ключами и секретами	616	190
AISAFE2-WEAK-041 · Недостаточный контроль доступа к системным инструкциям и секретам	635	216
AISAFE2-WEAK-094 · Небезопасное хранение и распространение делегированных учётных данных пользователя	716	312
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1150	44
AISAFE2-COMP-009 · Конвейер обучения и дообучения	1151	54
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1152	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1153	61
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1154	69
AISAFE2-COMP-020 · Контур управления внешним AI-провайдером	1155	79
AISAFE2-MIT-018 · Управление машинными идентичностями и минимальными полномочиями	1252	383
AISAFE2-MIT-015 · Защита конфигурации системных промптов и инструкций инструментов	1382	371

Мера защиты · AISAFE2-MIT-065

Защищённое сохранение материалов расследования

Краткое описание

Отобранные материалы расследования хранятся как минимальный изолированный набор с контролируемым доступом, сроком и проверяемой целостностью.

Область применения

- Общая безопасность

Механизмы

- Целостность
- Наблюдаемость
- Жизненный цикл
- Авторизация

Защитный механизм

Для каждого дела утверждаются цель, минимальный состав, владелец, читатели, место и срок. Материалы отделяются от рабочих журналов и связываются с делом, источником, временем, версией и хешем либо манифестом. Права записи, чтения, экспорта, удаления и администрирования разделяются; доступ регистрируется. Версионирование, неизменяемое хранение или эквивалентная защита обнаруживают замену и не позволяют рабочему администратору единолично изменить или удалить единственную копию.

Формы контроля

- Техническая
- Процессная
- Организационная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к журналам, трассам, выгрузкам, частям диалога и иным объектам, отобранным для расследования, поддержки, контроля качества или другой утверждённой цели. Состав и срок задаёт внешнее решение для категории данных и дела. Если хранится только минимальная телеметрия без содержимого, ветвь необработанного содержимого неприменима, но сохранённые события всё равно защищаются. Неприменимость допустима только для конкретного вида материала, который по утверждённому решению не создаётся и не хранится.

Предпосылки

- Требуются утверждённые цель, минимальный состав, владелец, круг читателей, место, срок и правило удаления для каждого вида материала.
- Должны быть определены идентификатор дела, источник, временная метка, точная версия объекта и способ проверки целостности.
- Полномочия записи, чтения, экспорта, удаления и администрирования должны быть различимы и проверяемы.

Реализация

Создают отдельное хранилище и связывают объект с делом. По умолчанию сохраняют идентификаторы, версии, решения и технические исходы; содержимое — только по явному решению и после минимизации. Источник, время, версия и хеш фиксируются у объекта или в отдельно защищённом манифесте. Писатель не получает чтение и удаление, расследователь не администрирует защиту, администратор исследуемой системы не управляет единственной копией. Включают версионирование, блокировку изменения или эквивалент; регистрируют чтение, экспорт и изменение политики, изолируют дела. По истечении срока объект удаляют, если нет утверждённого удержания.

Проверка

Выбирают синтетический объект и прослеживают его от исходного события до точной версии в архиве, сверяя хеш или манифест. Выполняют разрешённое чтение, чтение расследователем другого дела, изменение и удаление рабочим администратором, экспорт и штатное удаление после окончания тестового срока. Проверяют фактические полномочия, события доступа, сохранность единственной копии и поведение при действующем удержании.

Негативный тест

Рабочий администратор пытается изменить и удалить единственную защищённую копию, а расследователь другого дела — прочитать и экспортировать её: операции блокируются либо изменение однозначно обнаруживается без утраты исходной версии.

Подтверждения

Сохраняются решение о составе и сроке, идентификаторы дела и объектов, версии и хеши, фактические полномочия, настройки защиты, события чтения и экспорта и результаты проверок изменения, удаления и изоляции.

Поведение при отказе

Блокировка процессом

Сценарий отказа

При неизвестной целостности, правах, сроке или местонахождении материала расследование нельзя закрывать и нельзя уничтожать доступные исходные объекты до восстановления проверяемого состояния. Неотложная локализация опасного пути при этом не откладывается.

Исправление и повторная проверка

Восстанавливают проверенную версию, исправляют полномочия, изоляцию, срок или регистрацию доступа и повторяют те же операции чтения, изменения, экспорта и удаления. Решение о закрытии принимают только после подтверждения целостности и полноты требуемого набора.

Побочные эффекты

Неизменяемость и удержание увеличивают стоимость и могут препятствовать своевременному удалению; ошибочная область дела расширяет доступ, а чрезмерный состав повышает риск раскрытия.

Остаточный риск

Мера не определяет законность сбора, не восстанавливает отсутствующие события и защищает только учтённые объекты и доверенные метаданные источника. Скомпрометированный источник может записать ложные данные до архивирования, неучтённая копия — остаться вне защиты, а сговор разделённых ролей — обойти организационную границу.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-FORENSIC2

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Материалы изолируются по делу; чтение и экспорт доступны только назначенным участникам, протоколируются, ограничиваются сроком хранения и завершаются контролируемым удалением.

Ограничение: Карточка не задаёт шифрование архива и правовое основание отбора содержимого. Эти условия утверждают и проверяют отдельно; минимальная телеметрия не требует сбора полного диалога.

Условие применения: Если содержимое диалога полностью или частично сохраняется для конкретного расследования, поддержки, контроля качества или другой утверждённой цели.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-FORENSIC3

Редакция: Редакция 7, версия от 17.08.2026

Содержательное совпадение

Отделяет материалы от рабочих администраторов, разделяет запись, чтение и удаление, связывает объект с источником, временем и точной версией и проверяет его целостность по хешу или манифесту.

Ограничение: Мера не доказывает полноту или правомерность исходных материалов и не восстанавливает события, которые не были зафиксированы до сохранения.

Условие применения: Для журналов и выгрузок, отобранных как материалы расследования.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-048 · Недостаточное реагирование на инциденты и аварийное восстановление	650	229
AISAFE2-COMP-013 · Телеметрия и обнаружение	1156	64
AISAFE2-COMP-018 · Операционные данные, очереди и кэши	1157	75
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1267	400
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1383	410

Мера защиты · AISAFE2-MIT-066

Версионирование доверенной цели агента и повторная авторизация

Краткое описание

Агент исполняет типизированную версию цели и разрешённых возможностей из проверенного серверного источника, а любое изменение создаёт новую версию и проходит новую авторизацию.

Механизмы

- Авторизация
- Управление конфигурацией
- Целостность

Защитный механизм

Доверенный серверный объект связывает инициатора, владельца, цель, версии цели и политики, разрешённые возможности, ограничения и делегирование. Пользовательский текст, RAG, результат инструмента или другого агента остаются данными и не изменяют объект. Изменение выполняется отдельной операцией: создаёт неизменяемую версию, проходит повторную авторизацию и требуемое подтверждение; текущий цикл не продолжает работу с неявно расширенной целью.

Формы контроля

- Техническая

- Архитектурная

Ответственность

- Заказчик
- Поставщик

Применимость

Мера применяется к многошаговому агенту, где оркестратор хранит цель, меняет подцели, делегирует работу или выбирает внешние действия. Доверенный источник может находиться в приложении, агентной платформе или контуре политик, если проверяются его идентичность, полномочия и версия. Человеческое подтверждение зависит от риска, повторная авторизация изменения обязательна.

Предпосылки

- Требуются аутентифицированный серверный источник цели, владелец сценария и схема типизированного объекта цели, политики, ограничений и разрешённых возможностей.
- Должны быть определены полномочия на создание и изменение цели, правило конфликта с недоверенными данными и условия обязательного подтверждения.
- Каждое внешнее действие должно иметь независимую точку авторизации, способную использовать точную версию цели и делегирования.

Реализация

Оркестратор создаёт объект цели до цикла и передаёт модели ссылку или ограниченное представление без права изменения. Сборщик контекста маркирует пользователя, RAG, инструменты и других агентов как недоверенные источники; предложенная ими подцель остаётся данными до серверного перехода. Переход проверяет инициатора, возможности и ограничения, создаёт версию и при необходимости запрашивает подтверждение; старая версия сохраняется. Каждый шаг и решение авторизации связаны с точной версией. Неизвестная версия, конфликт доверия, потеря владельца или прямое изменение останавливают внешние действия.

Проверка

Запускают допустимую цель и проверяют версию каждого шага. Затем через пользователя, RAG, инструмент и другого агента по отдельности предлагают расширить цель, сменить владельца, добавить возможность или снять ограничение; ни один канал не создаёт полномочие или действие. Разрешённое изменение через доверенный путь должно создать новую версию, пройти повторную авторизацию и требуемое подтверждение; старое делегирование перестаёт действовать.

Негативный тест

RAG-фрагмент и ответ инструмента требуют добавить запрещённую возможность и продолжить без подтверждения: агент не получает полномочие, вызов целевой системы отсутствует, а попытка и действующая версия цели регистрируются.

Подтверждения

Сохраняются схема и версии цели и политики, инициатор и владелец, разрешённые возможности без секретов, решения авторизации и подтверждения, привязка шагов и результаты четырёх проверок недоверенных источников.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

При неизвестной версии, неавторизованном изменении, конфликте доверия или недоступности проверки агент прекращает новые внешние действия, сохраняет причину и при необходимости отзывает делегирование. Продолжение с расширенной целью по умолчанию запрещено.

Исправление и повторная проверка

Исправляют источник, разбор типов, полномочия изменения или правило перехода, создают новую проверенную версию и повторяют исходную попытку тем же каналом. Недоверенное расширение остаётся данными, разрешённый переход проходит новую авторизацию, а старое делегирование не действует.

Побочные эффекты

Повторная авторизация и подтверждение увеличивают задержку и могут остановить длительную задачу; строгая неизменяемость требует явной модели подцелей, миграции версии и восстановления после сбоя.

Остаточный риск

Мера не доказывает правильность плана и бизнес-инварианты и не заменяет авторизацию ресурса, операции и получателя непосредственно перед эффектом. Авторизованный владелец или скомпрометированный доверенный источник всё ещё могут выпустить опасную либо неоднозначную цель.

Сопоставления с внешними источниками

Сопоставление указывает на тематическую связь и само по себе не подтверждает соответствие требованиям.

Стандарт безопасности ИИ-систем в Yandex Cloud · AI-AGENT-GOAL1

Редакция: Редакция 7, версия от 17.08.2026

Частичное пересечение

Агент исполняет только типизированную версионируемую цель от доверенного серверного источника, связанную с инициатором и разрешёнными возможностями; любое изменение цели требует повторной авторизации до внешнего эффекта.

Ограничение: В этой редакции Стандарта любое изменение цели требует новой авторизации и подтверждения. В карточке человеческое подтверждение зависит от риска; порядок подтверждения каждого изменения нужно определить дополнительно.

Условие применения: Для многошагового агента с серверной целью, подцелями, делегированием или внешними действиями.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

> **LLM03:2026** · Поддерживает часть требования
>
> Версионированная цель и повторная авторизация её изменения не дают недоверенному тексту незаметно расширить полномочия текущего агентного запуска.
>
> **Граница связи:** Разрешённая цель тоже может привести к опасному плану; права на конкретный ресурс и действие проверяются отдельно.
>
> | | |
> |---|---|
> | **Субъект** | Аналитик угроз AI-приложения |
> | **Условие применения** | AI-агент получает инструменты или полномочия, достаточные для действия вне границ конкретной задачи. |
>
> Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-007 · Отсутствие границы доверия между инструкциями и данными	579	154
AISAFE2-WEAK-055 · Возложение бизнес-логики и обязательных ограничений на модель	661	245
AISAFE2-COMP-003 · Управление контекстом общения с моделью	1158	40
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1159	42
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1160	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1161	77
AISAFE2-MIT-003 · Разделение доверенных инструкций и недоверенного контента при сборке контекста	1202	324
AISAFE2-MIT-029 · Независимая авторизация действий агента перед внешним эффектом	1384	429

Мера защиты · AISAFE2-MIT-067

Эталон защищённой конфигурации и контроль отклонений

Краткое описание

Утверждённый эталон задаёт защищённые настройки для точных версии и среды компонента, а проверка при развёртывании и в эксплуатации блокирует неизвестные отклонения.

Область применения

- Общая безопасность

Механизмы

- Управление конфигурацией
- Шлюз допуска
- Жизненный цикл

Защитный механизм

Для типа компонента, версии и среды хранится утверждённый машиночитаемый эталон обязательных настроек, владельца и допустимых исключений. Развёртывание и периодическая проверка сравнивают с ним эффективную конфигурацию; неизвестное, просроченное или критичное отклонение блокирует выпуск либо изолирует затронутый экземпляр до решения.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Применяется ко всем конфигурируемым AI-компонентам, средам и управляемым сервисам в объёме наблюдаемых настроек. Для поставляемого сервиса допустимо проверяемое подтверждение доступных свойств, но невозможность увидеть эффективную настройку является ограничением, а не доказательством соответствия. Мера дополняет, но не заменяет управление уязвимостями и модель угроз.

Предпосылки

- Нужны инвентарь точных версий и сред, владельцы настроек и способ получить эффективную, а не только желаемую конфигурацию.
- Исключение должно иметь область, владельца риска, срок и обязательную повторную проверку.

Реализация

Эталон задаёт как минимум режим аутентификации, административные и отладочные функции, привилегии процесса, сетевые привязки, доступные плагины, экспорт и применимые безопасные значения. Профиль версионируется вместе с поддерживаемой версией компонента. Проверка выполняется до продвижения и повторяется после изменения или с периодичностью, определённой владельцем компонента. Сравнивается фактическая конфигурация всех штатных и обходных экземпляров. Допустимое отклонение не становится новым значением по умолчанию без отдельного пересмотра эталона.

Проверка

Для каждого компонента получают эффективную конфигурацию и сопоставляют её с точной редакцией эталона. Проверяют контрольный соответствующий экземпляр, небезопасное значение, неизвестное поле, просроченное исключение, изменение после выпуска и прямой административный путь. Мера неэффективна, если критичное или неизвестное отклонение достигает либо сохраняет рабочее состояние без блокировки, изоляции и решения владельца.

Негативный тест

Меняют одну обязательную настройку на небезопасную, добавляют неизвестное отклонение и просрочивают исключение, затем повторяют через штатный и прямой путь развёртывания. Контрольная конфигурация проходит, остальные не получают разрешение на работу.

Подтверждения

Редакции эталонов, связь с версиями и средами, эффективные конфигурации, результаты сравнения, решения по исключениям, события блокировки или изоляции и отчёты повторной проверки.

Поведение при отказе

Блокировка процессом

Сценарий отказа

Непроверенная или критично отклонённая конфигурация не продвигается. Выявленный после выпуска экземпляр изолируется либо переводится на последнюю подтверждённую конфигурацию; недоступность проверки не трактуется как соответствие.

Исправление и повторная проверка

Возвращают подтверждённое значение либо оформляют ограниченное исключение, устраняют обход и повторно получают эффективную конфигурацию. Исходное отклонение должно исчезнуть, контрольный экземпляр остаться рабочим, а новый результат быть связан с точными версией и средой.

Побочные эффекты

Строгая проверка замедляет срочные изменения и может изолировать рабочий экземпляр после несовместимого обновления; исключения требуют операционного сопровождения.

Остаточный риск

Эталон может быть неполным или устареть, а поставщик — скрывать эффективные значения; соответствие конфигурации не доказывает отсутствие дефекта кода, архитектурной ошибки или злоупотребления разрешённой функцией. Одобренная настройка, неучтённый экземпляр или скрытый параметр могут остаться небезопасными.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(b) · Поддерживает часть требования

Эталон и проверка отклонений помогают сохранять защитные настройки компонентов, через которые проходят персональные данные.

Граница связи: Наблюдаемая конфигурация не доказывает эффективность всех мер конфиденциальности, целостности, доступности и устойчивости.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Профиль управляет конфигурацией компонентов, обрабатывающих персональные данные; отклонения проверяются.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-084 · Отсутствие утверждённого эталона защищённой конфигурации AI-компонентов	697	294
AISAFE2-COMP-002 · Модель и интерфейс инференса	1162	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1163	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1164	44
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1165	56
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1166	58
AISAFE2-COMP-013 · Телеметрия и обнаружение	1167	64
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1168	67
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1169	86
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1385	415
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1395	573

Мера защиты · AISAFE2-MIT-068

Изолированные неизменяемые версии критичных AI-активов

Краткое описание

Критичный AI-актив имеет идентифицируемую неизменяемую версию и независимую копию вне рабочих полномочий, а восстановление требуемой версии регулярно проверяется.

Область применения

- Общая безопасность

Механизмы

- Восстановление
- Целостность
- Изоляция

Защитный механизм

Реестр связывает каждый критичный набор данных, артефакт модели, корпус или индекс базы знаний и критичную конфигурацию с неизменяемым идентификатором, манифестом и независимой копией в отдельной границе полномочий и отказа. Восстановление выбирает требуемую версию, проверяет состав и только затем возвращает её в рабочий контур.

Формы контроля

- Техническая
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Применяется к активам, утрата которых делает невозможными проверенное восстановление, воспроизведение или безопасную эксплуатацию: данным, весам и артефактам модели, базе знаний и индексам, системным инструкциям и критичной конфигурации. Глубина и срок определяются критичностью и допустимыми RPO/RTO. Мера не распространяется автоматически на временные очереди, кэши и воспроизводимое низкорисковое состояние. Низкорисковый объект исключают из области меры по документированному решению владельца только после оценки критичности и успешного воспроизведения требуемой версии из неизменяемых входов по проверяемому рецепту; требования к RPO/RTO и хранению при этом продолжают действовать.

Предпосылки

- Требуются реестр критичных активов, стабильные идентификаторы версий, манифесты и независимые полномочия на удаление и восстановление.
- Должны быть определены требуемые версия, RPO/RTO и безопасный путь контролируемого возврата в рабочий контур.

Реализация

Для каждой допущенной версии создают манифест и контрольные суммы. Как минимум одна копия защищена от изменения и удаления обычной рабочей или административной идентичностью и не разделяет с основным экземпляром единственную границу отказа. Удержание учитывает законное удаление и жизненный цикл. Учение блокирует основной домен и рабочие полномочия, восстанавливает не обязательно последнюю, а заданную версию в изолированную среду, проверяет манифест и пригодность, затем проводит контролируемое продвижение.

Проверка

Для каждого класса критичных активов удаляют или блокируют рабочий экземпляр и основной домен полномочий, затем восстанавливают заданную рабочую, но не обязательно последнюю версию. Проверяют манифест, контрольные суммы, доступность через заявленные RPO/RTO и невозможность одной рабочей идентичностью изменить все копии. Мера неэффективна, если нужная версия отсутствует, копия разделяет полномочия удаления или возврат не проверен.

Негативный тест

Рабочей идентичностью удаляют или портят основной экземпляр и пытаются изменить все копии, блокируют основной домен и запрашивают более раннюю точную версию. Все защищённые копии не должны быть уничтожены, а выбранная версия должна восстановиться и пройти проверку.

Подтверждения

Реестр активов и версий, манифесты и контрольные суммы, политика неизменяемости и удержания, схема независимых полномочий и доменов, журналы удаления и отчёты учений с измеренными RPO/RTO.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Неподтверждённая или повреждённая версия не возвращается в рабочий контур. Затронутая функция блокируется либо использует защищённую версию, ранее успешно прошедшую проверку восстановления; отсутствие проверенной копии немедленно эскалируется владельцу риска.

Исправление и повторная проверка

Создают независимую копию, исправляют манифест, права удаления или путь восстановления и повторяют исходное уничтожение и восстановление той же версии. Проверки состава и пригодности должны пройти в пределах решения, а рабочая идентичность не должна затронуть все копии.

Побочные эффекты

Независимые копии увеличивают стоимость и усложняют законное удаление, синхронизацию удержания и управление большим числом версий.

Остаточный риск

Неизменяемая копия может содержать вредоносное или ошибочное состояние либо разделять управляющую плоскость с рабочей средой; восстановление не подтверждает смысловую корректность актива и готовность реагирования. Общий поставщик, административная

плоскость или ошибка до фиксации версии могут затронуть несколько копий, а восстановленная версия — оказаться несовместимой с изменившейся средой.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(c) · Поддерживает часть требования

Изолированная проверенная копия AI-данных или индекса помогает вернуть доступ к персональным данным после инцидента.

Граница связи: Мера охватывает только выбранные AI-активы; восстановление остальных данных и соблюдение нужного срока подтверждаются отдельными испытаниями.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Условие связи с карточкой	Актив содержит персональные данные либо нужен для доступа к ним; RPO и RTO проверяются для затронутой операции.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-IMP-033 · Утрата критичных данных и артефактов AI-системы, кроме датасетов, в результате атаки	570	141
AISAFE2-WEAK-085 · Отсутствие изолированных неизменяемых версий критичных AI-активов	701	296
AISAFE2-COMP-008 · Данные для подсистемы обучения	1170	52
AISAFE2-COMP-010 · Хранилище и реестр артефактов модели	1171	56
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1172	67
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1173	86
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1386	410

Мера защиты · AISAFE2-MIT-069

Наследование класса чувствительности векторными представлениями

Краткое описание

Каждое векторное представление и его копии наследуют не менее строгий класс чувствительности источника и соответствующие правила вычисления, хранения, передачи и экспорта.

Механизмы

- Управление и надзор
- Управление конфигурацией

Защитный механизм

Конвейер связывает вектор с источником и его действующим классом чувствительности. Представление, индекс, экспорт и резервная копия получают не менее строгий класс; смешанный объект получает наиболее строгий применимый класс. Потеря происхождения или неизвестная классификация блокирует обработку либо применяет наиболее строгий режим, зафиксированный в схеме наследования.

Формы контроля

- Архитектурная
- Техническая
- Процессная

Ответственность

- Заказчик
- Поставщик

Применимость

Применяется при вычислении, хранении, передаче, экспорте или резервировании эмбеддингов и иных векторных представлений данных. Строгость обработки определяется классом источника и калиброванной оценкой реконструкции, связывания записей и вывода атрибутов; мера не утверждает универсальную обратимость представления.

Предпосылки

- Нужны классификация источников, устойчивое происхождение каждого представления и инвентарь рабочих, экспортных, временных и резервных копий.
- Политики хранения, передачи и внешней обработки должны принимать эффективный класс как обязательный вход.

Реализация

Класс присваивается при создании представления и сохраняется при пакетировании, индексировании, копировании, экспорте и восстановлении. Для смешанного батча или индекса используется наиболее строгий класс либо доказуемое разделение. Перед внешним

вычислением проверяют допустимость получателя и условий обработки. Изменение или отзыв класса источника запускает обновление либо удаление производных объектов. Кэши, аналитические выгрузки и резервные копии включаются в тот же инвентарь.

Проверка

Создают представления для источников разных классов и смешанного набора, затем проверяют рабочий индекс, внешнее вычисление, экспорт, кэш и восстановленную копию. Удаляют происхождение и меняют класс источника. Мера неэффективна, если любой производный объект получает более слабый режим, неизвестный объект продолжает обработку без консервативного решения или изменение не распространяется.

Негативный тест

Пытаются сохранить, передать и экспортировать вектор чувствительного источника как общедоступный, смешивают его с низким классом и удаляют происхождение. Все варианты должны блокироваться либо получать наиболее строгий класс; контрольный низкочувствительный вектор сохраняет свой режим.

Подтверждения

Версии схемы наследования, ссылки происхождения, классы источников и векторов, инвентарь копий и потоков, решения внешней обработки, события изменения и результаты негативных тестов.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Вектор с неизвестным источником или классом не передаётся и не экспортируется и до исправления происхождения обрабатывается только в наиболее строгом изолированном режиме, зафиксированном в схеме наследования.

Исправление и повторная проверка

Восстанавливают происхождение и класс, переклассифицируют или удаляют все производные копии и повторяют исходную передачу или экспорт. Чувствительный вариант должен применить требуемую защиту, а контрольный разрешённый поток остаться работоспособным.

Побочные эффекты

Консервативное наследование повышает стоимость хранения и передачи, ограничивает совместное индексирование и может заблокировать внешние сервисы без достаточных гарантий.

Остаточный риск

Классификация не измеряет обратимость представления и сама по себе не предотвращает раскрытие; её должны исполнять отдельные контроли доступа, хранения, транспорта и потоков. Неизвестная техника восстановления, ошибочный класс источника или злоупотребление разрешённого получателя могут оставить защиту недостаточной.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и

выполнение меры оцениваются для конкретной системы.

OWASP Top 10 for Large Language Model Applications 2026

Редакция: OWASP Top 10 for LLM Applications 2026, 4 августа 2026 года

LLM09:2026 · Поддерживает часть требования

Наследование класса чувствительности источника не позволяет обрабатывать эмбеддинги и их копии как заведомо безвредные числа.

Граница связи: Класс сам не блокирует чтение или инверсию; его должны исполнять права доступа, изоляция и защита экспорта.

Субъект	Аналитик угроз AI-приложения
Условие применения	В приложении используются векторное хранилище, поиск по сходству или векторные представления данных.

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-087 · Ненаследование чувствительности исходных данных векторными представлениями	707	299
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	1174	48
AISAFE2-COMP-016 · Адаптер внешнего AI-сервиса	1175	69
AISAFE2-COMP-023 · Подсистема формирования базы знаний	1176	86
AISAFE2-COMP-025 · Источники данных для инструментов	1177	90
AISAFE2-MIT-054 · Принудительное соблюдение утвержденных ограничений обработки и передачи данных	1387	519
AISAFE2-MIT-062 · Шифрование данных и артефактов при хранении	1388	547
AISAFE2-MIT-063 · Защищённая передача данных и артефактов между компонентами	1389	550

Мера защиты · AISAFE2-MIT-070

Сквозное активное тестирование AI-системы

Краткое описание

Сквозная программа red teaming сочетает репрезентативные атаки и ручную экспертную проверку системы и процессов, а её актуальные результаты блокируют неприемлемый выпуск или продолжение эксплуатации.

Механизмы

- Валидация
- Шлюз допуска
- Жизненный цикл

Защитный механизм

Из текущих архитектуры и модели угроз формируется версия области оценки и набор сценариев для архитектуры, поведения модели, RAG, агентов, инструментов, конфигурации, операционных мер и процессов. Активные атаки и ручная экспертная проверка выполняются против точно идентифицированного состава системы; каждой находке назначаются владелец, приоритет, решение и срок. Незакрытая находка выше утверждённого порога блокирует выпуск или продолжение эксплуатации, а исправление либо компенсирующая мера проверяются исходным сценарием.

Формы контроля

- Техническая
- Процессная
- Выполняемая человеком

Ответственность

- Заказчик

Применимость

Мера применяется перед значимым выпуском, после существенного изменения архитектуры, модели, промпта, конфигурации, компонента или процесса, а также периодически и после инцидента в объёме, определённом владельцем риска. Неохваченная область должна быть явно исключена с обоснованием и решением владельца риска. Изменение любого объекта, от которого зависел результат, аннулирует соответствующее подтверждение и требует новой оценки. Тестирование не доказывает отсутствие недостатков и само по себе их не устраняет.

Предпосылки

- Требуются актуальные версии архитектуры и модели угроз с перечнем границ доверия, компонентов, процессов, активов, ролей и допустимых последствий теста.
- Нужны идентификаторы точных версий системы, модели, промптов, конфигурации, компонентов и выпуска, а также изолированная или контролируемая среда для безопасного воспроизведения атак.
- До начала оценки утверждаются пороги блокировки, владельцы находок, правила исключений и проверяемые процедуры телеметрии, обнаружения, эскалации, локализации и восстановления.

Реализация

План оценки связывают с точными редакциями архитектуры и модели угроз и включают в него репрезентативные сценарии против каждой объявленной области: архитектуры, поведения модели, RAG, агентов, инструментов, конфигурации, операционных мер и человеческих процессов. Автоматизированные и ручные атаки дополняются независимой экспертной проверкой предположений, комбинированных путей и результатов. В ходе упражнений проверяют появление и корреляцию телеметрии, срабатывание обнаружения, передачу на эскалацию, локализацию, восстановление и исполнение соответствующих действий людьми. Отчёт связывают с версиями всех проверенных объектов; находки получают владельца, приоритет, решение и срок. Изменение проверенного состава автоматически лишает результат статуса актуального. Исправление или компенсирующий контроль не закрывает находку до повторного выполнения исходного сценария и необходимых регрессионных сценариев.

Проверка

Для каждой заявленной области выполняют минимум одну репрезентативную атаку и фиксируют ожидаемый и фактический результат, ручную экспертную оценку, телеметрию, обнаружение и действия эскалации, локализации и восстановления. Проверяют, что отчёт однозначно связан с точными версиями системы, модели, промпта, конфигурации, компонентов и выпуска, а изменение любой из них делает подтверждение недействующим. Находка выше утверждённого порога должна блокировать выпуск или продолжение эксплуатации; после исправления либо компенсирующей меры исходный сценарий должен быть выполнен повторно.

Негативный тест

Исключают из плана один обязательный компонент и один человеческий процесс, меняют конфигурацию после успешной оценки и оставляют находку выше порога без решения. Подтверждение должно стать недействительным, а выпуск или продолжение эксплуатации — блокироваться до полной оценки и повторного выполнения исходного сценария.

Подтверждения

Сохраняются версии архитектуры и модели угроз, область и исключения, идентификаторы системы, модели, промптов, конфигурации, компонентов и выпуска, сценарии и ручные экспертные заметки, техническая телеметрия, решения обнаружения и реагирования, реестр находок с владельцами, приоритетами, решениями и сроками, блокировки, исключения и результаты повторных тестов.

Поведение при отказе

Блокировка процессом

Сценарий отказа

При отсутствии актуального подтверждения, пропуске обязательной области или наличии незакрытой находки выше утверждённого порога выпуск либо продолжение эксплуатации блокируется. Для уже работающей системы допускается только ограниченное по времени безопасное состояние с компенсирующими мерами и явным решением владельца риска.

Исправление и повторная проверка

Устраняют выявленный дефект либо вводят утверждённую компенсирующую меру, обновляют точные версии и область оценки и повторяют исходный сценарий тем же способом вместе с затронутыми регрессионными сценариями. Закрытие находки допускается только после подтверждения, что прежний путь атаки больше не достигает недопустимого результата, а необходимые сигналы и действия реагирования сохраняются.

Побочные эффекты

Активные сценарии могут нарушить тестовую или рабочую среду, создать ложные срабатывания и раскрыть чувствительные детали; ручная проверка и повторные оценки увеличивают срок выпуска и требуют контролируемых правил проведения.

Остаточный риск

Репрезентативная выборка и ручная экспертиза не охватывают все комбинации состояний, неизвестные техники и будущие изменения. Успешный тест подтверждает только проверенный сценарий и точный состав версии; он не доказывает отсутствие недостатков и не заменяет исправление, постоянное наблюдение или реагирование.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

СолидЛаб — анализ защищённости систем с AI

Частичная поддержка. Проводит активный поиск уязвимостей согласованного AI-контура и повторный тест после исправлений.

Ограничение: Проверка охватывает согласованный объём и момент проведения; она не создаёт автоматический шлюз выпуска и не доказывает защищённость после изменения версии. Блокировку выпуска и повторный тест организует заказчик.

Внешние требования и рекомендации

Сопоставление показывает связь карточки с частью положения. Применимость требования и выполнение меры оцениваются для конкретной системы.

Федеральный закон от 27.07.2006 № 152-ФЗ О персональных данных

Редакция: Редакция с изменениями по Федеральному закону № 265-ФЗ от 26.07.2026

ст. 19, ч. 2, п. 4 · Поддерживает часть требования

До ввода системы активные испытания могут проверить эффективность выбранных мер против атак на её AI-компоненты.

Граница связи: Активные испытания охватывают согласованные сценарии атак; оценка эффективности остальных обязательных мер должна быть завершена до ввода ИСПДн в эксплуатацию.

Субъект	Оператор персональных данных
Условие применения	ИСПДн с AI-компонентами готовится к вводу в эксплуатацию, и оценивается эффективность принятых мер защиты персональных данных.
Условие связи с карточкой	Испытания выполняются до ввода в эксплуатацию и проверяют выбранные меры защиты на точном составе системы.

Открыть положение в источнике

Regulation (EU) 2016/679, General Data Protection Regulation

Редакция: Regulation (EU) 2016/679, официальная публикация от 4 мая 2016 года

Article 32(1)(d) · Поддерживает часть требования

Повторяемые активные испытания AI-системы и проверка исправлений могут войти в регулярную оценку действенности мер защиты.

Граница связи: Программа должна отдельно охватить другие технические и организационные меры обработки персональных данных.

Субъект	Контролёр и обработчик персональных данных по GDPR
Условие применения	Обработка персональных данных требует мер безопасности, соразмерных установленному риску.
Начало действия положения	2018-05-25

Открыть положение в источнике

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-COMP-002 · Модель и интерфейс инференса	1178	37
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1179	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1180	44
AISAFE2-COMP-006 · RAG и извлечение из базы знаний	1181	48
AISAFE2-COMP-011 · Вычислительная и сетевая среда выполнения	1182	58
AISAFE2-COMP-012 · Идентификация, авторизация, ключи и секреты	1183	61
AISAFE2-COMP-013 · Телеметрия и обнаружение	1184	64
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1185	67
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1186	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1187	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1188	82
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1370	514
AISAFE2-MIT-022 · Регистрация и сквозное связывание событий AI-цепочки	1390	400
AISAFE2-MIT-023 · Выявление цепочек AI-атак и злоупотреблений	1391	405
AISAFE2-MIT-024 · Локализация и восстановление после AI-инцидентов	1392	410
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1393	415
AISAFE2-MIT-026 · Управление уязвимостями AI-компонентов в эксплуатации	1394	419
AISAFE2-MIT-067 · Эталон защищённой конфигурации и контроль отклонений	1395	564
AISAFE2-MIT-071 · Проверка безопасности собственного кода и бизнес-логики перед выпуском	1396	577

Мера защиты · AISAFE2-MIT-071

Проверка безопасности собственного кода и бизнес-логики перед выпуском

Краткое описание

Проверка собственного кода сочетает ручное ревью и выбранные по модели угроз техники анализа, связывает результаты с точными commit и build и блокирует выпуск при неприемлемых находках или обходе контроля.

Механизмы

- Валидация
- Шлюз допуска
- Жизненный цикл

Защитный механизм

Для first-party кода интерфейса и приложения, backend, интеграций, оркестрации, исполнения политик, агентов и бизнес-логики конвейер выбирает по модели угроз и типу компонента ручное security review, SAST, DAST, фаззинг, поиск секретов и целевые abuse-case тесты. Критичные для безопасности изменения требуют независимого ручного ревью. Результаты связываются с точными commit и build; находка выше утверждённого порога, отсутствие обязательного результата или несовпадение выпускаемого артефакта блокируют выпуск, а исправление проверяется исходной проверкой.

Формы контроля

- Техническая
- Процессная
- Выполняемая человеком

Ответственность

- Заказчик

Применимость

Мера применяется к собственному исходному коду и бизнес-логике, которые организация разрабатывает или изменяет для UI и приложений, backend, интеграций, оркестрации, исполнения политик и агентных функций. Проверка компонентов цепочки поставки перед выпуском и управление уязвимостями AI-компонентов остаются за отдельными мерами. Эффективная конфигурация среды выполнения остаётся в границе отдельной меры по эталону защищённой конфигурации и контролю отклонений. Сквозное активное тестирование поведения модели, RAG, промптов и агента выполняется отдельной мерой и специализированными контролями. Один SAST не считается полной проверкой бизнес-логики.

Предпосылки

- Требуются актуальные модель угроз, классификация компонентов и перечень критичных для безопасности путей и бизнес-инвариантов собственного кода.
- Нужны воспроизводимая сборка с идентификаторами commit и build, защищённые правила ветвей и выпуска и возможность сопоставить развёртываемый артефакт с проверенным результатом.
- Должны быть утверждены применимость SAST, DAST, фаззинга, поиска секретов и abuse-case тестов, пороги блокировки, независимые рецензенты и ограниченный процесс исключений.

Реализация

Для каждого изменения определяют затронутые first-party поверхности и по модели угроз выбирают применимые SAST, DAST, фаззинг, поиск секретов и целевые abuse-case тесты; неприменимость техники обосновывают. Ручное security review проверяет границы доверия, авторизацию, валидацию, ошибки, гонки и бизнес-инварианты, а для критичных изменений выполняется независимо от автора. Все результаты, одобрения и исключения связывают с точными commit и build. Выпускной шлюз принимает только этот build и отклоняет прямой merge, ручную сборку, пропуск обязательной проверки и развёртывание иного артефакта. После исправления повторяют исходную проверку и затронутый набор регрессии.

Проверка

На репрезентативных изменениях UI или приложения, backend, интеграции, оркестрации, исполнения политик, агента и бизнес-логики проверяют применение выбранных техник и независимого ручного ревью для критичных изменений. Каждая находка и каждое одобрение должны ссылаться на точные commit и build; находка выше утверждённого порога или отсутствующий результат должны блокировать выпуск. Отдельно проверяют, что прямой merge, ручная сборка и развёртывание артефакта, отличного от проверенного build, не достигают рабочей среды.

Негативный тест

Создают критичное изменение бизнес-логики с находкой выше порога и пытаются выпустить его через прямой merge, ручную сборку и подмену проверенного артефакта. Все пути должны блокироваться; после исправления исходная ручная или автоматизированная проверка повторяется для нового commit и build.

Подтверждения

Сохраняются версия модели угроз, область изменения и обоснование выбранных техник, отчёты SAST, DAST, фаззинга, поиска секретов и abuse-case тестов, ручные комментарии и независимые одобрения, реестр находок и исключений, точные commit, build и digest артефакта, решения шлюза и результаты тестов обхода и повторной проверки.

Поведение при отказе

Отказ с блокировкой

Сценарий отказа

Выпуск блокируется при отсутствующей обязательной проверке или независимом ревью, находке выше порога, недоступности связи commit с build, попытке обхода либо несовпадении артефакта. Срочное исключение допускается только в заранее ограниченном процессе с владельцем риска, сроком и обязательной повторной проверкой.

Исправление и повторная проверка

Исправляют код или бизнес-логику, закрывают путь обхода либо восстанавливают связь commit и build, затем повторяют исходную ручную проверку, тот же анализатор или abuse-case сценарий и затронутую регрессию для нового commit. Выпуск разрешается только для идентичного проверенного build после закрытия находки или документированного ограниченного исключения.

Побочные эффекты

Строгий шлюз и независимое ревью увеличивают время выпуска, анализаторы создают ложные срабатывания, а фаззинг и динамические тесты требуют воспроизводимой среды и могут расходовать значительные ресурсы.

Остаточный риск

Выбранные анализаторы и рецензенты могут пропустить многокомпонентную ошибку, скрытый путь или редкое состояние. Успешные проверки подтверждают только заявленный first-party код и точный build и не покрывают сторонние зависимости, эффективную конфигурацию или поведение модели, RAG, промптов и агента; SAST сам по себе не доказывает корректность бизнес-логики.

Сервисы Yandex Cloud

Если вы используете облачную инфраструктуру, эти возможности могут помочь реализовать часть меры. Наличие сервиса само по себе не подтверждает его настройку или эффективность контроля.

SourceCraft Security

Частичная поддержка. SAST, анализ зависимостей и поиск секретов SourceCraft Security дают результаты автоматических проверок исходного кода, которые можно включить в блокирующую политику выпуска.

Ограничение: Автоматические проверки не заменяют независимое ревью кода, DAST, фаззинг и тесты бизнес-логики; соответствие выпущенного артефакта проверенному commit/build нужно обеспечить отдельно.

Yandex Managed Service for GitLab

Частичная поддержка. Обязательные правила ревью могут блокировать слияние в защищённую ветку до получения требуемых подтверждений и одобрений.

Ограничение: Правила согласования сами по себе не выполняют сканирование и ручную экспертизу безопасности, не доказывают качество проверки и не гарантируют, что развёрнутый артефакт соответствует проверенной ветке; обходы сборки и выпуска требуют отдельного контроля.

Связанные карточки по типам

Пояснения и ограничения приведены в реестре связей. Справа указаны номера связей и страница карточки.

Карточка	Связи	Стр.
AISAFE2-WEAK-044 · Недостатки управления изменениями и развёртыванием AI-системы	642	222
AISAFE2-WEAK-047 · Недостаточное управление уязвимостями	648	227
AISAFE2-COMP-001 · Пользовательский интерфейс и API приложения	1189	35
AISAFE2-COMP-004 · Ядро агента и управление шагами (Agent Core)	1190	42
AISAFE2-COMP-005 · Инструменты, MCP-серверы и коннекторы	1191	44
AISAFE2-COMP-015 · Контур разработки, тестирования и выпуска	1192	67
AISAFE2-COMP-017 · Серверная часть приложения и бизнес-логика	1193	72
AISAFE2-COMP-019 · Контур исполнения политик и защиты AI-запросов	1194	77
AISAFE2-COMP-021 · Межагентское взаимодействие и реестр агентов	1195	82
AISAFE2-MIT-053 · Прослеживаемая архитектура, модель угроз и пересмотр риска	1371	514
AISAFE2-MIT-070 · Сквозное активное тестирование AI-системы	1396	573
AISAFE2-MIT-025 · Блокирующий шлюз выпуска AI-системы	1397	415

Реестр связей

Каждая запись сохраняет отдельную связь между карточками. Несколько записей для одной пары могут описывать разные механизмы или условия. Номер записи служит для навигации внутри документа.

Связь 1 · **AISAFE2-WEAK-001** → **AISAFE2-COMP-001**

В компоненте «Пользовательский интерфейс и API приложения» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли» проявляется недостаток «Избыточное доверие пользователя к AI-выводу»: Убедительный AI-вывод становится непропорционально значимым основанием решения и снижает независимую проверку со стороны пользователя, даже если система прямо не гарантирует его правильность.

Связь 2 · **AISAFE2-WEAK-001** → **AISAFE2-COMP-014**

В компоненте «Контур человеческой проверки и подтверждения» на поверхности «Правила внедрения человеческой проверки при обработке запроса» проявляется недостаток «Избыточное доверие пользователя к AI-выводу»: Убедительный AI-вывод становится непропорционально значимым основанием решения и снижает независимую проверку со стороны пользователя, даже если система прямо не гарантирует его правильность.

Связь 3 · **AISAFE2-WEAK-002** → **AISAFE2-COMP-001**

В компоненте «Пользовательский интерфейс и API приложения» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли» проявляется недостаток «Отсутствие привязки утверждений AI-ответа к источникам»: Для утверждений AI-ответа отсутствует проверяемая связь с фактически использованными источниками и статусом подтверждения.

Связь 4 · **AISAFE2-WEAK-002** → **AISAFE2-COMP-006**

Запросное извлечение возвращает фрагмент без проверяемой ссылки на фактически использованный источник и без статуса его подтверждения.

Связь 5 · **AISAFE2-WEAK-003** → **AISAFE2-COMP-001**

Интерфейс или API выдаёт AI-результат потребителю без обязательной проверки схемы ответа и формализуемых ограничений содержимого; результат, нарушающий контракт, не отклоняется.

Связь 6 · **AISAFE2-WEAK-003** → **AISAFE2-COMP-005**

Инструмент принимает сформированный моделью результат как аргумент вызова без обязательной проверки схемы и формализуемых ограничений перед исполнением операции.

Связь 7 · **AISAFE2-WEAK-003** → **AISAFE2-COMP-017**

Серверная часть использует AI-вывод в прикладном сценарии без обязательной проверки заданного контракта и формализуемых ограничений; нарушающий их результат не отклоняется до дальнейшего действия.

Связь 8 · AISAFE2-WEAK-003 → AISAFE2-COMP-019

Недостаток непосредственно возникает, когда контур исполнения политик принимает вывод модели без обязательной проверки контракта перед дальнейшим действием.

Связь 9 · AISAFE2-WEAK-003 → AISAFE2-COMP-024

Сгенерированный код, документ, изображение или сохраняемый текст пересекает границу к бизнес-получателю без обязательной проверки контракта, содержания и способа использования.

Связь 10 · AISAFE2-WEAK-004 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли» проявляется недостаток «Отсутствие контекстного кодирования или изоляции»: Содержимое модели попадает в активный синтаксический или исполняемый контекст без соответствующего экранирования либо изоляции.

Связь 11 · AISAFE2-WEAK-004 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» на поверхности «Общие инструменты и их среды выполнения» проявляется недостаток «Отсутствие контекстного кодирования или изоляции»: Содержимое модели попадает в активный синтаксический или исполняемый контекст без соответствующего экранирования либо изоляции.

Связь 12 · AISAFE2-WEAK-004 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Отсутствие контекстного кодирования или изоляции»: Содержимое модели попадает в активный синтаксический или исполняемый контекст без соответствующего экранирования либо изоляции.

Связь 13 · AISAFE2-WEAK-004 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» на поверхности «Сервисный и прикладной слой приложения» проявляется недостаток «Отсутствие контекстного кодирования или изоляции»: Содержимое модели попадает в активный синтаксический или исполняемый контекст без соответствующего экранирования либо изоляции.

Связь 14 · AISAFE2-WEAK-005 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли» проявляется недостаток «Недостаточная маркировка и предупреждения в интерфейсе»: Представление ответа не позволяет пользователю различить AI-генерацию, подтверждённый факт и результат, требующий проверки перед применением.

Связь 15 · AISAFE2-WEAK-006 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли» проявляется недостаток «Отсутствие раскрытия факта взаимодействия с AI»: Пользователь не получает явного сигнала, что взаимодействует с AI или видит созданный им материал.

Связь 16 · AISAFE2-WEAK-007 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса» проявляется недостаток «Отсутствие границы доверия между инструкциями и данными»: Доверенные правила и недоверенный контент объединяются в одном контексте без сохраняемой типизации и происхождения.

Связь 17 · AISAFE2-WEAK-007 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата» проявляется недостаток «Отсутствие границы доверия между инструкциями и данными»: Доверенные правила и недоверенный контент объединяются в одном контексте без сохраняемой типизации и происхождения.

Связь 18 · AISAFE2-WEAK-007 → AISAFE2-COMP-006

RAG возвращает недоверенный фрагмент с инструктивным содержимым без сохраняемой типизации и меток доверия, из-за чего его можно смешать с управляющими инструкциями.

Связь 19 · AISAFE2-WEAK-007 → AISAFE2-COMP-007

В компоненте «Долговременная память» на поверхности «Извлечение данных для последующих запросов» проявляется недостаток «Отсутствие границы доверия между инструкциями и данными»: Доверенные правила и недоверенный контент объединяются в одном контексте без сохраняемой типизации и происхождения.

Связь 20 · AISAFE2-WEAK-008 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Фундаментальный конфликт между возможностями и безопасностью»: Полезная способность модели выполнять широкий класс инструкций остаётся применимой и к нежелательным целям, не образуя детерминированной внутренней границы безопасности.

Ограничение: Только для генеративной модели, которая обрабатывает инструкции и контекст; перенос на иные ML-модели требует отдельной проверки.

Связь 21 · AISAFE2-WEAK-009 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Ненадёжное соблюдение системных инструкций»: Модель разрешает конфликт инструкций и ограничений вероятностно, поэтому ожидаемое старшинство директив не является инвариантом её поведения.

Ограничение: Только для генеративной модели, которая обрабатывает инструкции и контекст; перенос на иные ML-модели требует отдельной проверки.

Связь 22 · AISAFE2-WEAK-010 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Размытие обученных ограничений через контекст»: Содержимое текущего контекста способно ослабить влияние обученных ограничений без изменения весов модели.

Ограничение: Только для генеративной модели, которая обрабатывает инструкции и контекст; перенос на иные ML-модели требует отдельной проверки.

Связь 23 · AISAFE2-WEAK-010 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса» проявляется недостаток «Размытие обученных ограничений через контекст»: Содержимое текущего контекста способно ослабить влияние обученных ограничений без изменения весов модели.

Ограничение: Только для генеративной модели, которая обрабатывает инструкции и контекст; перенос на иные ML-модели требует отдельной проверки.

Связь 24 · AISAFE2-WEAK-011 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Поверхностная настройка модели на безопасное поведение»: Безопасное поведение модели зависит от поверхностных признаков запроса и теряет устойчивость при переформулировке или смене условий.

Ограничение: Только для генеративной модели, которая обрабатывает инструкции и контекст; перенос на иные ML-модели требует отдельной проверки.

Связь 25 · AISAFE2-WEAK-011 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Поверхностная настройка модели на безопасное поведение»: Безопасное поведение модели зависит от поверхностных признаков запроса и теряет устойчивость при переформулировке или смене условий.

Связь 26 · AISAFE2-WEAK-012 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» на поверхности «Общие инструменты и их среды выполнения» проявляется недостаток «Непроверяемая или скомпрометированная цепочка поставки AI»: Точный состав, происхождение или целостность AI/ML-компонентов выпуска нельзя подтвердить до их использования.

Связь 27 · AISAFE2-WEAK-012 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Непроверяемая или скомпрометированная цепочка поставки AI»: Точный состав, происхождение или целостность AI/ML-компонентов выпуска нельзя подтвердить до их использования.

Связь 28 · AISAFE2-WEAK-012 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Непроверяемая или скомпрометированная цепочка поставки AI»: Точный состав, происхождение или целостность AI/ML-компонентов выпуска нельзя подтвердить до их использования.

Связь 29 · AISAFE2-WEAK-012 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Непроверяемая или скомпрометированная цепочка поставки AI»: Точный состав, происхождение или целостность AI/ML-компонентов выпуска нельзя подтвердить до их использования.

Связь 30 · AISAFE2-WEAK-012 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Непроверяемая или скомпрометированная цепочка поставки AI»: Точный состав, происхождение или целостность AI/ML-компонентов выпуска нельзя подтвердить до их использования.

Связь 31 · AISAFE2-WEAK-012 → AISAFE2-COMP-022

В компоненте «Программные артефакты и цепочка поставки» на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости» проявляется недостаток «Непроверяемая или скомпрометированная цепочка поставки AI»: Точный состав, происхождение или целостность AI/ML-компонентов выпуска нельзя подтвердить до их использования.

Связь 32 · AISAFE2-WEAK-013 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Отсутствие прослеживаемости происхождения и преобразований данных»: Для данных нельзя восстановить подтверждённые сведения об источнике, версии и значимых преобразованиях, с которыми они попали в AI-контур.

Связь 33 · AISAFE2-WEAK-013 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Отсутствие прослеживаемости происхождения и преобразований данных»: Для данных нельзя восстановить подтверждённые сведения об источнике, версии и значимых преобразованиях, с которыми они попали в AI-контур.

Связь 34 · AISAFE2-WEAK-013 → AISAFE2-COMP-023

Для подготовленной базы знаний нельзя восстановить подтверждённые сведения об источнике, версии и преобразованиях, с которыми данные были опубликованы.

Связь 35 · AISAFE2-WEAK-013 → AISAFE2-COMP-025

Для исходного набора данных или внешней поверхности инструмента нельзя подтвердить происхождение, версию и значимые изменения до поступления в AI-контур.

Связь 36 · AISAFE2-WEAK-014 → AISAFE2-COMP-007

В компоненте «Долговременная память» на поверхности «Правила записи, обновления, истечения и удаления» проявляется недостаток «Неконтролируемый допуск материала в источники знаний и контуры обратной связи»: Непроверенный внешний или сгенерированный материал допускается в источник знаний, память либо повторный цикл обработки.

Связь 37 · AISAFE2-WEAK-014 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Неконтролируемый допуск материала в источники знаний и контуры обратной связи»: Непроверенный внешний или сгенерированный материал допускается в источник знаний, память либо повторный цикл обработки.

Связь 38 · AISAFE2-WEAK-014 → AISAFE2-COMP-023

На этапе допуска и публикации непроверенный внешний или сгенерированный материал входит в подготовленную базу знаний и влияет на последующие ответы.

Связь 39 · AISAFE2-WEAK-015 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Переобучение модели»: Модель воспроизводит особенности обучающей или оценочной выборки вместо устойчивого поведения на целевых данных.

Связь 40 · AISAFE2-WEAK-015 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Переобучение модели»: Модель воспроизводит особенности обучающей или оценочной выборки вместо устойчивого поведения на целевых данных.

Связь 41 · AISAFE2-WEAK-015 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Переобучение модели»: Модель воспроизводит особенности обучающей или оценочной выборки вместо устойчивого поведения на целевых данных.

Связь 42 · AISAFE2-WEAK-016 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Слабый контроль доступа к обучающим данным, конвейерам и артефактам»: Машинное задание обучения или обработки артефакта получает более широкие полномочия, чем нужны для конкретного запуска и его входов или выходов.

Связь 43 · AISAFE2-WEAK-016 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Слабый контроль доступа к обучающим данным, конвейерам и артефактам»: Машинное задание обучения или обработки артефакта получает более широкие полномочия, чем нужны для конкретного запуска и его входов или выходов.

Связь 44 · AISAFE2-WEAK-016 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Слабый контроль доступа к обучающим данным, конвейерам и артефактам»: Машинное задание обучения или обработки артефакта получает более широкие полномочия, чем нужны для конкретного запуска и его входов или выходов.

Связь 45 · AISAFE2-WEAK-016 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» на поверхности «Каталоги пользователей и сервисных идентичностей» проявляется недостаток «Слабый контроль доступа к обучающим данным, конвейерам и артефактам»: Машинное задание обучения или обработки артефакта получает более широкие полномочия, чем нужны для конкретного запуска и его входов или выходов.

Связь 46 · AISAFE2-WEAK-017 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Отсутствие проверяемых критериев выбора модели»: Процесс выбора модели или конфигурации не задаёт проверяемые критерии пригодности и правила сравнения кандидатов для целевого сценария.

Связь 47 · AISAFE2-WEAK-018 → AISAFE2-COMP-002

Параметры выдачи делают ответы модели избыточно подробными и устойчивыми к сравнению; при доступе внешнего субъекта и отсутствии ограничений на систематический опрос это облегчает исследование поведения модели.

Связь 48 · AISAFE2-WEAK-018 → AISAFE2-COMP-002

Интерфейс инференса позволяет внешнему субъекту проводить адаптивные серии запросов и получать информативный устойчивый сигнал соответствия входов и выходов без ограничений на систематическое исследование модели.

Связь 49 · AISAFE2-WEAK-019 → AISAFE2-COMP-002

Интерфейс инференса возвращает недоверенному потребителю полный вектор оценок по альтернативам вместо минимального результата, раскрывая избыточный сигнал о поведении модели.

Связь 50 · AISAFE2-WEAK-019 → AISAFE2-COMP-002

Пользовательский или диагностический API возвращает недоверенному потребителю логиты либо логарифмы правдоподобия, не требующиеся для заявленной операции.

Связь 51 · AISAFE2-WEAK-019 → AISAFE2-COMP-002

Интерфейс инференса передаёт недоверенному потребителю детальные оценки уверенности сверх необходимого результата, раскрывая дополнительный сигнал о поведении модели.

Связь 52 · AISAFE2-WEAK-022 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» на поверхности «Каталоги пользователей и сервисных идентичностей» проявляется недостаток «Недостаточный контроль доступа на этапах обучения, хранения и эксплуатации»: Архитектура не задаёт согласованную политику авторизации по субъектам, операциям и стадиям жизненного цикла AI-системы.

Связь 53 · AISAFE2-WEAK-023 → AISAFE2-COMP-007

В компоненте «Долговременная память» на поверхности «Сохранённые единицы информации» проявляется недостаток «Незашифрованное хранение»: Постоянные данные или артефакты доступны в хранилище без криптографической защиты содержимого.

Связь 54 · AISAFE2-WEAK-023 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Незашифрованное хранение»: Постоянные данные или артефакты доступны в хранилище без криптографической защиты содержимого.

Связь 55 · AISAFE2-WEAK-023 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Незашифрованное хранение»: Постоянные данные или артефакты доступны в хранилище без криптографической защиты содержимого.

Связь 56 · AISAFE2-WEAK-023 → AISAFE2-COMP-011

Незащищённая временная копия на диске, снимке или носителе среды выполнения непосредственно реализует незашифрованное хранение.

Связь 57 · AISAFE2-WEAK-023 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» на поверхности «Сборщики журналов, трассировок и метрик» проявляется недостаток «Незашифрованное хранение»: Постоянные данные или артефакты доступны в хранилище без криптографической защиты содержимого.

Связь 58 · AISAFE2-WEAK-023 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» на поверхности «Операционные реляционные и документные базы данных» проявляется недостаток «Незашифрованное хранение»: Постоянные данные или артефакты доступны в хранилище без криптографической защиты содержимого.

Связь 59 · AISAFE2-WEAK-023 → AISAFE2-COMP-023

Подготовленный корпус, поисковый индекс или промежуточная копия базы знаний хранится без требуемой криптографической защиты содержимого.

Связь 60 · AISAFE2-WEAK-024 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» на поверхности «MCP-клиенты и MCP-серверы» проявляется недостаток «Незашифрованные каналы передачи»: Данные или артефакты передаются между компонентами без защищённого транспортного шифрования.

Связь 61 · AISAFE2-WEAK-024 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Незашифрованные каналы передачи»: Данные или артефакты передаются между компонентами без защищённого транспортного шифрования.

Связь 62 · AISAFE2-WEAK-024 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Незашифрованные каналы передачи»: Данные или артефакты передаются между компонентами без защищённого транспортного шифрования.

Связь 63 · AISAFE2-WEAK-024 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» на поверхности «Клиенты, прокси и адаптеры внешних AI API» проявляется недостаток «Незашифрованные каналы передачи»: Данные или артефакты передаются между компонентами без защищённого транспортного шифрования.

Связь 64 · AISAFE2-WEAK-024 → AISAFE2-COMP-021

При передаче сообщения или артефакта через A2A-шлюз либо брокер отсутствие защищённого транспортного канала открывает содержимое наблюдателю на маршруте.

Ограничение: Только для фактического межагентского канала с данными, требующими защиты; наличие реестра агентов само по себе не доказывает передачу.

Связь 65 · AISAFE2-WEAK-025 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» на поверхности «Каталоги пользователей и сервисных идентичностей» проявляется недостаток «Недостаточное управление ключами и секретами»: Ключи и секреты не отделены от защищаемых данных и операций по владельцу, назначению или сроку действия.

Связь 66 · AISAFE2-WEAK-026 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли» проявляется недостаток «Отсутствие ограничений частоты запросов»: Интерфейсы инференса принимают повторные запросы без сквозной квоты, привязанной к субъекту и совокупности его точек доступа.

Связь 67 · AISAFE2-WEAK-026 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «API и протокол обращения к модели» проявляется недостаток «Отсутствие ограничений частоты запросов»: Интерфейсы инференса принимают повторные запросы без сквозной квоты, привязанной к субъекту и совокупности его точек доступа.

Связь 68 · AISAFE2-WEAK-026 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» на поверхности «Клиенты, прокси и адаптеры внешних AI API» проявляется недостаток «Отсутствие ограничений частоты запросов»: Интерфейсы инференса принимают повторные запросы без сквозной квоты, привязанной к субъекту и совокупности его точек доступа.

Связь 69 · AISAFE2-WEAK-026 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» на поверхности «AI-шлюзы и прокси принудительного контроля» проявляется недостаток «Отсутствие ограничений частоты запросов»: Интерфейсы инференса принимают повторные запросы без сквозной квоты, привязанной к субъекту и совокупности его точек доступа.

Связь 70 · AISAFE2-WEAK-028 → AISAFE2-COMP-002

На поверхности «API и протокол обращения к модели» не сохраняются или не связываются события вызовов по субъекту и цепочке запросов, поэтому последовательность извлечения модели может затеряться среди обычного трафика.

Ограничение: Связь охватывает события на API инференса; обнаружение кампании требует корреляции и правил в отдельном контуре телеметрии.

Связь 71 · AISAFE2-WEAK-028 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» на поверхности «Сборщики журналов, трассировок и метрик» проявляется недостаток «Недостатки наблюдаемости и обнаружения злоупотреблений»: AI-цепочка не формирует или не связывает события, достаточные для выявления и расследования злоупотреблений.

Связь 72 · AISAFE2-WEAK-028 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» на поверхности «Сборщики журналов, трассировок и метрик» проявляется недостаток «Недостатки наблюдаемости и обнаружения злоупотреблений»: Запросы к модели не анализируются как связанная кампания, поэтому характерные последовательности извлечения теряются среди обычного трафика.

Связь 73 · AISAFE2-WEAK-028 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» на поверхности «Клиенты, прокси и адаптеры внешних AI API» проявляется недостаток «Недостатки наблюдаемости и обнаружения злоупотреблений»: Запросы к модели не анализируются как связанная кампания, поэтому характерные последовательности извлечения теряются среди обычного трафика.

Связь 74 · AISAFE2-WEAK-030 → AISAFE2-COMP-002

На поверхности «Подготовка и выдача результата инференса» характерные выходы модели не получают заранее заданной проверяемой метки, поэтому при расследовании труднее связать подозрительную копию с защищаемой версией.

Ограничение: Связь относится к маркировке выдаваемых результатов; маркировка весов и других артефактов модели проверяется на их собственных границах.

Связь 75 · AISAFE2-WEAK-030 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Отсутствие проверяемой маркировки модели и её выходов»: У модели, её артефактов или характерных выходов нет проверяемой контрольной метки, пригодной как дополнительный признак происхождения и несанкционированного копирования.

Связь 76 · AISAFE2-WEAK-031 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Недостаточная защита модели при поставке в недоверенную среду»: Артефакт модели передается для исполнения в среду, где владелец модели не контролирует хост, память и пути копирования.

Связь 77 · AISAFE2-WEAK-031 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Недостаточная защита модели при поставке в недоверенную среду»: Артефакт модели передается для исполнения в среду, где владелец модели не контролирует хост, память и пути копирования.

Связь 78 · AISAFE2-WEAK-032 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Недостаточная защита физического доступа к модели и инфраструктуре»: Физический или аппаратно-привилегированный доступ к хосту, носителю либо ускорителю может обойти логические ограничения на артефакт модели.

Связь 79 · AISAFE2-WEAK-032 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Недостаточная защита физического доступа к модели и инфраструктуре»: Физический или аппаратно-привилегированный доступ к хосту, носителю либо ускорителю может обойти логические ограничения на артефакт модели.

Связь 80 · AISAFE2-WEAK-033 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Раскрытие сведений об отдельных обучающих записях через модель»: Обучение на чувствительных записях может закрепить их содержание в параметрах так, что модель воспроизводит фрагменты при подходящих запросах.

Связь 81 · AISAFE2-WEAK-033 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Раскрытие сведений об отдельных обучающих записях через модель»: Поведение модели позволяет вывести атрибуты или участие отдельной обучающей записи даже без дословного воспроизведения исходного примера.

Связь 82 · AISAFE2-WEAK-033 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Раскрытие сведений об отдельных обучающих записях через модель»: Обучение на чувствительных записях может закрепить их содержание в параметрах так, что модель воспроизводит фрагменты при подходящих запросах.

Связь 83 · AISAFE2-WEAK-033 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Раскрытие сведений об отдельных обучающих записях через модель»: Обучение на чувствительных записях может закрепить их содержание в параметрах так, что модель воспроизводит фрагменты при подходящих запросах.

Связь 84 · AISAFE2-WEAK-035 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Недостатки управления жизненным циклом данных»: Решения об инвентаризации, классификации, допустимом использовании, доступе, хранении и изменении данных не имеют полного управляемого жизненного цикла.

Связь 85 · AISAFE2-WEAK-035 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостатки управления жизненным циклом данных»: Решения об инвентаризации, классификации, допустимом использовании, доступе, хранении и изменении данных не имеют полного управляемого жизненного цикла.

Связь 86 · AISAFE2-WEAK-035 → AISAFE2-COMP-025

Исходные данные инструмента не имеют полного инвентаря, владельца, класса чувствительности, правил доступа, допустимого использования и срока хранения.

Связь 87 · AISAFE2-WEAK-036 → AISAFE2-COMP-003

При сборке запроса в пользовательскую часть контекста включаются записи или поля, не нужные для цели обработки, хотя их можно исключить до обращения к модели.

Связь 88 · AISAFE2-WEAK-036 → AISAFE2-COMP-007

В долговременную память записываются данные, не нужные для её назначения, хотя их можно исключить до сохранения и дальнейшего включения в AI-контекст.

Связь 89 · AISAFE2-WEAK-036 → AISAFE2-COMP-008

В набор для обучения или оценки включаются записи или поля, не нужные для его заявленного назначения и исключаемые до использования набора.

Связь 90 · AISAFE2-WEAK-036 → AISAFE2-COMP-013

Сборщики журналов и трассировок сохраняют поля запросов или ответов, не нужные для предусмотренной диагностики и исключаемые до записи.

Связь 91 · AISAFE2-WEAK-036 → AISAFE2-COMP-016

Адаптер передаёт внешнему AI-сервису поля запроса, не нужные для модельной операции и исключаемые до отправки.

Связь 92 · AISAFE2-WEAK-036 → AISAFE2-COMP-018

Операционное хранилище или очередь передаёт в AI-обработку целые записи либо сообщения вместо необходимых полей, хотя остальные поля можно исключить до передачи.

Связь 93 · AISAFE2-WEAK-036 → AISAFE2-COMP-019

Контур исполнения политик пропускает в модельный или инструментальный маршрут данные сверх необходимых для цели обработки, хотя лишние поля можно исключить на этой границе.

Связь 94 · AISAFE2-WEAK-036 → AISAFE2-COMP-023

При подготовке корпуса сохраняются записи или поля, которые не нужны для заявленной цели базы знаний и должны быть исключены.

Связь 95 · AISAFE2-WEAK-037 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Непропорциональный вклад отдельных обучающих записей в результат обучения»: Отдельные или повторяющиеся обучающие записи оказывают непропорциональное влияние на поведение модели и повышают риск запоминания либо восстановления данных.

Связь 96 · AISAFE2-WEAK-037 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Непропорциональный вклад отдельных обучающих записей в результат обучения»: Отдельные или повторяющиеся обучающие записи оказывают непропорциональное влияние на поведение модели и повышают риск запоминания либо восстановления данных.

Связь 97 · AISAFE2-WEAK-039 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата» проявляется недостаток «Выдача инструментом чувствительных или избыточных данных»: Инструмент или коннектор возвращает получателю чувствительные либо избыточные данные без минимизации, проекции полей и сохраняемой метки чувствительности.

Связь 98 · AISAFE2-WEAK-040 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 99 · AISAFE2-WEAK-040 → AISAFE2-COMP-007

В компоненте «Долговременная память» на поверхности «Сохранённые единицы информации» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 100 · AISAFE2-WEAK-040 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» на поверхности «Обучающие и валидационные выборки» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 101 · AISAFE2-WEAK-040 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» на поверхности «Сборщики журналов, трассировок и метрик» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 102 · AISAFE2-WEAK-040 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» на поверхности «Клиенты, прокси и адаптеры внешних AI API» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 103 · AISAFE2-WEAK-040 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» на поверхности «Операционные реляционные и документные базы данных» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 104 · AISAFE2-WEAK-040 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» на поверхности «AI-шлюзы и прокси принудительного контроля» проявляется недостаток «Отсутствие маскирования или анонимизации чувствительных данных»: Необходимые для цели записи и поля сохраняются с чувствительными значениями в исходном виде вместо требуемого маскирования, токенизации или анонимизации.

Связь 105 · AISAFE2-WEAK-040 → AISAFE2-COMP-023

Необходимые для базы знаний чувствительные значения публикуются без требуемого маскирования, токенизации или анонимизации.

Связь 106 · AISAFE2-WEAK-041 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса» проявляется недостаток «Недостаточный контроль доступа к системным инструкциям и секретам»: Доступ к конфигурации системного промпта, инструкциям инструментов или связанным секретам не ограничен по субъекту и операции.

Связь 107 · AISAFE2-WEAK-041 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Недостаточный контроль доступа к системным инструкциям и секретам»: Доступ к конфигурации системного промпта, инструкциям инструментов или связанным секретам не ограничен по субъекту и операции.

Связь 108 · AISAFE2-WEAK-041 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» на поверхности «Каталоги пользователей и сервисных идентичностей» проявляется недостаток «Недостаточный контроль доступа к системным инструкциям и секретам»: Доступ к конфигурации системного промпта, инструкциям инструментов или связанным секретам не ограничен по субъекту и операции.

Связь 109 · AISAFE2-WEAK-041 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостаточный контроль доступа к системным инструкциям и секретам»: Доступ к конфигурации системного промпта, инструкциям инструментов или связанным секретам не ограничен по субъекту и операции.

Связь 110 · AISAFE2-WEAK-041 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» на поверхности «Операционные реляционные и документные базы данных» проявляется недостаток «Недостаточный контроль доступа к системным инструкциям и секретам»: Доступ к конфигурации системного промпта, инструкциям инструментов или связанным секретам не ограничен по субъекту и операции.

Связь 111 · AISAFE2-WEAK-042 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы» проявляется недостаток «Недостаточная пригодность модели для целевого сценария»: Конкретная версия модели не проходит фактическую репрезентативную проверку по заранее заданным критериям качества или безопасности целевого сценария.

Связь 112 · AISAFE2-WEAK-042 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостаточная пригодность модели для целевого сценария»: Конкретная версия модели не проходит фактическую репрезентативную проверку по заранее заданным критериям качества или безопасности целевого сценария.

Связь 113 · AISAFE2-WEAK-043 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса» проявляется недостаток «Недостаточные ограничения в системной инструкции»: Системная инструкция неполно, неоднозначно или противоречиво задаёт запреты и границы поведения, ожидаемые в целевом сценарии.

Связь 114 · AISAFE2-WEAK-044 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостатки управления изменениями и развёртыванием AI-системы»: Разработка и выпуск AI-системы допускают изменения без обязательных проверок безопасности, владельцев решения и подтверждений, связанных с точной версией.

Связь 115 · AISAFE2-WEAK-045 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Недостатки управления учётными записями и правами в контуре MLOps»: Человеческие и административные учётные записи MLOps получают общие, избыточные или долговечные права, которые не пересматриваются при изменении роли.

Связь 116 · AISAFE2-WEAK-045 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Недостатки управления учётными записями и правами в контуре MLOps»: Человеческие и административные учётные записи MLOps получают общие, избыточные или долговечные права, которые не пересматриваются при изменении роли.

Связь 117 · AISAFE2-WEAK-045 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» на поверхности «Каталоги пользователей и сервисных идентичностей» проявляется недостаток «Недостатки управления учётными записями и правами в контуре MLOps»: Человеческие и административные учётные записи MLOps получают общие, избыточные или долговечные права, которые не пересматриваются при изменении роли.

Связь 118 · AISAFE2-WEAK-045 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостатки управления учётными записями и правами в контуре MLOps»: Человеческие и административные учётные записи MLOps получают общие, избыточные или долговечные права, которые не пересматриваются при изменении роли.

Связь 119 · AISAFE2-WEAK-045 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» на поверхности «Организации, проекты и административные API провайдера» проявляется недостаток «Недостатки управления учётными записями и правами в контуре MLOps»: Человеческие и административные учётные записи MLOps получают общие, избыточные или долговечные права, которые не пересматриваются при изменении роли.

Связь 120 · AISAFE2-WEAK-046 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостаточное управление соответствием требованиям»: Применимые требования, владельцы и подтверждения соответствия не связаны с конкретной архитектурой и выпуском AI-системы.

Связь 121 · AISAFE2-WEAK-047 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» на поверхности «Подготовка данных и признаков для обучения» проявляется недостаток «Недостаточное управление уязвимостями»: Эксплуатируемые AI-компоненты не имеют полного инвентаря и процесса оценки, исправления и подтверждения устранения уязвимостей.

Связь 122 · AISAFE2-WEAK-047 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» на поверхности «Веса, конфигурация и формат модели» проявляется недостаток «Недостаточное управление уязвимостями»: Эксплуатируемые AI-компоненты не имеют полного инвентаря и процесса оценки, исправления и подтверждения устранения уязвимостей.

Связь 123 · AISAFE2-WEAK-047 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Недостаточное управление уязвимостями»: Эксплуатируемые AI-компоненты не имеют полного инвентаря и процесса оценки, исправления и подтверждения устранения уязвимостей.

Связь 124 · AISAFE2-WEAK-047 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» на поверхности «Репозитории изменений, рецензирование и CI/CD» проявляется недостаток «Недостаточное управление уязвимостями»: Эксплуатируемые AI-компоненты не имеют полного инвентаря и процесса оценки, исправления и подтверждения устранения уязвимостей.

Связь 125 · AISAFE2-WEAK-047 → AISAFE2-COMP-022

В компоненте «Программные артефакты и цепочка поставки» на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости» проявляется недостаток «Недостаточное управление уязвимостями»: Эксплуатируемые AI-компоненты не имеют полного инвентаря и процесса оценки, исправления и подтверждения устранения уязвимостей.

Связь 126 · AISAFE2-WEAK-049 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» на поверхности «Каталоги пользователей и сервисных идентичностей» проявляется недостаток «Слабые или отсутствующие механизмы установления доверия в A2A»: Межагентское сообщение принимается без достаточной проверки отправителя, целостности, адресата, задачи, свежести или повторного использования.

Связь 127 · AISAFE2-WEAK-049 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» на поверхности «Реестры агентов и описание их возможностей» проявляется недостаток «Слабые или отсутствующие механизмы установления доверия в A2A»: Межагентское сообщение принимается без достаточной проверки отправителя, целостности, адресата, задачи, свежести или повторного использования.

Связь 128 · AISAFE2-WEAK-050 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» на поверхности «Виртуальные машины, контейнеры и оркестраторы» проявляется недостаток «Остаточные данные в общей инфраструктуре исполнения моделей»: Данные предыдущего процесса, задания или арендатора остаются доступными следующему потребителю общего ресурса исполнения моделей.

Связь 129 · AISAFE2-WEAK-051 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» на поверхности «Сервисный и прикладной слой приложения» проявляется недостаток «Выдача результата без завершения обязательных стадий вычислений»: Система выпускает обычный, успешный или разрешающий результат, хотя одна или несколько обязательных стадий анализа не получили корректного конечного состояния.

Связь 130 · AISAFE2-WEAK-051 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» на поверхности «AI-шлюзы и прокси принудительного контроля» проявляется недостаток «Выдача результата без завершения обязательных стадий вычислений»: Система выпускает обычный, успешный или разрешающий результат, хотя одна или несколько обязательных стадий анализа не получили корректного конечного состояния.

Связь 131 · AISAFE2-WEAK-052 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» на поверхности «Планирование и последовательность шагов» проявляется недостаток «Неограниченные внутренние циклы и усиление нагрузки»: Один допустимый запрос способен породить неограниченные агентные циклы, повторные попытки или разветвление и занять несоразмерный объём ресурсов до результата.

Связь 132 · AISAFE2-WEAK-052 → AISAFE2-COMP-011

Среда выполнения не создаёт агентный цикл, но при отсутствии жёстких лимитов времени, памяти и параллелизма позволяет циклу, возникшему в приложении или Agent Core, занять несоразмерный объём ресурсов.

Ограничение: Только если задания проходят через эту среду и её ресурсные пределы не ограничивают совокупную операцию.

Связь 133 · AISAFE2-WEAK-052 → AISAFE2-COMP-016

Неограниченные повторные обращения и обработка отказа внешнего провайдера непосредственно создают внутреннюю работу на поверхности адаптера.

Связь 134 · AISAFE2-WEAK-052 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» на поверхности «Сервисный и прикладной слой приложения» проявляется недостаток «Неограниченные внутренние циклы и усиление нагрузки»: Один допустимый запрос способен породить неограниченные агентные циклы, повторные попытки или разветвление и занять несоразмерный объём ресурсов до результата.

Связь 135 · AISAFE2-WEAK-052 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» на поверхности «Очереди сообщений, задания и событийные шины» недостаток проявляется, если одно задание без жёсткого предела повторов или порождений создаёт новые задания и несоразмерную нагрузку.

Ограничение: Только для очередей и фоновых заданий, где число повторов, потомков или общий бюджет операции не ограничены; хранение данных само по себе не создаёт цикл.

Связь 136 · AISAFE2-WEAK-053 → AISAFE2-COMP-001

Незакрытая конечная точка или API оставляет возможность новых вызовов после вывода приложения.

Связь 137 · AISAFE2-WEAK-053 → AISAFE2-COMP-002

Исполняемая модель или её интерфейс остаются доступными после решения о выводе.

Связь 138 · AISAFE2-WEAK-053 → AISAFE2-COMP-004

Ядро агента сохраняет возможность планировать и выполнять шаги после вывода агента.

Связь 139 · AISAFE2-WEAK-053 → AISAFE2-COMP-005

Инструмент, MCP-сервер или коннектор сохраняет маршрут либо возможность действия после вывода.

Связь 140 · AISAFE2-WEAK-053 → AISAFE2-COMP-007

Записи долговременной памяти остаются доступными после прекращения назначения компонента.

Связь 141 · AISAFE2-WEAK-053 → AISAFE2-COMP-008

Обучающие или оценочные наборы остаются доступными сверх принятого решения о хранении или удалении.

Связь 142 · AISAFE2-WEAK-053 → AISAFE2-COMP-009

После вывода остаётся запускаемый конвейер, расписание или задача обучения.

Связь 143 · AISAFE2-WEAK-053 → AISAFE2-COMP-010

Артефакт модели или запись реестра остаются доступными после вывода версии.

Связь 144 · AISAFE2-WEAK-053 → AISAFE2-COMP-011

После вывода остаются вычислительные ресурсы, маршруты либо временные объекты среды выполнения.

Связь 145 · AISAFE2-WEAK-053 → AISAFE2-COMP-012

Идентичность, полномочие, ключ или учётные данные продолжают действовать после вывода компонента.

Связь 146 · AISAFE2-WEAK-053 → AISAFE2-COMP-013

После вывода компонента продолжают работать неотозванные сборщики, правила или маршруты телеметрии либо его хранилище остаётся доступным сверх утверждённого срока и круга лиц.

Ограничение: Только для неотозванного рабочего маршрута или доступа вне утверждённого срока и круга лиц; изолированный архив по утверждённой политике допустим.

Связь 147 · AISAFE2-WEAK-053 → AISAFE2-COMP-015

Процесс вывода в контуре разработки и выпуска не отзывает полномочия и маршруты прекращающего работу компонента либо оставляет его управляемые копии доступными вне утверждённого срока и круга лиц.

Связь 148 · AISAFE2-WEAK-053 → AISAFE2-COMP-016

Адаптер сохраняет маршрут к внешнему AI-сервису после прекращения назначения интеграции.

Связь 149 · AISAFE2-WEAK-053 → AISAFE2-COMP-017

Серверный маршрут или бизнес-действие остаются доступными после вывода приложения.

Связь 150 · AISAFE2-WEAK-053 → AISAFE2-COMP-018

Рабочая запись, очередь, кэш или производная копия остаются доступными после вывода компонента.

Связь 151 · AISAFE2-WEAK-053 → AISAFE2-COMP-019

После вывода остаются действующие правила, маршруты или исключения контура исполнения политик.

Связь 152 · AISAFE2-WEAK-053 → AISAFE2-COMP-020

После вывода интеграции в административном контуре провайдера остаются действующие полномочия или маршруты, через которые можно продолжать обращения от имени выведенного компонента.

Связь 153 · AISAFE2-WEAK-053 → AISAFE2-COMP-021

Межагентский маршрут, регистрационная запись или полномочие агента остаются действующими после вывода.

Связь 154 · AISAFE2-WEAK-053 → AISAFE2-COMP-022

После вывода программный артефакт сохраняет исполняемый маршрут или может быть повторно развёрнут без проверки и допуска точной версии.

Ограничение: Только при действующем пути исполнения или повторного развёртывания без допуска; изолированное архивное хранение само по себе не образует недостаток.

Связь 155 · AISAFE2-WEAK-053 → AISAFE2-COMP-023

Опубликованный корпус, индекс или производная копия остаётся доступной после вывода подсистемы формирования базы знаний.

Связь 156 · AISAFE2-WEAK-054 → AISAFE2-COMP-002

Генеративная модель может выдать правдоподобный, но неверный результат при штатном запросе и при детерминированном декодировании; интерфейс инференса сам по себе не проверяет истинность.

Ограничение: Связь относится к генеративным моделям; для иной ML-модели механизм ошибки нужно устанавливать отдельно.

Связь 157 · AISAFE2-WEAK-055 → AISAFE2-COMP-004

Ядро агента превращает интерпретированное моделью правило в выбор шага, если обязательный инвариант не проверяется независимым механизмом до действия.

Связь 158 · AISAFE2-WEAK-055 → AISAFE2-COMP-017

Серверная бизнес-логика передаёт модели решение по обязательному правилу, когда не применяет это правило детерминированно перед прикладным эффектом.

Связь 159 · AISAFE2-WEAK-056 → AISAFE2-COMP-005

Инструмент получает сформированные моделью аргументы без проверки схемы, типов, диапазонов и допустимых сочетаний непосредственно перед исполнением.

Связь 160 · AISAFE2-WEAK-057 → AISAFE2-COMP-001

Пользовательский интерфейс допускает чувствительное действие без отдельного подтверждения точного получателя, параметров и ожидаемого эффекта.

Связь 161 · AISAFE2-WEAK-057 → AISAFE2-COMP-014

Контур человеческой проверки не создаёт блокирующего решения для неизменённого предложения до исполнения чувствительного действия.

Связь 162 · AISAFE2-WEAK-058 → AISAFE2-COMP-005

Инструмент выполняет запрос под общей сервисной идентичностью, не сохраняя права и область доступа инициатора конкретной сессии.

Связь 163 · AISAFE2-WEAK-058 → AISAFE2-COMP-012

Контур идентификации и авторизации не передаёт или не проверяет контекст инициатора для конкретного объекта и операции инструмента.

Связь 164 · AISAFE2-WEAK-059 → AISAFE2-COMP-002

Общая дообученная модель выдаёт всем арендаторам поведение, сформированное объединёнными данными, без возможности применить исходные права к параметрам ответа.

Связь 165 · AISAFE2-WEAK-059 → AISAFE2-COMP-008

Данные нескольких арендаторов объединяются в один набор дообучения без сохраняемой границы принадлежности и разрешённого использования.

Связь 166 · AISAFE2-WEAK-059 → AISAFE2-COMP-009

Конвейер дообучения создаёт общий артефакт из данных нескольких арендаторов без раздельных заданий, выходов или проверяемого правила изоляции.

Связь 167 · AISAFE2-WEAK-060 → AISAFE2-COMP-001

Внешний API или интерфейс возвращает технические детали внутренней ошибки вместо устойчивого пользовательского сообщения.

Связь 168 · AISAFE2-WEAK-060 → AISAFE2-COMP-002

Интерфейс модели или инференса раскрывает внутренние сведения через ненормализованную ошибку вместо ограниченного внешнего сообщения.

Связь 169 · AISAFE2-WEAK-060 → AISAFE2-COMP-005

Инструмент или коннектор передаёт необработанную диагностическую ошибку следующему компоненту либо пользователю.

Связь 170 · AISAFE2-WEAK-060 → AISAFE2-COMP-006

Интерфейс извлечения раскрывает через ошибку сведения о запросе, поисковом маршруте, индексе или внутреннем источнике.

Связь 171 · AISAFE2-WEAK-060 → AISAFE2-COMP-016

Адаптер внешнего AI-сервиса передаёт наружу либо в модель необработанные ошибки и служебные сведения поставщика.

Связь 172 · AISAFE2-WEAK-060 → AISAFE2-COMP-017

Серверная часть приложения не преобразует внутреннюю ошибку в ограниченный внешний контракт на границе ответа.

Связь 173 · AISAFE2-WEAK-061 → AISAFE2-COMP-013

Контур телеметрии публикует трассировки, метрики или служебные поля во внешний канал без отдельной минимизации и решения доступа.

Связь 174 · AISAFE2-WEAK-062 → AISAFE2-COMP-004

Ядро агента расходует шаги, время, токены и вызовы инструментов без общего остатка бюджета задачи, сохраняемого между итерациями.

Связь 175 · AISAFE2-WEAK-062 → AISAFE2-COMP-011

Среда выполнения допускает потребление вычислений, памяти, очередей или временного хранения без связи с общим бюджетом субъекта и задачи.

Связь 176 · AISAFE2-WEAK-062 → AISAFE2-COMP-017

Серверная часть учитывает локальные запросы, но не связывает расходы сессии, пользователя и внутренних AI-операций в единый предел.

Связь 177 · AISAFE2-WEAK-062 → AISAFE2-COMP-019

Контур исполнения политик не применяет единый бюджет до каждой учитываемой операции на всех охваченных маршрутах.

Связь 178 · AISAFE2-WEAK-062 → AISAFE2-COMP-020

Квоты и стоимостные ограничения внешнего AI-провайдера не связаны с бюджетами приложения, пользователя и агентной задачи.

Связь 179 · AISAFE2-WEAK-063 → AISAFE2-COMP-001

Интерфейс не показывает пользователю оценку или верхнюю границу стоимости, текущий расход, предупреждение и доступный предел до платной операции.

Связь 180 · AISAFE2-WEAK-063 → AISAFE2-COMP-017

Серверная часть тарифицирует либо инициирует платные операции без пользовательского бюджета, прозрачного учёта и обязательного подтверждения установленного порога.

Связь 181 · AISAFE2-WEAK-064 → AISAFE2-COMP-005

Отказ или задержка внешнего инструмента распространяется дальше, если вызов не ограничен по времени, числу повторов и параллельным обращениям.

Связь 182 · AISAFE2-WEAK-064 → AISAFE2-COMP-016

Адаптер внешнего AI-сервиса не ограничивает ожидание, повторы и параллельные обращения и не выбирает заранее заданный режим при отказе провайдера.

Связь 183 · AISAFE2-WEAK-064 → AISAFE2-COMP-017

Серверная часть не изолирует отказ внешней зависимости от полного пользовательского маршрута и распространяет задержку или ошибку на весь результат.

Связь 184 · AISAFE2-WEAK-065 → AISAFE2-COMP-014

Правила маршрутизации контура человеческой проверки направляют оператору лишние, слишком мелкие или повторяющиеся задачи без переиспользования эквивалентного решения.

Связь 185 · AISAFE2-WEAK-067 → AISAFE2-COMP-014

Интерфейс и рабочий процесс человеческой проверки не предоставляют оператору необходимый контекст, историю сходных решений и безопасные групповые действия.

Связь 186 · AISAFE2-WEAK-068 → AISAFE2-COMP-014

Контур человеческой проверки не измеряет качество решений и признаки утомления и не запускает ротацию или независимую выборочную проверку при перегрузке.

Связь 187 · AISAFE2-WEAK-069 → AISAFE2-COMP-014

Контур человеческой проверки не задаёт блокирование, приоритетную деградацию, замещение и восстановление при переполнении очереди или недоступности оператора.

Связь 188 · AISAFE2-WEAK-072 → AISAFE2-COMP-019

Контур исполнения политик не сопоставляет фактический поток с утверждённой доверенной зоной и не блокирует неизвестный или запрещённый переход данных.

Связь 189 · AISAFE2-WEAK-073 → AISAFE2-COMP-001

Пользовательский интерфейс или API передаёт результат без независимо сформированного материала, достаточного потребителю для содержательной проверки до применения.

Связь 190 · AISAFE2-WEAK-073 → AISAFE2-COMP-017

Серверный контур постобработки не формирует проверочный материал детерминированными средствами вне модели до передачи результата потребителю.

Связь 191 · AISAFE2-WEAK-074 → AISAFE2-COMP-002

Модель обрабатывает чувствительные элементы контекста и формирует ответ через общий выводной канал без технического разграничения доступа между контекстом и получателем.

Связь 192 · AISAFE2-WEAK-074 → AISAFE2-COMP-003

Контекст запроса объединяет элементы с разными ожиданиями конфиденциальности, но не задаёт принудительную границу их доступности через вывод модели.

Связь 193 · AISAFE2-WEAK-075 → AISAFE2-COMP-011

Автоматическое выделение вычислительных и сетевых ресурсов не имеет принудительного совокупного потолка, поэтому рост нагрузки продолжает увеличивать ресурсы и расходы.

Связь 194 · AISAFE2-WEAK-075 → AISAFE2-COMP-020

Квоты, пределы масштабирования и бюджетные пороги внешнего AI-провайдера не связаны с автоматическим ограничением совокупного потребления владельца.

Связь 195 · AISAFE2-WEAK-076 → AISAFE2-COMP-002

Разделяемая мощность инференса не резервирует и не ограничивает ресурсную долю арендатора, поэтому активность одного арендатора ухудшает задержки и завершение запросов других.

Связь 196 · AISAFE2-WEAK-076 → AISAFE2-COMP-011

Планирование общих вычислительных и сетевых ресурсов не обеспечивает изоляцию производительности между арендаторами при конкуренции за насыщенный ресурс.

Связь 197 · AISAFE2-WEAK-076 → AISAFE2-COMP-018

Общие очереди и рабочие каналы не разделяют ресурсные доли арендаторов, поэтому нагрузка одного субъекта задерживает задания остальных.

Связь 198 · AISAFE2-WEAK-077 → AISAFE2-COMP-011

Ёмкость и механизм наращивания вычислительной среды рассчитаны без достаточного запаса и своевременной реакции на ожидаемую пиковую нагрузку.

Связь 199 · AISAFE2-WEAK-077 → AISAFE2-COMP-020

Квоты и параметры масштабирования внешнего AI-провайдера не согласованы с профилем пикового спроса и целевым временем ответа.

Связь 200 · AISAFE2-WEAK-078 → AISAFE2-COMP-001

Входная граница продолжает принимать запросы сверх доступной ёмкости вместо ограничения или отклонения избыточной нагрузки.

Связь 201 · AISAFE2-WEAK-078 → AISAFE2-COMP-011

Среда выполнения не локализует насыщение ресурсов, поэтому перегрузка одного участка распространяется на остальные компоненты обработки.

Связь 202 · AISAFE2-WEAK-078 → AISAFE2-COMP-017

Прикладной контур не применяет согласованные правила допуска, сброса нагрузки и приоритета критичных потребителей при перегрузке.

Связь 203 · AISAFE2-WEAK-078 → AISAFE2-COMP-018

Очереди не имеют жёсткой границы и правил приоритизации, поэтому при спросе выше ёмкости растут без ограничения и задерживают всех потребителей.

Связь 204 · AISAFE2-WEAK-079 → AISAFE2-COMP-005

Для внешнего инструмента, MCP-сервера или коннектора не подготовлены заменяющий сервис, переносимая конфигурация и проверенный порядок переключения.

Связь 205 · AISAFE2-WEAK-079 → AISAFE2-COMP-011

Вычислительная, сетевая или площадочная среда зависит от одного поставщика без подготовленной альтернативы и переносимого способа размещения.

Связь 206 · AISAFE2-WEAK-079 → AISAFE2-COMP-016

Адаптер связан с единственным внешним AI-сервисом и не имеет подготовленного совместимого маршрута переключения на замену.

Связь 207 · AISAFE2-WEAK-079 → AISAFE2-COMP-020

Контур управления провайдером не содержит утверждённой альтернативы, переносимой конфигурации и готового порядка замещения единственного поставщика.

Связь 208 · AISAFE2-WEAK-080 → AISAFE2-COMP-002

Для произвольного смыслового свойства генеративного вывода нельзя задать полную машинно-проверяемую границу, охватывающую все семантически эквивалентные формы выражения.

Связь 209 · AISAFE2-WEAK-080 → AISAFE2-COMP-019

Контур исполнения политик распознаёт только формализованные или известные формы вывода и не обеспечивает полноту контроля произвольного смыслового свойства при парафразе, переводе или косвенном выражении.

Связь 210 · AISAFE2-WEAK-081 → AISAFE2-COMP-001

Внешний интерфейс или API принимает защищённую операцию без проверки идентичности субъекта, хотя её контракт требует установить инициатора до выполнения.

Ограничение: Только для операций, требующих установленной идентичности инициатора; утверждённые анонимные маршруты не входят в область связи.

Связь 211 · AISAFE2-WEAK-081 → AISAFE2-COMP-017

Прямой маршрут к внутреннему сервису является поверхностью обхода аутентификации внешнего входа.

Связь 212 · AISAFE2-WEAK-082 → AISAFE2-COMP-012

Контур идентификации является поверхностью недостаточно стойкой аутентификации привилегированного человека.

Связь 213 · AISAFE2-WEAK-082 → AISAFE2-COMP-020

Контур управления внешним AI-провайдером является целевой административной поверхностью, на которой недостаточно стойкая аутентификация привилегированного человека допускает вход или повышение полномочий.

Связь 214 · AISAFE2-WEAK-083 → AISAFE2-COMP-002

Внутренняя точка инференса принимает соединение из недоверенной или избыточно широкой сети в обход утверждённого прикладного маршрута.

Связь 215 · AISAFE2-WEAK-083 → AISAFE2-COMP-005

Служебный интерфейс инструмента или MCP-сервера сетево достижим из источника, которому он не предназначен.

Связь 216 · AISAFE2-WEAK-083 → AISAFE2-COMP-010

Административный интерфейс реестра модели или хранилища артефактов доступен вне утверждённого административного маршрута.

Связь 217 · AISAFE2-WEAK-083 → AISAFE2-COMP-011

Эффективная сеть среды выполнения не блокирует соединение между неутверждёнными источником и служебным получателем.

Связь 218 · AISAFE2-WEAK-084 → AISAFE2-COMP-002

Для сервера инференса нет применимого эталона безопасных настроек либо его эффективная конфигурация не проверяется.

Связь 219 · AISAFE2-WEAK-084 → AISAFE2-COMP-004

Агентная среда развёрнута без утверждённого эталона функций, привилегий процесса и служебных режимов.

Связь 220 · AISAFE2-WEAK-084 → AISAFE2-COMP-005

Сервер инструмента или коннектора сохраняет небезопасную настройку по умолчанию либо неизвестное отклонение от эталона.

Связь 221 · AISAFE2-WEAK-084 → AISAFE2-COMP-010

Реестр и хранилище артефактов не имеют проверяемого эталона настроек для точных версии и среды.

Связь 222 · AISAFE2-WEAK-084 → AISAFE2-COMP-011

Среда выполнения не сравнивает эффективные сетевые, процессные и административные настройки с утверждённым эталоном.

Связь 223 · AISAFE2-WEAK-084 → AISAFE2-COMP-013

Служебные режимы и интерфейсы телеметрии не ограничены применимым эталоном конфигурации.

Связь 224 · AISAFE2-WEAK-084 → AISAFE2-COMP-015

Контур выпуска не блокирует развёртывание без результата проверки точной конфигурации по применимому эталону.

Связь 225 · AISAFE2-WEAK-084 → AISAFE2-COMP-023

Подсистема формирования базы знаний и её хранилища работают без проверяемого эталона защищённых настроек.

Связь 226 · AISAFE2-WEAK-085 → AISAFE2-COMP-008

Для точной версии обучающего или оценочного набора нет независимой неизменяемой и проверенно восстанавливаемой копии.

Связь 227 · AISAFE2-WEAK-085 → AISAFE2-COMP-010

Рабочие полномочия могут уничтожить основной артефакт модели и все его восстановимые версии.

Связь 228 · AISAFE2-WEAK-085 → AISAFE2-COMP-015

Критичная конфигурация или системная инструкция в управляемом пути изменений не имеет защищённой версии вне рабочих полномочий.

Связь 229 · AISAFE2-WEAK-085 → AISAFE2-COMP-023

Корпус или индекс базы знаний не имеет независимой неизменяемой версии, восстановление которой проверено.

Связь 230 · AISAFE2-WEAK-086 → AISAFE2-COMP-023

Конвейер разбиения и индексирования теряет либо не синхронизирует атрибуты доступа исходного материала у фрагмента и производного объекта.

Связь 231 · AISAFE2-WEAK-087 → AISAFE2-COMP-006

Запросное извлечение использует векторные представления, чей класс обработки слабее класса исходного материала.

Связь 232 · AISAFE2-WEAK-087 → AISAFE2-COMP-016

Внешнему сервису вычисления представлений передаются векторы без наследованных требований обработки источника.

Связь 233 · AISAFE2-WEAK-087 → AISAFE2-COMP-023

При вычислении, хранении, экспорте или резервировании индекса представление не наследует чувствительность источника.

Связь 234 · AISAFE2-WEAK-088 → AISAFE2-COMP-015

Контур допуска и пересмотра выпуска не требует утверждённых этических критериев и связанных с точной версией результатов оценки.

Связь 235 · AISAFE2-WEAK-089 → AISAFE2-COMP-004

Ядро агента предоставляет модели набор возможностей, не суженный до текущих задачи, роли и этапа.

Связь 236 · AISAFE2-WEAK-089 → AISAFE2-COMP-005

Реестр инструментов допускает функции и операции сверх необходимого эффекта текущей задачи.

Связь 237 · AISAFE2-WEAK-089 → AISAFE2-COMP-017

Прикладная логика не выводит эффективный набор инструментов из роли, задачи и состояния сценария.

Связь 238 · AISAFE2-WEAK-089 → AISAFE2-COMP-019

Контур исполнения политик не блокирует выбор инструмента или операции вне модели разрешённых для текущего шага возможностей.

Связь 239 · AISAFE2-WEAK-090 → AISAFE2-COMP-005

Инструмент, парсер или сгенерированный код исполняется с доступом к ресурсам хоста сверх явно выданной области.

Связь 240 · AISAFE2-WEAK-090 → AISAFE2-COMP-011

Среда выполнения не ограничивает процессные, файловые, сетевые, секретные, межарендаторные и ресурсные границы исполнителя.

Связь 241 · AISAFE2-WEAK-091 → AISAFE2-COMP-002

Процесс инференса загружает исполняемый формат модели с полномочиями рабочего сервиса.

Связь 242 · AISAFE2-WEAK-091 → AISAFE2-COMP-009

Конвейер обучения или дообучения десериализует артефакт либо сопровождающий код как доверенные данные.

Связь 243 · AISAFE2-WEAK-091 → AISAFE2-COMP-010

Реестр допускает формат модели, загрузка которого может выполнить код, несмотря на отдельно подтверждённую идентичность артефакта.

Связь 244 · AISAFE2-WEAK-092 → AISAFE2-COMP-004

Ядро агента повторно использует делегированное пользовательское полномочие вне исходных задачи, получателя или срока.

Связь 245 · AISAFE2-WEAK-092 → AISAFE2-COMP-005

Коннектор получает полномочие без принудительных ограничений аудитории, операции, ресурса и получателя.

Связь 246 · AISAFE2-WEAK-092 → AISAFE2-COMP-012

Контур авторизации выдаёт или принимает делегированный грант без сужения, срока и проверяемого отзыва.

Связь 247 · AISAFE2-WEAK-093 → AISAFE2-COMP-017

Прикладная серверная логика теряет сквозную связь арендатора с операцией между измерением потребления и формированием начисления.

Связь 248 · AISAFE2-WEAK-094 → AISAFE2-COMP-012

Контур идентичности и секретов не удерживает исходное пользовательское учётное значение в минимальной области хранения и доставки, ограниченной задачей.

Связь 249 · AISAFE2-WEAK-095 → AISAFE2-COMP-005

Коннектор к платёжному получателю принимает делегированное полномочие без принудительно проверяемых ограничений суммы, частоты, назначения и срока.

Связь 250 · AISAFE2-WEAK-001 → AISAFE2-IMP-001

Избыточное доверие пользователя к AI-выводу может привести к выполнению пользователем предложенного системой вредоносного действия, потому что уверенная форма ответа и приписываемая системе компетентность воспринимаются как признак достоверности, и пользователь запускает предложенный код, команду или настройку без независимой проверки.

Связь 251 · AISAFE2-WEAK-001 → AISAFE2-IMP-002

Избыточное доверие пользователя к AI-выводу может привести к ущербу от решения, принятого на основе результата AI-системы, потому что пользователь доверяет системе как эксперту и принимает решение по уверенно сформулированному ответу без независимой перепроверки, превращая ошибку результата в ошибку решения.

Связь 252 · AISAFE2-WEAK-001 → AISAFE2-IMP-003

Избыточное доверие пользователя к AI-выводу может привести к вовлечению пользователя в злонамеренную схему, потому что пользователь тем охотнее следует указаниям, что доверяет уверенным ответам системы, и при воздействии злоумышленника легче вовлекается в мошеннические схемы.

Связь 253 · AISAFE2-WEAK-001 → AISAFE2-IMP-004

Избыточное доверие пользователя к AI-выводу может привести к передаче потребителем чувствительных данных, потому что пользователь доверяет системе как помощнику и добровольно передаёт секреты и персональные данные, когда их запрос выглядит естественной частью диалога.

Связь 254 · AISAFE2-WEAK-002 → AISAFE2-IMP-001

Отсутствие привязки утверждений AI-ответа к источникам может привести к выполнению пользователем предложенного системой вредоносного действия, потому что без проверяемой связи утверждений с фактически использованными источниками пользователь не может восстановить основание рекомендации и действует по непроверяемому выводу, принятому за факт.

Ограничение: Только если отсутствие проверяемого источника повлияло на доверие к совету, пользователь выполнил предложенное действие и понёс вред

Связь 255 · AISAFE2-WEAK-002 → AISAFE2-IMP-002

Отсутствие привязки утверждений AI-ответа к источникам может привести к ущербу от решения, принятого на основе результата AI-системы, потому что утверждения ответа не сопровождаются надёжными ссылками на фактически использованные источники и статусом подтверждения, по которым решение можно было бы перепроверить.

Ограничение: Только если отсутствие проверяемого источника стало существенным для исполненного решения по AI-результату, после которого наступил вред

Связь 256 · AISAFE2-WEAK-003 → AISAFE2-IMP-001

Недостаточная проверка AI-вывода на границе доверия может привести к выполнению пользователем предложенного системой вредоносного действия, потому что между формированием вывода и его выдачей нет обязательного рубежа, проверяющего допустимость содержания и отклоняющего недопустимый результат, и опасное предложение беспрепятственно достигает пользователя.

Ограничение: Только если опасный совет прошёл непроверенную границу выдачи, пользователь его выполнил и понёс вред

Связь 257 · AISAFE2-WEAK-003 → AISAFE2-IMP-003

Недостаточная проверка AI-вывода на границе доверия может привести к вовлечению пользователя в злонамеренную схему, потому что система не отфильтровывает убеждающий обманный текст на границе выдачи, и склоняющее к действиям в интересах злоумышленника содержимое беспрепятственно достигает пользователя.

Ограничение: Только при подтверждённом изменении поведения пользователя в интересах злонамеренной схемы; создание вредоносного текста без вовлечения человека не входит.

Связь 258 · AISAFE2-WEAK-003 → AISAFE2-IMP-005

Недостаточная проверка AI-вывода на границе доверия может привести к нарушению этических норм по отношению к пользователю, потому что ответ не проходит обязательную проверку допустимости содержания перед выдачей, и оскорбительное или недопустимое обращение беспрепятственно достигает человека.

Ограничение: Только если пропущенный без проверки AI-вывод вызвал ответ, решение или обращение с пользователем, нарушившие утверждённые для сценария этические нормы и причинившие ему вред

Связь 259 · AISAFE2-WEAK-004 → AISAFE2-IMP-003

Недостаточность санитизации или изоляции вывода модели может привести к вовлечению пользователя в злонамеренную схему, потому что не нейтрализованная активная разметка из вывода модели исполняется в браузере пользователя (XSS) и в доверенном интерфейсе системы показывает ему обманное содержимое, формы и ссылки, уводящие в мошеннический сценарий.

Ограничение: Только если активный вывод создал обманное содержимое в доверенном интерфейсе и пользователь после этого вступил в злонамеренную схему; одно срабатывание XSS без вовлечения недостаточно

Связь 260 · AISAFE2-WEAK-004 → AISAFE2-IMP-004

Неизолированный активный вывод создаёт обманную форму в доверенном интерфейсе; пользователь фактически вводит чувствительное значение, которое затем попадает в модель либо инструмент без допустимого основания.

Ограничение: Только когда пользователь действительно вводит чувствительное значение в созданную активным выводом форму и это значение входит в контекст модели или инструмента без допустимого основания. Непосредственная XSS-эксфильтрация остаётся ветвью I024, а обман без передачи значения — I003.

Связь 261 · AISAFE2-WEAK-004 → AISAFE2-IMP-007

Недостаточность санитизации или изоляции вывода модели может привести к обходу ограничений и предусмотренной логики агента, потому что вывод модели, разбираемый оркестратором как структурированный формат, шаблон или конфигурация, без нейтрализации позволяет внедрённым конструкциям вмешаться в дальнейший ход сценария.

Связь 262 · AISAFE2-WEAK-004 → AISAFE2-IMP-008

Недостаточность санитизации или изоляции вывода модели может привести к неожидаемому поведению слоя инструментов, потому что сформированное моделью содержимое попадает во входные контексты инструментов и интерпретаторов без обработки по правилам этих контекстов, и внедрённые конструкции заставляют исправный компонент выполнить непредусмотренную операцию.

Связь 263 · AISAFE2-WEAK-004 → AISAFE2-IMP-015

Недостаточность санитизации или изоляции вывода модели может привести к несанкционированному чтению или изменению хранимых данных, потому что внедрённые в вывод конструкции запросов и путей (SQL-инъекция, обход файловых путей) исполняются хранилищем и файловой системой с правами приложения, открывая чтение и изменение данных вне полномочий инициатора.

Связь 264 · AISAFE2-WEAK-004 → AISAFE2-IMP-016

Сформированный моделью вывод попадает в командный или иной исполняемый контекст инструмента без обработки по правилам этого контекста и выполняется с полномочиями принимающего сервиса; это позволяет инициатору вызвать инфраструктурную функцию, на которую у него нет права.

Ограничение: Только при фактическом вызове неразрешённой инициатору инфраструктурной функции; отображение разметки или исполнение в браузере само по себе этого не доказывает.

Связь 265 · AISAFE2-WEAK-004 → AISAFE2-IMP-024

Если ответ модели вставляет чувствительное значение в адрес автоматически загружаемого ресурса или вызывает переход по такому адресу, небезопасная отрисовка может отправить значение внешнему узлу.

Ограничение: Только если клиент действительно выполняет сетевой запрос к получателю, которому значение передавать нельзя. Обычная ссылка без перехода и отображение текста не доказывают утечку.

Связь 266 · AISAFE2-WEAK-005 → AISAFE2-IMP-001

Недостаточность маркировки и предупреждений в интерфейсе может привести к выполнению пользователем предложенного системой вредоносного действия, потому что интерфейс не помечает результат как сгенерированный и требующий проверки и не создаёт точку осознанного решения перед значимым шагом, и человек запускает непроверенное предложение как готовое.

Ограничение: Только если отсутствие маркировки или предупреждения повлияло на выполнение опасного совета и действие причинило вред

Связь 267 · AISAFE2-WEAK-005 → AISAFE2-IMP-002

Недостаточность маркировки и предупреждений в интерфейсе может привести к ущербу от решения, принятого на основе результата AI-системы, потому что представление ответа не предупреждает об ограничениях, статусе проверки и рисках дальнейшего использования, и человек утверждает и передаёт непроверенный результат как готовый.

Ограничение: Только если неуказанный статус проверки или риск повлиял на исполненное решение и наступило вредное последствие

Связь 268 · AISAFE2-WEAK-005 → AISAFE2-IMP-003

Если злоумышленник подсовывает AI-результат как подтверждённое основание, отсутствие ясной маркировки и статуса проверки повышает доверие пользователя к непроверенным сведениям и может вовлечь его в обман.

Ограничение: Только при подтверждённом изменении поведения пользователя в интересах злонамеренной схемы; создание вредоносного текста без вовлечения человека не входит.

Связь 269 · AISAFE2-WEAK-005 → AISAFE2-IMP-004

Недостаточная маркировка и предупреждения в интерфейсе могут привести к передаче потребителем чувствительных данных: пользователь не видит границ сценария обработки и передаёт секреты или персональные сведения, не понимая, что они не нужны или не разрешены для этого сценария.

Ограничение: Только когда чувствительное значение из пользовательского ввода или файла действительно попадает в контекст без допустимого основания.

Связь 270 · AISAFE2-WEAK-006 → AISAFE2-IMP-001

Отсутствие раскрытия факта взаимодействия с AI может привести к выполнению пользователем предложенного системой вредоносного действия, потому что без явного сигнала о взаимодействии с AI пользователь относится к сгенерированному содержимому как к проверенному людьми и неверно калибрует доверие при принятии решения.

Ограничение: Только если пользователь не знал об AI-происхождении совета, из-за этого доверился ему и выполнил опасное действие с вредом

Связь 271 · AISAFE2-WEAK-006 → AISAFE2-IMP-002

Отсутствие раскрытия факта взаимодействия с AI может привести к ущербу от решения, принятого на основе результата AI-системы, потому что без явного сигнала о взаимодействии с AI пользователь относится к сгенерированному содержимому как к проверенному людьми и неверно калибрует доверие при принятии решения.

Ограничение: Только если сокрытие AI-происхождения результата повлияло на исполненное решение и наступил вред

Связь 272 · AISAFE2-WEAK-006 → AISAFE2-IMP-003

Отсутствие раскрытия факта взаимодействия с AI может привести к вовлечению пользователя в злонамеренную схему, потому что пользователь не осознаёт, что общается с управляемой AI-системой, и злоумышленнику проще выдавать агента за человека в схемах обмана и социальной инженерии.

Ограничение: Только если скрытое AI-происхождение общения способствовало вовлечению человека в конкретную злонамеренную схему с наблюдаемым изменением поведения

Связь 273 · AISAFE2-WEAK-006 → AISAFE2-IMP-004

Отсутствие раскрытия факта взаимодействия с AI может привести к передаче потребителем чувствительных данных, потому что не зная, что собеседник — управляемый AI-агент, пользователь сообщает ему сведения, которые не раскрыл бы в распознанном сценарии социальной инженерии злоумышленника.

Ограничение: Только если сокрытие AI-статуса побудило пользователя передать секреты или персональные данные этому агенту без допустимого основания

Связь 274 · AISAFE2-WEAK-007 → AISAFE2-IMP-003

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к вовлечению пользователя в злонамеренную схему, потому что инструкции злоумышленника, внедрённые в документы и ответы инструментов, исполняются моделью как команды и позволяют управлять ответами системы, склоняющими пользователя к нужным атакующему действиям.

Ограничение: Только когда человек фактически вовлечён в конкретную злонамеренную схему и его наблюдаемое поведение изменилось; одного обманного сообщения недостаточно.

Связь 275 · AISAFE2-WEAK-007 → AISAFE2-IMP-004

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к передаче потребителем чувствительных данных, потому что внедрённые в контекст инструкции атакующего заставляют систему запрашивать у пользователя секреты или направлять его в указанный внешний канал под видом естественного диалога.

Ограничение: Только если внедрённая инструкция побудила пользователя передать чувствительные сведения модели, инструменту или через систему внешнему адресату без допустимого основания

Связь 276 · AISAFE2-WEAK-007 → AISAFE2-IMP-006

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к неожидаемым для разработчика ответам модели, потому что любой текст, попавший в контекст, потенциально исполняется как инструкция, и недоверенное содержимое документов и ответов инструментов преодолевает заложенные ограничения, меняя поведение модели.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 277 · AISAFE2-WEAK-007 → AISAFE2-IMP-007

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к обходу ограничений и предусмотренной логики агента, потому что инструкция, проведённая через обрабатываемый документ или ответ инструмента, исполняется моделью как команда и меняет порядок действий, выбор инструментов и цель агента без изменения кода.

Связь 278 · AISAFE2-WEAK-007 → AISAFE2-IMP-009

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что недоверенный текст контекста меняет объём раскрытия его содержимого, и внедрённая инструкция заставляет модель выдать закрытые данные, полученные от инструментов.

Связь 279 · AISAFE2-WEAK-007 → AISAFE2-IMP-010

Отсутствие границы между инструкциями и данными во входных данных для модели может привести к раскрытию системного промпта и инструкций инструментов, потому что внедрённая в контекст инструкция меняет объём раскрытия его содержимого, и модель пересказывает системные инструкции по команде из недоверенного текста.

Связь 280 · AISAFE2-WEAK-007 → AISAFE2-IMP-021

Внедрённая через недоверенное содержимое инструкция может удерживать агентную обработку одного запроса в повторяющихся шагах, если завершение зависит от решения модели и жёсткий предел шагов отсутствует.

Ограничение: Только для повторяющегося выполнения одного запроса или цепочки, которое не завершается по штатному условию.

Связь 281 · AISAFE2-WEAK-008 → AISAFE2-IMP-005

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к нарушению этических норм по отношению к пользователю, потому что внутри универсальной модели нет детерминированной границы между разрешённым и запрещённым, и без внешнего контроля результата она способна сама сформировать недопустимый ответ.

Ограничение: Только если нежелательное использование способности модели проявилось в ответе, решении или обращении с пользователем, нарушивших утверждённые для сценария этические нормы и причинивших ему вред

Связь 282 · AISAFE2-WEAK-008 → AISAFE2-IMP-006

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к неожидаемым для разработчика ответам модели, потому что ограничения модели формируются обучением и не образуют строгого правила, и в пограничных или намеренно подобранных контекстах она выполняет недопустимую задачу или формирует вредный результат.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 283 · AISAFE2-WEAK-008 → AISAFE2-IMP-007

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к обходу ограничений и предусмотренной логики агента, потому что ограничения агента удерживаются лишь вероятностным поведением модели и преодолеваются подобранным запросом, а недопустимое действие без внешнего детерминированного контроля результата исключить нельзя.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

Связь 284 · AISAFE2-WEAK-008 → AISAFE2-IMP-009

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что нераскрытие полученных от инструментов сведений держится только на вероятностном поведении модели и преодолевается подобранным запросом.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

Связь 285 · AISAFE2-WEAK-008 → AISAFE2-IMP-010

Отсутствие детерминированной внутренней границы безопасности в возможностях модели может привести к раскрытию системного промпта и инструкций инструментов, потому что запрет на их выдачу не является детерминированной границей возможностей модели и обходится специально построенным запросом.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 286 · AISAFE2-WEAK-009 → AISAFE2-IMP-006

Ненадёжное соблюдение моделью системных инструкций может привести к неожидаемым для разработчика ответам модели, потому что старшинство системных директив не является инвариантом поведения, конфликт инструкций разрешается вероятностно, и ограничения преодолеваются как подобранными запросами, так и самопроизвольно.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 287 · AISAFE2-WEAK-009 → AISAFE2-IMP-007

Ненадёжное соблюдение моделью системных инструкций может привести к обходу ограничений и предусмотренной логики агента, потому что соблюдение системных директив вероятностно и зависит от всего контекста, и при конфликте инструкций агент отклоняется от предусмотренной логики, обходя заложенные правила.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

Связь 288 · AISAFE2-WEAK-009 → AISAFE2-IMP-009

Ненадёжное соблюдение моделью системных инструкций может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что указания о нераскрытии соблюдаются моделью вероятностно, а не как принуждение, и защищаемые сведения из контекста раскрываются вопреки системным директивам.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

Связь 289 · AISAFE2-WEAK-009 → AISAFE2-IMP-010

Ненадёжное соблюдение моделью системных инструкций может привести к раскрытию системного промпта и инструкций инструментов, потому что запрет на раскрытие — лишь одна из конкурирующих директив, соблюдаемая вероятностно, и подобранный запрос преодолевает его вопреки ожидаемому старшинству системного промпта.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 290 · AISAFE2-WEAK-010 → AISAFE2-IMP-005

Преодоление обученных ограничений содержимым контекста может привести к нарушению этических норм по отношению к пользователю, потому что накопленные в контексте примеры, назначенная роль и сложившийся шаблон беседы повышают вероятность продолжения недопустимого шаблона, и обученные ограничения преодолеваются в результате диалога.

Ограничение: Только если контекстный обход обученного ограничения привёл к ответу, решению или обращению с пользователем, нарушившим утверждённые для сценария этические нормы и причинившим ему вред

Связь 291 · AISAFE2-WEAK-010 → AISAFE2-IMP-006

Преодоление обученных ограничений содержимым контекста может привести к неожидаемым для разработчика ответам модели, потому что содержимое окна контекста меняет поведение модели без изменения весов, и подобранным запросом снимаются ограничения на запрещённые темы.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 292 · AISAFE2-WEAK-010 → AISAFE2-IMP-007

Преодоление обученных ограничений содержимым контекста может привести к обходу ограничений и предусмотренной логики агента, потому что постепенно выстроенный контекст переопределяет фактическое поведение агента, и обученные ограничения перестают удерживать сценарий его действий.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

Связь 293 · AISAFE2-WEAK-010 → AISAFE2-IMP-009

Преодоление обученных ограничений содержимым контекста может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что сложившийся в контексте шаблон согласия повышает вероятность выдачи полученных от инструментов защищаемых сведений вопреки обученному запрету.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

Связь 294 · AISAFE2-WEAK-010 → AISAFE2-IMP-010

Преодоление обученных ограничений содержимым контекста может привести к раскрытию системного промпта и инструкций инструментов, потому что накопленный контекст постепенно подводит модель к пересказу защищённых инструкций, в котором при прямом запросе она отказала бы.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 295 · AISAFE2-WEAK-011 → AISAFE2-IMP-005

Слабая настройка модели на безопасное поведение может привести к нарушению этических норм по отношению к пользователю, потому что безопасное поведение обобщается хуже способностей модели, и в необычных формулировках, языках и форматах ограничитель не срабатывает, пропуская недопустимое содержание пользователю.

Ограничение: Только если сбой настройки безопасного поведения при смене языка, формата или условий привёл к ответу, решению или обращению с пользователем, нарушившим утверждённые для сценария этические нормы и причинившим ему вред

Связь 296 · AISAFE2-WEAK-011 → AISAFE2-IMP-006

Слабая настройка модели на безопасное поведение может привести к неожиданным для разработчика ответам модели, потому что данные дообучения безопасному поведению покрывают более узкое распределение, чем способности модели, и часть обходов ограничений (редкие языки, кодировки, форматы) обнаруживается уже после выпуска.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 297 · AISAFE2-WEAK-011 → AISAFE2-IMP-007

Слабая настройка модели на безопасное поведение может привести к обходу ограничений и предусмотренной логики агента, потому что смена формы запроса (язык, кодировка, формат) выводит задачу из покрытого настройкой распределения, и заложенные ограничения поведения агента не срабатывают.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

Связь 298 · AISAFE2-WEAK-011 → AISAFE2-IMP-009

Слабая настройка модели на безопасное поведение может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что ограничитель, обученный на узком распределении, не срабатывает на просьбах пересказать, перевести или закодировать сведения, и защита нераскрытия обходится сменой формы запроса.

Ограничение: Только если сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; одно чтение источника не означает раскрытия

Связь 299 · AISAFE2-WEAK-011 → AISAFE2-IMP-010

Слабая настройка модели на безопасное поведение может привести к раскрытию системного промпта и инструкций инструментов, потому что запрет на их выдачу, выученный на типовых формулировках, обходится нетипичной формой запроса — переводом, кодированием, косвенным пересказом.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 300 · AISAFE2-WEAK-012 → AISAFE2-IMP-003

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к вовлечению пользователя в злонамеренную схему, потому что подменённый или заражённый компонент цепочки поставки действует изнутри доверенного контура и позволяет злоумышленнику управлять ответами и поведением системы.

Ограничение: Только если скомпрометированный компонент сформировал или передал убеждающее сообщение и человек изменил поведение в интересах злонамеренной схемы

Связь 301 · AISAFE2-WEAK-012 → AISAFE2-IMP-006

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к неожидаемым для разработчика ответам модели, потому что компоненты с заранее встроенным поведением — заражённые модели, отравленные наборы данных — включаются без проверки происхождения и целостности и демонстрируют неожиданные свойства изнутри доверенного контура.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 302 · AISAFE2-WEAK-012 → AISAFE2-IMP-007

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к обходу ограничений и предусмотренной логики агента, потому что скомпрометированный компонент, встроенный в контур агента, меняет его решения и действия изнутри в обход задуманной логики.

Ограничение: Только если скомпрометированный AI-компонент фактически изменил план, действие или состояние Agent Core; его присутствия в цепочке поставки недостаточно

Связь 303 · AISAFE2-WEAK-012 → AISAFE2-IMP-008

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к неожидаемому поведению слоя инструментов, потому что инструменты и их описания подключаются и обновляются без проверки происхождения и целостности (включая динамическую регистрацию через MCP), и в составе доступных появляется незаявленный или заражённый компонент.

Ограничение: Только когда инструмент, RAG или сторонний агент выдаёт непредусмотренный результат либо выполняет действие, достигшее следующего компонента или внешней системы.

Связь 304 · AISAFE2-WEAK-012 → AISAFE2-IMP-011

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к компрометации базы знаний, потому что скомпрометированный компонент цепочки поставки действует изнутри доверенного контура и может передавать содержимое базы знаний поставщику и третьим сторонам.

Ограничение: Только если скомпрометированный компонент передал неуполномоченному получателю существенную часть базы знаний, её метаданных или индекса

Связь 305 · AISAFE2-WEAK-012 → AISAFE2-IMP-012

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к краже модели, потому что внедрённый в цепочку поставки код действует изнутри контура с доступом к весам и артефактам модели и передаёт их атакующему.

Ограничение: Только если внедрённый через цепочку поставки компонент фактически передал веса или связанные артефакты субъекту без разрешения

Связь 306 · AISAFE2-WEAK-012 → AISAFE2-IMP-014

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к раскрытию сведений о технологиях, алгоритмах и источниках данных системы, потому что скомпрометированный компонент (например, инструмент загрузки наборов данных) сообщает сведения о системе и её данных наружу — поставщику или третьим сторонам.

Ограничение: Только при фактической передаче наружу сведений об архитектуре, технологиях, алгоритмах, конфигурации или источниках данных; внутреннего доступа недостаточно.

Связь 307 · AISAFE2-WEAK-012 → AISAFE2-IMP-015

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к несанкционированному чтению или изменению хранимых данных, потому что заражённый компонент исполняется с правами доверенной системы и читает или изменяет достижимые ей хранимые данные в обход контроля полномочий.

Ограничение: Только если заражённый компонент с правами доверенной системы фактически прочитал или изменил хранимый объект вне полномочий атакующего

Связь 308 · AISAFE2-WEAK-012 → AISAFE2-IMP-016

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что закладка в компоненте вызывает функции инфраструктуры от имени системы, на которые инициатор воздействия полномочий не имеет.

Ограничение: Только при успешном вызове инфраструктурной функции за пределами полномочий субъекта или с полномочиями скомпрометированного компонента.

Связь 309 · AISAFE2-WEAK-012 → AISAFE2-IMP-024

Недоверенная или скомпрометированная цепочка поставки AI-компонентов может привести к утечке чувствительных данных в недоверенный контур, потому что скомпрометированный или недоверенный компонент передаёт обрабатываемые данные поставщику и третьим сторонам изнутри доверенного контура.

Ограничение: Только при фактическом пересечении чувствительными данными границы доверия без разрешённого перехода к точному получателю или среде.

Связь 310 · AISAFE2-WEAK-013 → AISAFE2-IMP-006

Набор без восстановимой трассы источника и преобразований проходит допуск до проверки переходов; если в нём присутствует отравлённый материал, отсутствие трассы позволяет использовать точную изменённую версию в обучении, после чего модель формирует и передаёт неожиданный ответ.

Ограничение: Только когда конкретная отравлённая версия проходит допуск именно из-за отсутствия восстановимой трассы, используется в обучении и приводит к фактически переданному неожиданному ответу; отсутствие такой трассируемости без этого пути недостаточно.

Связь 311 · AISAFE2-WEAK-014 → AISAFE2-IMP-003

Неконтролируемый допуск материала в источники знаний и контуры обратной связи может привести к вовлечению пользователя в злонамеренную схему, потому что непроверенный материал, внедрённый атакующим в базы знаний и контуры обратной связи, сохраняется и воспроизводится, позволяя управлять ответами и действиями системы в его интересах.

Связь 312 · AISAFE2-WEAK-014 → AISAFE2-IMP-006

Если ошибки или вредоносные примеры без контроля попадают в обучающие данные либо контур обратной связи, последующее обучение может изменить поведение модели и вызвать неожиданные ответы на рабочих запросах.

Ограничение: Загрязнение RAG-корпуса без изменения модели относится к ответу системы и рассматривается отдельно; здесь требуется фактический путь в обучение.

Связь 313 · AISAFE2-WEAK-014 → AISAFE2-IMP-007

Неконтролируемый допуск материала в источники знаний и контуры обратной связи может привести к обходу ограничений и предусмотренной логики агента, потому что внедрённый в базы знаний и долговременную память материал воспроизводится при последующих обращениях и позволяет управлять планами и действиями агента через содержимое источников.

Связь 314 · AISAFE2-WEAK-014 → AISAFE2-IMP-008

Неконтролируемый допуск материала в источники знаний и контуры обратной связи может привести к неожидаемому поведению слоя инструментов, потому что база знаний пополняется из источников, доступных злоумышленнику, без проверки происхождения и качества материала, и RAG возвращает подложные или отравленные данные.

Связь 315 · AISAFE2-WEAK-015 → AISAFE2-IMP-006

Переобученность модели может привести к неожиданным для разработчика ответам модели, потому что модель воспроизводит частные особенности обучающей выборки вместо устойчивых закономерностей, и на новых и редких входах качество систематически ниже измеренного при тестировании.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 316 · AISAFE2-WEAK-015 → AISAFE2-IMP-013

Наблюдаемая зависимость переобученной модели от знакомой выборки позволяет воспроизводимо определить участие чувствительной записи либо восстановить её атрибут посредством атаки определения принадлежности или проверки восстановления.

Ограничение: Только при подтверждённом воспроизведении, определении принадлежности, логическом выводе или восстановлении конкретной чувствительной записи либо её атрибута.

Связь 317 · AISAFE2-WEAK-016 → AISAFE2-IMP-006

Избыточные права машинного задания позволяют фактически изменить обучающие данные или артефакт вне назначенного пути; изменённая версия проходит обучение и выпуск и формирует неожиданный ответ.

Ограничение: Только когда точная версия модели формирует и передаёт ответ вне утверждённых границ содержания, формата, качества или поведения; одной возможности недостаточно.

Связь 318 · AISAFE2-WEAK-016 → AISAFE2-IMP-012

Слабый контроль доступа к обучающим данным, конвейерам и артефактам может привести к краже модели, потому что компрометация одного задания конвейера с избыточными сервисными полномочиями открывает хищение всех достижимых артефактов модели.

Связь 319 · AISAFE2-WEAK-016 → AISAFE2-IMP-015

Слабый контроль доступа к обучающим данным, конвейерам и артефактам может привести к несанкционированному чтению или изменению хранимых данных, потому что сервисные учётные записи конвейеров имеют избыточные полномочия, и компрометация одного задания позволяет читать и подменять достижимые наборы и модели, включая отравление будущих версий.

Связь 320 · AISAFE2-WEAK-016 → AISAFE2-IMP-031

Недостаток позволяет атакующему изменить или удалить датасет обучения, тестирования либо валидации вместе с доступными доверенными копиями, вследствие чего требуемая версия датасета становится невосстановимой.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения доступных доверенных копий.

Связь 321 · AISAFE2-WEAK-016 → AISAFE2-IMP-033

Недостаток позволяет атакующему изменить или удалить веса, базу знаний, индекс либо критичную конфигурацию вместе с доступными доверенными версиями, вследствие чего требуемый AI-артефакт становится невосстановимым.

Ограничение: Только для вызванной атакой невосстановимой утраты критичного AI-артефакта, кроме датасета, когда отсутствует доступная доверенная версия или воспроизводимый рецепт.

Связь 322 · AISAFE2-WEAK-017 → AISAFE2-IMP-006

Модель выбирается и допускается без проверяемых критериев содержания, формата и качества; выбранная точная версия затем фактически выдаёт результат вне предусмотренной границы.

Ограничение: Только когда точная версия модели формирует и передаёт ответ вне утверждённых границ содержания, формата, качества или поведения; одной возможности недостаточно.

Связь 323 · AISAFE2-WEAK-018 → AISAFE2-IMP-012

Наблюдаемость и воспроизводимость поведения модели для внешнего субъекта могут привести к краже модели, потому что штатный доступ к ответам раскрывает функциональную зависимость выхода от входа, и накопленные пары запрос — ответ позволяют обучить функциональную копию модели без доступа к весам.

Ограничение: Только если собранные через интерфейс пары запрос—ответ позволили подтвердить функциональное воспроизведение ценного поведения модели без разрешения

Связь 324 · AISAFE2-WEAK-018 → AISAFE2-IMP-014

Наблюдаемость и воспроизводимость поведения модели для внешнего субъекта могут привести к раскрытию сведений о технологиях, алгоритмах и источниках данных системы, потому что воспроизводимый сигнал вход — выход позволяет адаптивными сериями запросов определить семейство и версию модели, настройки, защитные механизмы и подключённые источники данных.

Ограничение: Только при подтверждённом неразрешённом получении закрытой информации о технологиях, архитектуре, алгоритмах или источниках данных.

Связь 325 · AISAFE2-WEAK-019 → AISAFE2-IMP-012

Возврат логитов или распределения вероятностей раскрывает более подробный сигнал о поведении модели и может облегчить построение функционального аналога при массовых запросах.

Ограничение: Только если раскрытые оценки модели помогли подтвердить функциональное воспроизведение её ценного поведения без разрешения

Связь 326 · AISAFE2-WEAK-019 → AISAFE2-IMP-013

Избыточное раскрытие оценок уверенности в ответе модели может привести к раскрытию чувствительных данных из обучающего датасета, потому что серия детальных оценок уверенности раскрывает различимость отдельных примеров и повышает точность вывода о принадлежности предъявленного примера обучающей выборке.

Ограничение: Только если раскрытые оценки позволили подтвердить участие конкретной чувствительной записи в обучении либо восстановить её атрибут; предположение по сигналу недостаточно

Связь 327 · AISAFE2-WEAK-022 → AISAFE2-IMP-011

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к компрометации базы знаний, потому что несогласованные по стадиям и средам правила доступа создают участки с избыточными правами и транзитивный доступ к данным и их конвейерам между контурами.

Связь 328 · AISAFE2-WEAK-022 → AISAFE2-IMP-012

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к краже модели, потому что транзитивный доступ между контурами и избыточные права на стыках стадий жизненного цикла позволяют считывать артефакты модели вне полномочий.

Связь 329 · AISAFE2-WEAK-022 → AISAFE2-IMP-015

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к несанкционированному чтению или изменению хранимых данных, потому что на стыках стадий и сред возникают участки с неопределёнными и избыточными правами и транзитивный доступ, допускающие чтение, изменение и удаление данных вне полномочий.

Связь 330 · AISAFE2-WEAK-022 → AISAFE2-IMP-016

Несогласованный контроль доступа к данным и артефактам на протяжении жизненного цикла AI-системы может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что транзитивный доступ между контурами позволяет субъекту одной стадии вызывать функции компонентов другой, на которые у него нет полномочий.

Ограничение: Только если транзитивное право одной стадии действительно позволило субъекту вызвать функцию другой стадии, на которую у него нет полномочий

Связь 331 · AISAFE2-WEAK-022 → AISAFE2-IMP-031

Недостаток позволяет атакующему изменить или удалить датасет обучения, тестирования либо валидации вместе с доступными доверенными копиями, вследствие чего требуемая версия датасета становится невосстановимой.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения доступных доверенных копий.

Связь 332 · AISAFE2-WEAK-022 → AISAFE2-IMP-033

Недостаток позволяет атакующему изменить или удалить веса, базу знаний, индекс либо критичную конфигурацию вместе с доступными доверенными версиями, вследствие чего требуемый AI-артефакт становится невосстановимым.

Ограничение: Только для вызванной атакой невосстановимой утраты критичного AI-артефакта, кроме датасета, когда отсутствует доступная доверенная версия или воспроизводимый рецепт.

Связь 333 · AISAFE2-WEAK-023 → AISAFE2-IMP-011

Незащищенное хранение данных и артефактов может привести к компрометации базы знаний, потому что база знаний и её индекс хранятся без криптографической защиты содержимого, и субъект с доступом к месту хранения считывает корпус напрямую и целиком.

Ограничение: Только если субъект без полномочий прочитал из незащищённого хранилища существенную часть корпуса или индекса; возможность доступа недостаточна

Связь 334 · AISAFE2-WEAK-023 → AISAFE2-IMP-012

Незащищенное хранение данных и артефактов может привести к краже модели, потому что веса и контрольные точки хранятся без криптографической защиты, и артефакты модели копируются целиком из незашифрованного хранилища.

Ограничение: Только если субъект без разрешения скопировал веса или связанные артефакты из незащищённого хранилища

Связь 335 · AISAFE2-WEAK-023 → AISAFE2-IMP-015

Незащищенное хранение данных и артефактов может привести к несанкционированному чтению или изменению хранимых данных, потому что без криптографической защиты хранения единственным барьером остаётся логическое разграничение, и субъект с доступом к месту хранения читает содержимое напрямую.

Ограничение: Только если субъект без полномочий фактически прочитал хранимый объект через незащищённое хранилище; возможность доступа недостаточна

Связь 336 · AISAFE2-WEAK-024 → AISAFE2-IMP-011

Незащищенные каналы передачи данных могут привести к компрометации базы знаний, потому что данные передаются между компонентами без шифрования транспорта и проверки подлинности сторон, и содержимое базы перехватывается в канале без наблюдаемых следов.

Ограничение: Только если через незащищённый канал прошёл и был перехвачен объём, позволяющий получить существенную часть базы знаний, её метаданных или индекса

Связь 337 · AISAFE2-WEAK-024 → AISAFE2-IMP-012

Незащищенные каналы передачи данных могут привести к краже модели, потому что артефакты передаются по каналам без транспортного шифрования, и веса модели перехватываются при передаче без наблюдаемых следов, включая этап поставки.

Ограничение: Только если веса или связанные артефакты были фактически перехвачены в незащищённом канале субъектом без разрешения

Связь 338 · AISAFE2-WEAK-024 → AISAFE2-IMP-015

Атакующий изменяет данные или артефакт в незащищённом канале, а конечная точка принимает и сохраняет подменённое содержимое, что реализует несанкционированное изменение хранимого объекта.

Ограничение: Только когда атакующий изменяет данные или артефакт в незащищённом канале, конечная точка принимает подмену и сохраняет её как хранимый объект вне полномочий атакующего; пассивное чтение трафика само по себе не образует AISAFE2-IMP-015.

Связь 339 · AISAFE2-WEAK-024 → AISAFE2-IMP-016

Атакующий подменяет или повторно воспроизводит запрос в незащищённом канале, а конечная точка исполняет инфраструктурную функцию вне полномочий атакующего.

Ограничение: Только если подменённый или повторный запрос был принят конечной точкой и вызвал функцию вне полномочий атакующего

Связь 340 · AISAFE2-WEAK-025 → AISAFE2-IMP-010

Компрометация разделяемого, встроенного или долгоживущего административного секрета даёт атакующему доступ к штатной конфигурации, через которую он фактически читает системный промпт либо инструкцию инструмента.

Ограничение: Только если скомпрометированный ключ или секрет дал неуполномоченному субъекту фактически прочитать системный промпт либо инструкцию инструмента

Связь 341 · AISAFE2-WEAK-025 → AISAFE2-IMP-011

Компрометация криптографического ключа позволяет расшифровать хранимый корпус, а компрометация разделяемого, встроенного или долгоживущего административного секрета — экспортировать через функции управления существенную часть базы знаний, её метаданных или индекса.

Ограничение: Только если скомпрометированный ключ или секрет позволил фактически расшифровать либо экспортировать существенную часть базы знаний или индекса

Связь 342 · AISAFE2-WEAK-025 → AISAFE2-IMP-012

Компрометация криптографического ключа снимает защиту хранимых или передаваемых артефактов модели, а компрометация разделяемого, встроенного или долгоживущего административного секрета даёт доступ к реестру артефактов; в обоих случаях атакующий может несанкционированно скопировать веса или связанные артефакты.

Ограничение: Только если скомпрометированный ключ или секрет привёл к фактическому копированию весов либо связанных артефактов без разрешения

Связь 343 · AISAFE2-WEAK-025 → AISAFE2-IMP-015

Компрометация криптографического ключа либо привилегированного секрета позволяет атакующему расшифровать или через штатный интерфейс прочитать либо изменить хранимый объект вне предоставленных ему полномочий.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

Связь 344 · AISAFE2-WEAK-025 → AISAFE2-IMP-016

Компрометация разделяемого, встроенного или долгоживущего учётного секрета позволяет атакующему действовать от имени легитимного привилегированного субъекта и успешно вызвать функцию инфраструктурного компонента за пределами собственных полномочий.

Ограничение: Только если скомпрометированный секрет использован для успешного вызова функции вне собственных полномочий атакующего

Связь 345 · AISAFE2-WEAK-026 → AISAFE2-IMP-011

Недостаточность механизмов ограничения числа и частоты запросов может привести к компрометации базы знаний, потому что интерфейсы инференса принимают обращения без сквозной квоты субъекта, и массовый систематический опрос воспроизводит содержимое и свойства базы знаний серией запросов.

Ограничение: Только если серия запросов без действенной квоты позволила воспроизводимо собрать существенную часть базы знаний или индекса

Связь 346 · AISAFE2-WEAK-026 → AISAFE2-IMP-012

Недостаточность механизмов ограничения числа и частоты запросов может привести к краже модели, потому что отсутствие сквозных ограничений числа и частоты запросов делает возможным массовый опрос для восстановления функционального аналога модели.

Ограничение: Только если массовый опрос без действенной квоты позволил подтвердить функциональный аналог ценного поведения модели

Связь 347 · AISAFE2-WEAK-026 → AISAFE2-IMP-013

Недостаточность механизмов ограничения числа и частоты запросов может привести к раскрытию чувствительных данных из обучающего датасета, потому что неограниченное число запросов через штатный интерфейс позволяет систематически извлекать свойства и записи обучающих данных.

Ограничение: Только если серия запросов позволила подтвердить раскрытие или реконструкцию конкретной чувствительной записи, её атрибута либо участия в обучении

Связь 348 · AISAFE2-WEAK-026 → AISAFE2-IMP-019

Недостаточность механизмов ограничения числа и частоты запросов может привести к чрезмерному потреблению вычислительных ресурсов системы, потому что отсутствие сквозных квот числа и частоты запросов позволяет создавать избыточную вычислительную нагрузку на инфраструктуру владельца.

Ограничение: Только при фактическом превышении утверждённого бюджета вычислений, памяти, сети или ускорителей собственной среды AI-системы.

Связь 349 · AISAFE2-WEAK-026 → AISAFE2-IMP-022

Недостаточность механизмов ограничения числа и частоты запросов может привести к недоступности системы или непредоставлению результата в директивный срок, потому что неограниченный по числу и частоте поток запросов создаёт нагрузку, ведущую к отказу в обслуживании легитимных потребителей.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 350 · AISAFE2-WEAK-027 → AISAFE2-IMP-012

Отсутствие экономического ограничителя позволяет продолжать массовый опрос, стоимость которого ниже ценности извлекаемого поведения, до подтверждённого построения функционального аналога модели.

Ограничение: Только если экономически доступная кампания запросов дошла до подтверждённого функционального аналога модели

Связь 351 · AISAFE2-WEAK-027 → AISAFE2-IMP-019

Близкая к нулю предельная стоимость поддерживает массовую кампанию, совокупная нагрузка которой превышает бюджет вычислений владельца.

Ограничение: Только при фактическом превышении утверждённого бюджета вычислений, памяти, сети или ускорителей и установленной связи с продолжающейся кампанией.

Связь 352 · AISAFE2-WEAK-028 → AISAFE2-IMP-011

Неразличимая распределённая кампания извлечения базы знаний продолжается до достижения порога компрометации, когда связанный с детектированием контур реагирования в противном случае остановил бы её раньше.

Ограничение: Только если незамеченная кампания достигла порога компрометации, а действующий контур реагирования остановил бы её по своевременному сигналу; нехватка телеметрии без этого пути недостаточна

Связь 353 · AISAFE2-WEAK-028 → AISAFE2-IMP-012

Неразличимая распределённая кампания накопления пар запрос—ответ продолжается до функционального извлечения модели, когда связанный контур реагирования в противном случае остановил бы её раньше.

Ограничение: Только при доказанной кампании извлечения, заранее действующем контуре реагирования и подтверждённом функциональном аналоге, полученном из-за продолжения незамеченной кампании.

Связь 354 · AISAFE2-WEAK-028 → AISAFE2-IMP-019

Реальная аномальная или вредоносная кампания повторных обращений к модели остаётся необнаруженной, хотя достаточный сигнал должен был запустить предусмотренный контур реагирования; продолжение кампании накапливает токенный расход запросов и связанную вычислительную нагрузку до превышения утверждённого бюджета ресурсов среды.

Ограничение: Только если незамеченная кампания повторных запросов вызвала измеренный перерасход ресурсов, а действующий контур реагирования остановил бы её по своевременному сигналу; токенный предел одного запроса может сохраняться

Связь 355 · AISAFE2-WEAK-031 → AISAFE2-IMP-012

Недостаточная защита модели при поставке в недоверенную среду может привести к краже модели, потому что артефакт исполняется в среде, которую контролирует получатель, расшифрованные веса неизбежно оказываются в памяти процесса и копируются из дистрибутива, памяти или временных файлов.

Ограничение: Только если из дистрибутива, памяти или временных файлов в недоверенной среде были фактически скопированы веса либо связанные артефакты без разрешения

Связь 356 · AISAFE2-WEAK-032 → AISAFE2-IMP-011

Недостаточная защита физического доступа к модели и инфраструктуре может привести к компрометации базы знаний, потому что физический доступ к носителям и серверам хранилищ позволяет скопировать корпус и индекс в обход логических средств защиты.

Ограничение: Только если физический доступ привёл к копированию существенной части базы знаний или индекса субъектом без полномочий

Связь 357 · AISAFE2-WEAK-032 → AISAFE2-IMP-012

Физический или аппаратно-привилегированный доступ может открыть путь к артефакту модели, если защита носителя, ключей и локального доступа не перекрывает этот путь.

Ограничение: Только если физический доступ привёл к неразрешённому копированию весов или связанных артефактов модели

Связь 358 · AISAFE2-WEAK-032 → AISAFE2-IMP-015

При физическом или аппаратно-привилегированном доступе субъект может прочитать либо изменить хранимые данные, если защита носителя, ключей и локального доступа допускает такую операцию.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

Связь 359 · AISAFE2-WEAK-032 → AISAFE2-IMP-016

Недостаточная защита физического доступа к модели и инфраструктуре может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что субъект с физическим доступом вмешивается в работу оборудования и компрометирует компоненты инфраструктуры в обход логических средств контроля.

Ограничение: Только если физический доступ привёл к вызову инфраструктурной функции без полномочий; вмешательство в оборудование без такого вызова сюда не входит

Связь 360 · AISAFE2-WEAK-033 → AISAFE2-IMP-013

Модель сохраняет сведения об отдельной чувствительной обучающей записи; целевые запросы к выпущенной версии позволяют неуполномоченному субъекту воспроизвести её фрагмент, вывести атрибут или подтвердить участие записи в обучении. Удаление исходной записи из датасета само по себе не меняет выпущенную модель.

Связь 361 · AISAFE2-WEAK-035 → AISAFE2-IMP-009

Набор без владельца, классификации и решения о допустимости включается в инструментальный источник, после чего чувствительное содержание фактически выдаётся неуполномоченному получателю.

Ограничение: Только при фактической выдаче или воспроизводимой реконструкции чувствительных сведений из инструментального источника неуполномоченным получателем.

Связь 362 · AISAFE2-WEAK-035 → AISAFE2-IMP-013

Чувствительный набор без классификации и решения о допустимости включается в обучение, после чего запись или атрибут воспроизводимо раскрывается через модель.

Ограничение: Только при подтверждённом воспроизведении, определении принадлежности, логическом выводе или восстановлении конкретной чувствительной записи либо её атрибута.

Связь 363 · AISAFE2-WEAK-036 → AISAFE2-IMP-009

Лишние чувствительные поля или записи не удаляются до передачи в RAG или контекст и фактически выдаются получателю без полномочий.

Ограничение: Только при фактической выдаче или воспроизводимой реконструкции чувствительных сведений из инструментального источника неуполномоченным получателем.

Связь 364 · AISAFE2-WEAK-036 → AISAFE2-IMP-013

Не нужные для цели чувствительные записи включаются в обучение и затем воспроизводимо выводятся либо восстанавливаются через модель.

Ограничение: Только при подтверждённом воспроизведении, определении принадлежности, логическом выводе или восстановлении конкретной чувствительной записи либо её атрибута.

Связь 365 · AISAFE2-WEAK-036 → AISAFE2-IMP-024

Избыточные чувствительные поля могут попасть к внешнему провайдеру, которому передача именно этих данных не разрешена утверждённой политикой; фактический запрос тогда пересекает запрещённую границу.

Ограничение: Только если избыточные данные фактически переданы провайдеру, не утверждённому для этого класса данных и цели; избыток в разрешённом контуре сам по себе не утечка

Связь 366 · AISAFE2-WEAK-037 → AISAFE2-IMP-006

Непропорциональный вклад отдельных обучающих записей в результат обучения может привести к неожидаемым для разработчика ответам модели, потому что из-за отсутствия дедупликации немногочисленные намеренно внедрённые записи непропорционально влияют на обучение и позволяют целенаправленно изменить поведение модели, включая внедрение закладки.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 367 · AISAFE2-WEAK-037 → AISAFE2-IMP-013

Непропорциональный вклад отдельных обучающих записей в результат обучения может привести к раскрытию чувствительных данных из обучающего датасета, потому что дубликаты и связанные записи не ограничиваются по суммарному вкладу, запоминание конкретных примеров усиливается, и становится возможным их извлечение и определение участия в обучении.

Связь 368 · AISAFE2-WEAK-039 → AISAFE2-IMP-009

Инструмент возвращает чувствительные или избыточные поля без проекции, усечения и сохранения метки чувствительности; результат поступает в модель или интерфейс и раскрывает данные неуполномоченному получателю.

Связь 369 · AISAFE2-WEAK-039 → AISAFE2-IMP-024

Инструмент передаёт чувствительные данные без минимизации или проверки меток внешней модели либо иному получателю, которому эти данные не разрешены; фактическая передача пересекает запрещённую границу.

Ограничение: Только если инструмент выдал чувствительные поля модели или получателю, которым их передавать нельзя; избыточный объём при разрешённом получателе сам по себе не утечка

Связь 370 · AISAFE2-WEAK-040 → AISAFE2-IMP-009

Отсутствие маскирования или анонимизации чувствительных данных может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что необходимые для задачи данные циркулируют в исходном виде вместо маскированного, и попавшие в контекст и ответы значения раскрывают чувствительные сведения без потери функции защищённой формы.

Ограничение: Только если немаскированные сведения из инструментального источника выданы или воспроизводимо восстановлены получателем без полномочий; само хранение исходных значений недостаточно

Связь 371 · AISAFE2-WEAK-040 → AISAFE2-IMP-013

Отсутствие маскирования или анонимизации чувствительных данных может привести к раскрытию чувствительных данных из обучающего датасета, потому что необходимые записи сохраняются в обучающих наборах с чувствительными значениями в исходном виде вместо маскированного, и запоминание значений моделью раскрывает их.

Ограничение: Только если исходное немаскированное значение из обучающего набора воспроизведено или восстановлено через модель; его хранение в датасете недостаточно

Связь 372 · AISAFE2-WEAK-040 → AISAFE2-IMP-024

Исходные чувствительные значения из контекста, журнала или набора данных без маскирования передаются получателю, которому их обрабатывать нельзя, и пересекают запрещённую границу.

Ограничение: Только если исходные чувствительные значения из журнала, набора или контекста ушли получателю или среде без разрешения; отсутствие маскирования внутри доверенного контура недостаточно

Связь 373 · AISAFE2-WEAK-041 → AISAFE2-IMP-010

Недостаточный контроль доступа к системным инструкциям и секретам может привести к раскрытию системного промпта и инструкций инструментов, потому что доступ к конфигурации системного промпта и инструкциям инструментов не ограничен по субъекту и операции, и их читает более широкий круг, чем необходимо.

Ограничение: Только если субъект без соответствующего права фактически прочитал системный промпт либо инструкцию инструмента через конфигурацию

Связь 374 · AISAFE2-WEAK-041 → AISAFE2-IMP-015

Недостаточный контроль доступа к системным инструкциям и секретам может привести к несанкционированному чтению или изменению хранимых данных, потому что круг субъектов, способных читать и изменять системные инструкции, описания инструментов и связанные секреты, шире необходимого, и несанкционированное изменение меняет поведение системы.

Связь 375 · AISAFE2-WEAK-042 → AISAFE2-IMP-002

Недостаточная пригодность модели для целевого сценария может привести к ущербу от решения, принятого на основе результата AI-системы, потому что модель, не соответствующая целевому сценарию по критериям качества, систематически ошибается именно там, где применяется, и пользователи принимают решения на недостоверных результатах.

Ограничение: Только если непригодность модели для целевого сценария дала ошибочный результат, на котором существенно основывалось исполненное вредное решение

Связь 376 · AISAFE2-WEAK-042 → AISAFE2-IMP-006

Недостаточная пригодность модели для целевого сценария может привести к неожидаемым для разработчика ответам модели, потому что конкретная версия модели проваливает граничные случаи и теряет требуемые свойства, систематически выдавая результаты за предусмотренными границами качества и поведения.

Связь 377 · AISAFE2-WEAK-043 → AISAFE2-IMP-006

Недостаточные ограничения в системных инструкциях могут привести к неожидаемым для разработчика ответам модели, потому что неполные, неоднозначные или противоречивые системные инструкции оставляют формально не запрещённые запросы обслуживаемыми штатно, за пределами задуманных разработчиком границ.

Связь 378 · AISAFE2-WEAK-043 → AISAFE2-IMP-007

Недостаточные ограничения в системных инструкциях могут привести к обходу ограничений и предусмотренной логики агента, потому что границы поведения заданы в системных инструкциях неполно или противоречиво, и формально не запрещённые запросы уводят агента от предусмотренного сценария работы.

Ограничение: Только когда модель, контекст и Agent Core фактически меняют предусмотренный план, действие или переход состояния; отдельный сбой инструмента не входит.

Связь 379 · AISAFE2-WEAK-043 → AISAFE2-IMP-010

Недостаточные ограничения в системных инструкциях могут привести к раскрытию системного промпта и инструкций инструментов, потому что системная инструкция не запрещает запросы вида повтори всё, что было выше, и содержимое контекста раскрывается формально не запрещёнными просьбами.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 380 · AISAFE2-WEAK-043 → AISAFE2-IMP-024

Системная инструкция не задаёт требуемый запрет внешней передачи, и модель или агент фактически отправляет чувствительные данные недоверенному получателю.

Ограничение: Только если утверждённая политика запрещала конкретную передачу, этот запрет не был задан агенту и данные фактически ушли запрещённому адресату

Связь 381 · AISAFE2-WEAK-044 → AISAFE2-IMP-006

Недостатки управления изменениями и развёртыванием AI-системы могут привести к неожидаемым для разработчика ответам модели, потому что изменения модели, библиотек и конфигурации попадают в рабочую среду без обязательных проверок защищённости, и в эксплуатации появляются нежелательно ведущие себя версии.

Ограничение: Только когда неожиданный ответ модели сформирован и передан потребителю или следующему компоненту; возможность ошибки без выданного ответа недостаточна.

Связь 382 · AISAFE2-WEAK-044 → AISAFE2-IMP-022

Недостатки управления изменениями и развёртыванием AI-системы могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что изменения публикуются без поэтапного развёртывания и плана отката, и дефектную версию нельзя быстро откатить, продлевая недоступность сервиса.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 383 · AISAFE2-WEAK-045 → AISAFE2-IMP-011

Недостатки управления учётными записями и правами в контуре MLOps могут привести к компрометации базы знаний, потому что избыточные и неотозванные права учётных записей в контуре MLOps дают широкий несанкционированный доступ к данным и хранилищам базы знаний.

Ограничение: Только если избыточное право MLOps распространялось на базу знаний и было использовано для получения её существенной части

Связь 384 · AISAFE2-WEAK-045 → AISAFE2-IMP-012

Недостатки управления учётными записями и правами в контуре MLOps могут привести к краже модели, потому что накопленные избыточные и неотозванные права в контуре MLOps дают несанкционированный доступ к артефактам модели в конвейерах и хранилищах.

Связь 385 · AISAFE2-WEAK-045 → AISAFE2-IMP-015

Недостатки управления учётными записями и правами в контуре MLOps могут привести к несанкционированному чтению или изменению хранимых данных, потому что сменившие роль и уволившиеся сохраняют права в конвейерах и хранилищах, и компрометация такой учётной записи даёт широкий доступ к данным, артефактам и конфигурации.

Связь 386 · AISAFE2-WEAK-045 → AISAFE2-IMP-031

Устаревшее либо избыточное человеческое или административное право MLOps используется для удаления или повреждения датасета вместе со всеми доступными доверенными копиями, вследствие чего датасет нельзя восстановить для обучения, тестирования или валидации.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения всех доступных доверенных копий.

Связь 387 · AISAFE2-WEAK-045 → AISAFE2-IMP-033

Устаревшее либо избыточное человеческое или административное право MLOps используется для удаления или повреждения весов, базы знаний, индекса либо критичной конфигурации вместе с доступными версиями и рецептами восстановления.

Ограничение: Только для вызванной атакой невосстановимой утраты весов, базы знаний, индекса или критичной конфигурации при отсутствии пригодной версии либо доверенного рецепта восстановления.

Связь 388 · AISAFE2-WEAK-047 → AISAFE2-IMP-015

Недостаточное управление уязвимостями может привести к несанкционированному чтению или изменению хранимых данных, потому что известные уязвимости остаются применимыми к рабочим компонентам неопределённо долго, и их эксплуатация даёт несанкционированный доступ к хранимым данным.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

Связь 389 · AISAFE2-WEAK-047 → AISAFE2-IMP-016

Недостаточное управление уязвимостями может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что известные, но не устранённые уязвимости компонентов эксплуатируются для выполнения функций и произвольного кода вне полномочий.

Ограничение: Только если эксплуатация неустранённой уязвимости привела к успешному вызову функции вне полномочий атакующего

Связь 390 · AISAFE2-WEAK-047 → AISAFE2-IMP-022

Недостаточное управление уязвимостями может привести к недоступности системы или непредоставлению результата в директивный срок, потому что эксплуатация неустранённых уязвимостей выводит компоненты из строя, останавливая обслуживание потребителей.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 391 · AISAFE2-WEAK-047 → AISAFE2-IMP-031

Атакующий эксплуатирует неустранённую уязвимость для разрушительного доступа к точному датасету и уничтожает все пригодные доверенные копии и пути воспроизводимого восстановления.

Ограничение: Только если неустранённая уязвимость использована для уничтожения точного датасета и всех пригодных копий или путей восстановления; невосстановимость подтверждена

Связь 392 · AISAFE2-WEAK-047 → AISAFE2-IMP-033

Атакующий эксплуатирует неустранённую уязвимость для разрушительного доступа к весам, базе знаний, индексу либо критичной конфигурации и уничтожает все пригодные доверенные копии и пути воспроизводимого восстановления.

Ограничение: Только если неустранённая уязвимость использована для уничтожения точных весов, базы знаний, индекса или конфигурации и всех пригодных путей восстановления; невосстановимость подтверждена

Связь 393 · AISAFE2-WEAK-048 → AISAFE2-IMP-022

Недостатки реагирования на инциденты и процесса аварийного восстановления могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что сценарии локализации и восстановления не отработаны и не обеспечены проверенными копиями, операторы импровизируют, и простой растягивается сверх директивных интервалов.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 394 · AISAFE2-WEAK-048 → AISAFE2-IMP-031

После разрушительного воздействия атаки на точный датасет неотработанный процесс восстановления и отсутствие проверенных доверенных копий превращают повреждение в невосстановимую утрату.

Ограничение: Только если атака повредила точный датасет, пригодных доверенных копий или рецептов нет и проверенное восстановление не удалось; временный простой исключён

Связь 395 · AISAFE2-WEAK-048 → AISAFE2-IMP-033

После разрушительного воздействия атаки на веса, базу знаний, индекс либо критичную конфигурацию неотработанное восстановление и отсутствие проверенных копий делают объект невосстановимым.

Ограничение: Только если атака повредила точный критичный артефакт, кроме датасета, пригодных версий или рецептов нет и проверенное восстановление не удалось

Связь 396 · AISAFE2-WEAK-049 → AISAFE2-IMP-007

Слабое установление доверия в межагентном взаимодействии может привести к обходу ограничений и предусмотренной логики агента, потому что поддельное, изменённое, ошибочно адресованное или повторно воспроизведённое сообщение другого агента принимается как действующее поручение и меняет планы и действия системы.

Связь 397 · AISAFE2-WEAK-049 → AISAFE2-IMP-008

Получатель принимает поддельное, изменённое или повторно воспроизведённое межагентское поручение за действительное и передаёт его инструментальному слою; инструмент или сторонний агент возвращает непредусмотренный результат либо выполняет действие вне контракта.

Ограничение: Только если такое поручение фактически принято и вызвало непредусмотренный результат или действие, достигшее следующего компонента либо внешней системы.

Связь 398 · AISAFE2-WEAK-049 → AISAFE2-IMP-024

Слабое установление доверия в межагентном взаимодействии может привести к утечке чувствительных данных в недоверенный контур, потому что данные передаются агенту, идентичность и полномочия которого не подтверждены, то есть фактически неизвестному получателю за границей доверия.

Ограничение: Только если чувствительные данные фактически переданы агенту, чьи идентичность и полномочия не установлены либо не разрешены для этих данных

Связь 399 · AISAFE2-WEAK-050 → AISAFE2-IMP-032

Неочищенная память, кэши, буферы или временные файлы общей среды передают состояние предыдущего задания следующему изолированному контексту, что напрямую смешивает краткоживущие артефакты разных заданий или арендаторов.

Связь 400 · AISAFE2-WEAK-051 → AISAFE2-IMP-007

Выдача результата без завершения обязательных стадий вычислений может привести к обходу ограничений и предусмотренной логики агента, потому что пропуск или сбой обязательной стадии неотличим от успешного завершения, и агент пропускает обязательный шаг (например, согласование), выдавая неполное выполнение за проверенное.

Связь 401 · AISAFE2-WEAK-052 → AISAFE2-IMP-017

Неограниченные внутренние циклы и усиление нагрузки могут привести к неограниченному потреблению токенов, потому что внутренние шаги планирования, повторы и самопроверка не ограничены бюджетом задачи, и расход внутри одного запроса растёт незаметно для лимитов на число запросов.

Ограничение: Только для одной траектории обработки с повторными обращениями к языковой модели, в которой не срабатывает утверждённый токенный предел.

Связь 402 · AISAFE2-WEAK-052 → AISAFE2-IMP-018

Неограниченные внутренние циклы и усиление нагрузки могут привести к незапланированным расходам на платные инструменты, потому что повторные попытки, веерные вызовы инструментов и циклы агента множат обращения к тарифицируемым сервисам без ограничения числа шагов и бюджета задачи.

Ограничение: Только когда цикл или действие содержит тарифицируемый инструмент либо внешний сервис и создаёт подтверждённое начисление вне утверждённого бюджета.

Связь 403 · AISAFE2-WEAK-052 → AISAFE2-IMP-019

Неограниченные внутренние циклы и усиление нагрузки могут привести к чрезмерному потреблению вычислительных ресурсов системы, потому что один запрос порождает неограниченную траекторию вычислений — рекурсию, самопроверку, веерные вызовы, разрастание контекста, — занимая несоразмерный объём ресурсов.

Связь 404 · AISAFE2-WEAK-052 → AISAFE2-IMP-021

Если внутренние шаги одного запроса повторяются без достижения нового состояния, отсутствие жёсткого предела итераций, глубины вложенности или времени не останавливает цикл и позволяет ему продолжаться до внешнего прерывания.

Ограничение: Только при наблюдаемом повторении шагов без штатного завершения; само отсутствие лимита не доказывает зацикливание.

Связь 405 · AISAFE2-WEAK-052 → AISAFE2-IMP-022

Неограниченные внутренние циклы и усиление нагрузки могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что неограниченные внутренние циклы удерживают ресурсы и квоты вплоть до зацикливания системы и отказа в обслуживании.

Связь 406 · AISAFE2-WEAK-053 → AISAFE2-IMP-012

Небезопасный вывод AI-компонента из эксплуатации может привести к краже модели, потому что после вывода компонента из эксплуатации забытые копии моделей остаются доступными для хищения в несопровождаемом и неучтённом контуре.

Ограничение: Только если из оставшейся после вывода из эксплуатации копии фактически получены веса или связанные артефакты без разрешения

Связь 407 · AISAFE2-WEAK-053 → AISAFE2-IMP-015

Небезопасный вывод AI-компонента из эксплуатации может привести к несанкционированному чтению или изменению хранимых данных, потому что у выведенного из эксплуатации компонента сохраняются доступы, учётные данные и данные, и связанные системы продолжают предоставлять доступ от его имени.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

Связь 408 · AISAFE2-WEAK-053 → AISAFE2-IMP-016

Небезопасный вывод AI-компонента из эксплуатации может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что выведенный компонент остаётся неучтённой точкой входа: связанные системы продолжают принимать вызовы и предоставлять доступ от его имени.

Ограничение: Только если оставшийся после вывода путь позволил вызвать инфраструктурную функцию от имени снятого компонента без полномочий

Связь 409 · AISAFE2-WEAK-054 → AISAFE2-IMP-001

Стохастическая недостоверность вывода генеративной модели может привести к выполнению пользователем предложенного системой вредоносного действия, потому что модель формирует правдоподобный, но не обязательно верный или безопасный результат при штатной работе, и ошибочная рекомендация (несуществующая функция, вредная команда) появляется без атаки и отказа.

Ограничение: Только если ошибочная рекомендация была прямо предложена как действие, пользователь её выполнил и понёс вред

Связь 410 · AISAFE2-WEAK-054 → AISAFE2-IMP-002

Стохастическая недостоверность вывода генеративной модели может привести к ущербу от решения, принятого на основе результата AI-системы, потому что модель выдаёт правдоподобный текст, а не проверенный факт, и решение, существенно опирающееся на такой ответ (сделка, отмена заказа, утверждение расчёта), оборачивается ущербом.

Ограничение: Только если недостоверный ответ стал существенным основанием исполненного решения и вреда; сама генерация ошибочного ответа недостаточна

Связь 411 · AISAFE2-WEAK-054 → AISAFE2-IMP-006

Генеративная модель может при штатной работе выдать правдоподобный, но неверный ответ, вымышленный источник или ошибочную рекомендацию; детерминированное декодирование само по себе не обеспечивает истинность.

Связь 412 · AISAFE2-WEAK-055 → AISAFE2-IMP-007

Возложение бизнес-логики и обязательных ограничений на модель может привести к обходу ограничений и предусмотренной логики агента, потому что обязательные правила существуют только как текст для модели, её решение напрямую становится действием системы, и убеждение, переформулирование или инъекция меняют поведение агента.

Связь 413 · AISAFE2-WEAK-056 → AISAFE2-IMP-007

Недостаточная валидация аргументов вызова инструментов может привести к обходу ограничений и предусмотренной логики агента, потому что аргументы вызова, собранные моделью из недоверенного текста, не проверяются перед исполнением, и подмена объектов операции (идентификаторов, адресатов, сумм) уводит выполнение за предусмотренные рамки.

Связь 414 · AISAFE2-WEAK-056 → AISAFE2-IMP-008

Недостаточная валидация аргументов вызова инструментов может привести к неожидаемому поведению слоя инструментов, потому что через непроверенные аргументы вызова, собранные из недоверенного текста, даже исправный инструмент можно заставить выполнить непредусмотренные действия за пределами контракта.

Связь 415 · AISAFE2-WEAK-056 → AISAFE2-IMP-015

Недостаточная валидация аргументов вызова инструментов может привести к несанкционированному чтению или изменению хранимых данных, потому что через непроверенные аргументы в нижележащие системы проходят инъекции запросов к базам данных и обход файловых путей, которые исполняются с правами инструмента и открывают чтение и изменение данных вне полномочий инициатора.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

Связь 416 · AISAFE2-WEAK-056 → AISAFE2-IMP-016

Недостаточная валидация аргументов вызова инструментов может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что непроверенные аргументы, собранные моделью из недоверенного текста, реализуют инъекции в нижележащие системы (SQL, команды ОС, обход путей) вплоть до компрометации инструмента и смежных компонентов.

Ограничение: Только если непроверенный аргумент привёл к инъекции или обходу, давшему вызов функции вне полномочий инициатора

Связь 417 · AISAFE2-WEAK-057 → AISAFE2-IMP-007

Отсутствие обязательного подтверждения чувствительных действий может привести к обходу ограничений и предусмотренной логики агента, потому что предусмотренный контур согласования высокорисковых операций фактически не исполняется: необратимые и внешне значимые действия совершаются без явного решения уполномоченного субъекта по их точным параметрам, и логика обязательных подтверждений обходится.

Связь 418 · AISAFE2-WEAK-057 → AISAFE2-IMP-018

Отсутствие обязательного подтверждения чувствительных действий может привести к незапланированным расходам на платные инструменты, потому что дорогостоящие операции не требуют явного подтверждения уполномоченным субъектом, и платные вызовы совершаются без осознанного решения по их параметрам.

Ограничение: Только если тарифицируемый вызов требовал обязательного подтверждения, прошёл без него и создал начисление сверх утверждённого бюджета

Связь 419 · AISAFE2-WEAK-057 → AISAFE2-IMP-020

Отсутствие обязательного подтверждения чувствительных действий позволяет агенту совершать дорогие операции от имени пользователя без его согласия, что в тарифицируемом контексте напрямую конвертируется в неожиданные для пользователя расходы.

Ограничение: Только если стоимость операции несёт пользователь и дорогой вызов прошёл без обязательного подтверждения; вне пользовательского тарифа этот недостаток не создаёт его расхода

Связь 420 · AISAFE2-WEAK-057 → AISAFE2-IMP-024

Отсутствие обязательного подтверждения чувствительных действий может привести к утечке чувствительных данных в недоверенный контур, потому что действия передачи данных вовне — публикация, отправка, выгрузка — исполняются от имени пользователя без явного решения по точным параметрам операции, и ошибочная или навязанная атакующим передача пересекает границу доверия раньше, чем пользователь о ней узнаёт.

Ограничение: Только если требующая подтверждения передача чувствительных данных прошла без решения уполномоченного пользователя к неразрешённому адресату

Связь 421 · AISAFE2-WEAK-058 → AISAFE2-IMP-009

Вызов инструментов без учёта идентичности и полномочий инициатора может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что инструмент действует от общей служебной учётной записи, не проверяя идентичность и полномочия инициатора, и возвращает данные, на которые у запрашивающего субъекта нет прав.

Связь 422 · AISAFE2-WEAK-058 → AISAFE2-IMP-015

Вызов инструментов без учёта идентичности и полномочий инициатора может привести к несанкционированному чтению или изменению хранимых данных, потому что запрос менее привилегированного субъекта наследует широкие права сервисной учётной записи инструмента и читает или меняет чужие хранимые данные.

Связь 423 · AISAFE2-WEAK-058 → AISAFE2-IMP-016

Вызов инструментов без учёта идентичности и полномочий инициатора может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что инструмент не проверяет идентичность и полномочия инициатора и выполняет функции с широкими сервисными правами по запросу менее привилегированного субъекта.

Связь 424 · AISAFE2-WEAK-059 → AISAFE2-IMP-013

Недостаточная изоляция арендаторов при дообучении модели может привести к раскрытию чувствительных данных из обучающего датасета, потому что при дообучении данные разных арендаторов смешиваются без сохранения принадлежности в параметрах модели, и через штатный интерфейс общего сервиса извлекаются чужие данные.

Ограничение: Только если данные другого арендатора фактически извлечены или восстановлены через общую дообученную модель; совместное обучение без такого результата недостаточно

Связь 425 · AISAFE2-WEAK-060 → AISAFE2-IMP-014

Нефильтрованное внутреннее сообщение об ошибке, содержащее закрытые сведения о стеке, компонентах, конфигурации, алгоритмах или источниках данных, пересекает внешнюю границу и становится доступно неуполномоченному получателю.

Связь 426 · AISAFE2-WEAK-061 → AISAFE2-IMP-014

Избыточно открытая внутренняя телеметрия может привести к раскрытию сведений о технологиях, алгоритмах и источниках данных системы, потому что трассировки, метрики и служебные заголовки доступны внешнему получателю через ответы, интерфейс и экспорт аналитики и раскрывают сведения об устройстве и работе системы.

Ограничение: Только при подтверждённом неразрешённом получении закрытой информации о технологиях, архитектуре, алгоритмах или источниках данных.

Связь 427 · AISAFE2-WEAK-061 → AISAFE2-IMP-024

Избыточно открытая внутренняя телеметрия может привести к утечке чувствительных данных в недоверенный контур, потому что телеметрия и служебные данные экспортируются во внешнюю аналитику и попадают в ответы и заголовки, вынося за контур сведения и чувствительное содержимое запросов.

Ограничение: Только если телеметрия с чувствительным содержимым запроса была экспортирована или выдана получателю без разрешения

Связь 428 · AISAFE2-WEAK-062 → AISAFE2-IMP-017

Если число внутренних обращений к модели не ограничено, отсутствие сквозного токенного бюджета позволяет составной операции продолжать расходовать токены без верхней границы, хотя каждый отдельный вызов соблюдает локальный лимит.

Ограничение: Только если у запроса, сессии или цепочки нет другого действующего конечного предела общего токенного расхода; конечная цепочка с локальными лимитами может превысить бюджет, но не образует неограниченное потребление.

Связь 429 · AISAFE2-WEAK-062 → AISAFE2-IMP-018

Отсутствие сквозного учёта стоимости между участниками одной операции позволяет каждому компоненту соблюдать локальный лимит, тогда как совокупные вызовы платных инструментов превышают утверждённый бюджет и создают незапланированные начисления.

Связь 430 · AISAFE2-WEAK-062 → AISAFE2-IMP-019

Отсутствие общего ресурсного бюджета для составной операции позволяет отдельным компонентам продолжать вычисления в пределах локальных лимитов, тогда как их совокупное потребление вычислений, памяти и сети превышает утверждённый предел среды.

Связь 431 · AISAFE2-WEAK-063 → AISAFE2-IMP-020

Если пользователь оплачивает потребление, отсутствие доступного ему бюджета, оценки и уведомления может привести к счёту, существенно превышающему показанную или разумно ожидаемую стоимость.

Ограничение: Только при оплате потребления пользователем и подтверждённом существенном превышении ожидаемой стоимости; само увеличение тарифицируемого объёма не достаточно.

Связь 432 · AISAFE2-WEAK-064 → AISAFE2-IMP-018

Неизолированные отказы внешних зависимостей могут привести к незапланированным расходам на платные инструменты, потому что безусловные повторы при отказах и замедлениях внешних зависимостей порождают лавину дополнительных тарифицируемых обращений к платным сервисам.

Ограничение: Только когда цикл или действие содержит тарифицируемый инструмент либо внешний сервис и создаёт подтверждённое начисление вне утверждённого бюджета.

Связь 433 · AISAFE2-WEAK-064 → AISAFE2-IMP-022

Неизолированные отказы внешних зависимостей не компенсируются переключением и планом обработки недоступности, что позволяет отказу или деградации одного звена распространиться на весь цикл обработки запроса и привести к недоступности AI-системы.

Связь 434 · AISAFE2-WEAK-065 → AISAFE2-IMP-023

Избыточная, слишком мелкая и повторная маршрутизация задач на ручную проверку для эквивалентных случаев генерирует поток обращений, превышающий пропускную способность контура HITL, что напрямую повышает нагрузку на человеческие ресурсы.

Связь 435 · AISAFE2-WEAK-066 → AISAFE2-IMP-023

Численность, график и резерв операторов, рассчитанные ниже пиковой потребности, не выдерживают всплеска обращений, что приводит к росту очереди и задержек в контуре HITL и напрямую повышает нагрузку на человеческие ресурсы.

Связь 436 · AISAFE2-WEAK-067 → AISAFE2-IMP-023

Недостаточная инструментальная поддержка увеличивает время обработки задания; когда поток превышает снизившуюся пропускную способность операторов, очередь и нагрузка растут.

Ограничение: При потоке ниже доступной пропускной способности увеличение времени обработки само по себе не означает перегрузку человеческого контура.

Связь 437 · AISAFE2-WEAK-069 → AISAFE2-IMP-007

Неопределённое поведение при перегрузке или недоступности HITL-контура может привести к обходу ограничений и предусмотренной логики агента, потому что при переполнении очереди проверки система пропускает обращения без проверки, превращая перегрузку HITL-контура в обход обязательного контроля.

Ограничение: Только для сценария продолжения работы без обязательной проверки: при недоступной человеческой проверке операция, которая по утверждённому маршруту обязана ждать решения HITL, продолжает выполнение без этой проверки и меняет предусмотренный план, действие или переход Agent Core.

Связь 438 · AISAFE2-WEAK-069 → AISAFE2-IMP-022

Неопределённое поведение при перегрузке или недоступности HITL-контура может привести к недоступности системы или непредоставлению результата в директивный срок, потому что при недоступности или переполнении HITL-контура вариант полной остановки обработки оборачивается отказом в обслуживании легитимных пользователей.

Ограничение: Только для сценария безопасного отказа с блокировкой операции: при недоступной человеческой проверке операция остаётся заблокированной, и потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 439 · AISAFE2-WEAK-070 → AISAFE2-IMP-024

Отсутствие управляемой политики допустимого использования AI-сервисов может привести к утечке чувствительных данных в недоверенный контур, потому что при отсутствии правил, обучения и внутренних альтернатив сотрудники в рамках обычной работы отправляют служебные документы в публичные AI-сервисы.

Ограничение: Только если сотрудник отправил в публичный AI-сервис чувствительные данные, для которых этот сервис и цель обработки не разрешены; отсутствие политики без передачи недостаточно

Связь 440 · AISAFE2-WEAK-071 → AISAFE2-IMP-024

Ошибочно классифицированный маршрут может передать чувствительные данные из AI-контура получателю, которому их передача запрещена действующим решением, если проверка зоны и потока не срабатывает до отправки.

Ограничение: Только если для конкретного класса данных и получателя отдельно установлено, что внешний поток не разрешён; отсутствие схемы зон само по себе не доказывает утечку

Связь 441 · AISAFE2-WEAK-072 → AISAFE2-IMP-024

Обход обязательного межзонного контроля позволяет фактически передать чувствительные данные получателю, которому такой поток запрещён утверждённой политикой.

Ограничение: Только при подтверждённой передаче запрещённому получателю; обход шлюза на разрешённом потоке сам по себе не образует утечку.

Связь 442 · AISAFE2-WEAK-073 → AISAFE2-IMP-001

Независимая проверка выявляет конкретную опасность в явно предложенном действии, но находка не доставляется потребителю. Из-за этого не срабатывает заранее установленный обязательный пересмотр или отказ; действие выполняется и причиняет подтверждённый вред.

Ограничение: Только если подтверждены предупреждающая находка, обязательный маршрут пересмотра или отказа при ней, его пропуск из-за недоставки, выполненное действие и вред.

Связь 443 · AISAFE2-WEAK-073 → AISAFE2-IMP-002

Независимая проверка выявляет конкретную ошибку в AI-результате, но находка не доставляется до решения. Из-за этого не срабатывает заранее установленный обязательный пересмотр или отказ; результат становится существенным основанием исполненного решения с подтверждённым вредом.

Ограничение: Только если подтверждены находка, обязательный маршрут пересмотра или отказа при ней, его пропуск из-за недоставки, существенное влияние результата на исполненное решение и вред.

Связь 444 · AISAFE2-WEAK-074 → AISAFE2-IMP-009

Отсутствие гарантий нераскрытия содержимого контекста модели может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что между содержимым контекста и выводным каналом нет технического разграничения доступа, и данные других участников обработки раскрываются через ответ дословно или в преобразованной форме.

Ограничение: Только если сведения из инструментального контекста раскрыты через вывод модели получателю без полномочий; доступ модели к контексту сам по себе недостаточен

Связь 445 · AISAFE2-WEAK-074 → AISAFE2-IMP-010

Отсутствие гарантий нераскрытия содержимого контекста модели может привести к раскрытию системного промпта и инструкций инструментов, потому что системная инструкция и описания инструментов — такой же текст однородного контекста без меток доступа, и их раскрытие через выводной канал технически не исключено.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 446 · AISAFE2-WEAK-074 → AISAFE2-IMP-024

Модель может воспроизвести чувствительный фрагмент контекста в ответе или вызове инструмента; если этот вывод фактически отправлен получателю без права на данные, происходит запрещённая передача.

Ограничение: Только если содержимое контекста с чувствительными данными фактически выдано моделью получателю без разрешения; отсутствие технического разграничения само по себе недостаточно

Связь 447 · AISAFE2-WEAK-075 → AISAFE2-IMP-019

Отсутствие совокупных пределов масштабирования и расходов инфраструктуры может привести к чрезмерному потреблению вычислительных ресурсов системы, потому что автоматическое масштабирование не имеет принудительного верхнего предела, а бюджетные пороги не связаны с защитными действиями, и выделение ресурсов растёт до ручного вмешательства.

Связь 448 · AISAFE2-WEAK-076 → AISAFE2-IMP-022

Отсутствие ресурсной изоляции арендаторов позволяет нагрузке одного арендатора занимать общие очереди и вычислительные ресурсы, из-за чего другие арендаторы не получают результат в установленный директивный интервал.

Связь 449 · AISAFE2-WEAK-077 → AISAFE2-IMP-022

Недостаточное планирование ёмкости и масштабирования под пиковую нагрузку может привести к недоступности системы или непредоставлению результата в директивный срок, потому что ёмкость и механизм масштабирования рассчитаны без запаса на пиковую нагрузку, и штатный рост обращений выводит время ответа за директивные пределы.

Связь 450 · AISAFE2-WEAK-078 → AISAFE2-IMP-022

Недостатки управления нагрузкой, очередями и приоритетами при перегрузке могут привести к недоступности системы или непредоставлению результата в директивный срок, потому что при спросе выше ёмкости приём не ограничивается, очереди растут неограниченно, и деградация распространяется на всех потребителей вплоть до отказа в обслуживании.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 451 · AISAFE2-WEAK-079 → AISAFE2-IMP-022

Зависимость от единственного поставщика сервисов или площадки размещения может привести к недоступности системы или непредоставлению результата в директивный срок, потому что доступность ограничена единственным поставщиком без подготовленной замены и переключения, и его авария или регламентные работы означают простой на неконтролируемый владельцем срок.

Ограничение: Только когда потребитель не получает доступ или результат к окончанию установленного директивного интервала.

Связь 452 · AISAFE2-WEAK-080 → AISAFE2-IMP-005

Недопустимое содержание выражается парафразом, переводом или кодированием, не распознаётся конечным набором признаков внешнего фильтра и фактически достигает пользователя.

Ограничение: Только для конкретного воздействия на пользователя по заранее утверждённым нормам, с указанной затронутой группой и подтверждённым вредом.

Связь 453 · AISAFE2-WEAK-080 → AISAFE2-IMP-009

Преобразованное чувствительное содержание из инструментального источника не распознаётся семантическим фильтром и выдаётся неуполномоченному получателю.

Ограничение: Только при фактической выдаче или воспроизводимой реконструкции чувствительных сведений из инструментального источника неуполномоченным получателем.

Связь 454 · AISAFE2-WEAK-080 → AISAFE2-IMP-010

Системная инструкция передаётся парафразом, переводом или кодированием, которые внешний фильтр не связывает с защищаемым исходным содержанием.

Ограничение: Только при дословной выдаче или содержательно полной реконструкции системного промпта либо инструкции инструмента.

Связь 455 · AISAFE2-WEAK-080 → AISAFE2-IMP-024

Чувствительные сведения в преобразованной форме обходят внешний семантический фильтр и фактически пересекают границу доверия.

Ограничение: Только если преобразованный вывод всё ещё раскрывает чувствительные сведения получателю, которому их передавать нельзя

Связь 456 · AISAFE2-WEAK-082 → AISAFE2-IMP-010

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и фактически прочитать системный промпт либо инструкцию инструмента через штатную конфигурацию.

Ограничение: Только если атакующий, принятый за привилегированного пользователя, фактически прочитал системный промпт либо инструкцию инструмента через конфигурацию

Связь 457 · AISAFE2-WEAK-082 → AISAFE2-IMP-011

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и фактически экспортировать существенную часть базы знаний, её метаданных либо индекса через функции управления.

Ограничение: Только если атакующий под видом привилегированного пользователя экспортировал существенную часть базы знаний, её метаданных или индекса

Связь 458 · AISAFE2-WEAK-082 → AISAFE2-IMP-012

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и несанкционированно скопировать веса либо связанные артефакты модели.

Ограничение: Только если атакующий под видом привилегированного пользователя фактически скопировал веса или связанные артефакты модели

Связь 459 · AISAFE2-WEAK-082 → AISAFE2-IMP-015

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и успешно прочитать либо изменить хранимый объект вне собственных полномочий.

Ограничение: Только при успешном чтении или изменении хранимого объекта вне действующей политики доступа; наличие пути без выполненной операции недостаточно.

Связь 460 · AISAFE2-WEAK-082 → AISAFE2-IMP-016

Недостаточно стойкая аутентификация привилегированного человека позволяет атакующему быть принятым за администратора и успешно вызвать инфраструктурную функцию вне собственных полномочий.

Ограничение: Только если подмена привилегированного пользователя привела к успешному вызову функции, недоступной атакующему

Связь 461 · AISAFE2-WEAK-083 → AISAFE2-IMP-011

Сетевая достижимость служебного хранилища из недоверенной сети превращает отдельно установленный дефект аутентификации или конфигурации в удалённый путь, по которому атакующий фактически получает существенную часть базы знаний.

Ограничение: Только при отдельно установленном дефекте аутентификации или конфигурации, фактическом удалённом получении существенной части базы знаний и доказательстве, что недоверенный сетевой путь был необходимым условием доступа.

Связь 462 · AISAFE2-WEAK-083 → AISAFE2-IMP-012

Сетевая достижимость внутренней точки инференса или хранилища из недоверенной сети позволяет использовать отдельно установленный дефект доступа или отсутствие штатного ограничения и довести кампанию до получения весов либо функционального извлечения модели.

Ограничение: Только при отдельно установленном дефекте доступа или обходе штатного ограничения, фактическом использовании недоверенного сетевого пути и подтверждённом получении весов либо функциональном извлечении модели.

Связь 463 · AISAFE2-WEAK-083 → AISAFE2-IMP-015

Сетевая достижимость служебного интерфейса из недоверенной сети делает отдельно установленный дефект аутентификации или конфигурации удалённо применимым, после чего атакующий успешно читает либо изменяет хранимый объект.

Ограничение: Только при отдельно установленном дефекте аутентификации или конфигурации, успешной удалённой операции чтения либо изменения и доказательстве, что недоверенный сетевой путь сделал её достижимой.

Связь 464 · AISAFE2-WEAK-083 → AISAFE2-IMP-016

Сетевая достижимость административного интерфейса из недоверенной сети делает отдельно установленный дефект аутентификации, авторизации или кода удалённо применимым, после чего атакующий успешно вызывает запрещённую функцию.

Ограничение: Только при отдельно установленном дефекте аутентификации, авторизации или кода, успешном удалённом вызове запрещённой функции и доказательстве, что недоверенный сетевой путь сделал её достижимой.

Связь 465 · AISAFE2-WEAK-083 → AISAFE2-IMP-019

Фактически достижимая из недоверенной сети внутренняя точка инференса позволяет обходить штатный внешний контур ограничения частоты; при отдельно установленном недостатке внутренних квот по субъекту или совокупного бюджета повторные прямые вызовы доводят потребление вычислений, памяти, сети либо ускорителей собственной среды выше утверждённого порога.

Ограничение: Только если доступный извне внутренний интерфейс обошёл внешний лимитер, а недостаточные внутренние квоты позволили повторными вызовами измеримо превысить бюджет ресурсов; токенный предел отдельного запроса может сохраняться

Связь 466 · AISAFE2-WEAK-084 → AISAFE2-IMP-015

Отсутствие применимого эталона оставляет в рабочей среде конкретную небезопасную настройку, например анонимный доступ; через неё субъект без полномочий успешно читает либо изменяет хранимые данные.

Ограничение: Только для точной небезопасной настройки, которая действует из-за отсутствия применимого эталона и фактически приводит к успешному чтению либо изменению хранимого объекта субъектом без полномочий.

Связь 467 · AISAFE2-WEAK-084 → AISAFE2-IMP-016

Отсутствие применимого эталона оставляет открытой конкретную служебную функцию или избыточное право процесса; через эту настройку субъект без полномочий успешно вызывает инфраструктурную функцию.

Ограничение: Только для точной открытой функции или избыточного права процесса, которые действуют из-за отсутствия применимого эталона и фактически приводят к успешному запрещённому вызову.

Связь 468 · AISAFE2-WEAK-085 → AISAFE2-IMP-031

Недостаток позволяет атакующему изменить или удалить датасет обучения, тестирования либо валидации вместе с доступными доверенными копиями, вследствие чего требуемая версия датасета становится невосстановимой.

Ограничение: Только для вызванной атакой невосстановимой утраты датасета обучения, тестирования или валидации после уничтожения доступных доверенных копий.

Связь 469 · AISAFE2-WEAK-085 → AISAFE2-IMP-033

Недостаток позволяет атакующему изменить или удалить веса, базу знаний, индекс либо критичную конфигурацию вместе с доступными доверенными версиями, вследствие чего требуемый AI-артефакт становится невосстановимым.

Ограничение: Только для вызванной атакой невосстановимой утраты критичного AI-артефакта, кроме датасета, когда отсутствует доступная доверенная версия или воспроизводимый рецепт.

Связь 470 · AISAFE2-WEAK-086 → AISAFE2-IMP-009

Утрата атрибутов доступа исходных документов при включении в базу знаний может привести к раскрытию чувствительных данных, участвующих в формировании ответов инструментов, потому что фрагменты индексируются без переноса атрибутов доступа исходных документов, извлечение не может фильтровать выдачу по правам субъекта, и закрытые для него документы попадают в контекст и ответ.

Связь 471 · AISAFE2-WEAK-087 → AISAFE2-IMP-011

Ослабленная защита производного векторного представления позволяет неразрешённому субъекту по результатам теста для точной модели и конфигурации получить либо воспроизводимо реконструировать существенную часть базы знаний или индекса.

Ограничение: Только при тесте, относящемся к конкретному представлению, с точными моделью и версией, формой вектора или параметрами квантования, доступом атакующего, вспомогательными данными, корпусом, методом и измеренным порогом существенного восстановления. Универсальная обратимость векторных представлений не заявляется.

Связь 472 · AISAFE2-WEAK-087 → AISAFE2-IMP-024

Когда эмбеддинг чувствительного источника ошибочно обрабатывается как нечувствительный, он может быть передан получателю, которому исходный класс данных не разрешён.

Ограничение: Только при доказанной связи с чувствительным источником, наследовании класса обработки, фактическом переходе вектора через неутверждённую границу доверия и отсутствии разрешения у точного получателя. Универсальное восстановление исходных данных из векторных представлений не требуется и не заявляется.

Связь 473 · AISAFE2-WEAK-088 → AISAFE2-IMP-005

Отсутствие проверяемой этической оценки позволяет выпустить точную версию, не проверенную по утверждённым нормам и затрагиваемым группам; эта версия затем фактически причиняет пользователю конкретный этически недопустимый вред.

Ограничение: Только когда отсутствие обязательной оценки позволяет выпустить точную версию, а затем зафиксировано конкретное воздействие этой версии на пользователя по заранее определённым нормам и для указанной группы.

Связь 474 · AISAFE2-WEAK-088 → AISAFE2-IMP-028

Утверждённая этическая оценка обнаружила бы конкретный эффект и заблокировала, сузила или исправила бы выпуск, но версия выходит без оценки и фактически причиняет устойчивый запрещённый либо вредный эффект группе.

Ограничение: Только если показано, что обязательная оценка обнаружила бы эффект и изменила решение о выпуске, а версия без неё фактически причинила вред затронутой группе

Связь 475 · AISAFE2-WEAK-089 → AISAFE2-IMP-007

Избыточный состав функций и полномочий инструментов, доступных агенту, может привести к обходу ограничений и предусмотренной логики агента, потому что набор инструментов и операций не ограничен задачей, ролью и этапом выполнения, и ошибка модели или инъекция приводит к действиям, которых сценарий не предусматривал.

Связь 476 · AISAFE2-WEAK-089 → AISAFE2-IMP-016

Избыточный состав функций и полномочий инструментов, доступных агенту, может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что инструменты подключены с максимальными правами и доступны на всех этапах выполнения, и вызов через агента выполняет операции с правами выше прав инициатора.

Связь 477 · AISAFE2-WEAK-089 → AISAFE2-IMP-018

Избыточные полномочия дают агенту вызвать тарифицируемые инструменты и внешние сервисы, не нужные задаче; ошибка или инъекция приводит к подтверждённым начислениям вне утверждённого бюджета.

Ограничение: Только когда цикл или действие содержит тарифицируемый инструмент либо внешний сервис и создаёт подтверждённое начисление вне утверждённого бюджета.

Связь 478 · AISAFE2-WEAK-089 → AISAFE2-IMP-024

Избыточный состав функций и полномочий инструментов, доступных агенту, может привести к утечке чувствительных данных в недоверенный контур, потому что агенту доступны инструменты отправки данных во внешние каналы независимо от задачи, и ошибка модели или инъекция выводит данные наружу.

Связь 479 · AISAFE2-WEAK-089 → AISAFE2-IMP-031

Внедрённая атакующим инструкция использует избыточно доступную разрушительную операцию для уничтожения точного набора данных и всех пригодных доверенных копий или путей воспроизводимого восстановления.

Ограничение: Только если избыточно доступный инструмент использован атакой для уничтожения точного датасета и всех пригодных копий или путей восстановления; невосстановимость подтверждена

Связь 480 · AISAFE2-WEAK-089 → AISAFE2-IMP-033

Внедрённая атакующим инструкция использует избыточно доступную разрушительную операцию для уничтожения весов, базы знаний, индекса либо критичной конфигурации и всех пригодных копий или путей воспроизводимого восстановления.

Ограничение: Только если избыточно доступный инструмент использован атакой для уничтожения точного критичного артефакта, кроме датасета, и всех пригодных путей восстановления; невосстановимость подтверждена

Связь 481 · AISAFE2-WEAK-090 → AISAFE2-IMP-015

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к несанкционированному чтению или изменению хранимых данных, потому что скомпрометированный инструмент на общем хосте получает доступ к файловой системе, секретам и данным других арендаторов с полномочиями ядра агента.

Связь 482 · AISAFE2-WEAK-090 → AISAFE2-IMP-016

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к несанкционированному доступу к функциональности инфраструктурных компонентов, потому что исполнители работают с полномочиями ядра агента и общего хоста, и компрометация одного компонента (запуск произвольного кода) распространяется на функциональность всей системы.

Связь 483 · AISAFE2-WEAK-090 → AISAFE2-IMP-019

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к чрезмерному потреблению вычислительных ресурсов системы, потому что на исполнителей не наложены ограничения вычислений, памяти и времени выполнения, и скомпрометированный или сбойный компонент исчерпывает ресурсы общего хоста.

Связь 484 · AISAFE2-WEAK-090 → AISAFE2-IMP-024

Недостаточная изоляция среды исполнения инструментов и сгенерированного кода может привести к утечке чувствительных данных в недоверенный контур, потому что скомпрометированный исполнитель, обладая сетевым доступом хоста, выводит данные во внешние каналы за пределы контролируемой зоны.

Связь 485 · AISAFE2-WEAK-091 → AISAFE2-IMP-012

Использование форматов артефактов модели, допускающих исполнение кода при загрузке, может привести к краже модели, потому что подменённый или вредоносный артефакт исполняет код при загрузке с полномочиями сервиса, и из инфраструктуры похищаются модели.

Связь 486 · AISAFE2-WEAK-091 → AISAFE2-IMP-015

Загрузчик исполняет код, встроенный в артефакт модели, непосредственно при десериализации; код действует с полномочиями сервиса и успешно читает или изменяет достижимый хранимый объект вне полномочий атакующего.

Связь 487 · AISAFE2-WEAK-091 → AISAFE2-IMP-016

Загрузчик исполняет код, встроенный в артефакт модели, непосредственно при десериализации; код действует с полномочиями сервиса и вызывает достижимую инфраструктурную функцию вне полномочий атакующего.

Связь 488 · AISAFE2-WEAK-093 → AISAFE2-IMP-020

Операция одного арендатора ошибочно получает платёжную идентичность другого; её измеренное потребление начисляется другому плательщику сверх заранее показанной или подтверждённой суммы.

Ограничение: Только если установлены фактический инициатор, ошибочно начисленный плательщик, точная операция и реальное списание выше согласованной суммы без отдельного согласия плательщика.

Связь 489 · AISAFE2-WEAK-095 → AISAFE2-IMP-020

Делегированное платёжное полномочие без лимитов суммы и частоты, привязанных к задаче, позволяет агенту фактически выполнить списания выше показанной или подтверждённой пользователем границы.

Ограничение: Только при фактически исполненном списании выше заранее показанной либо подтверждённой суммы или лимита и при отсутствии отдельного подтверждения точных параметров. Наличие платёжного средства без списания недостаточно.

Связь 490 · AISAFE2-WEAK-095 → AISAFE2-IMP-025

Неограниченное по сумме и частоте делегированное платёжное полномочие позволяет агенту выполнить денежную операцию от имени пользователя, после которой остаётся точный атрибутированный невозвратный чистый убыток.

Ограничение: Только если операция без нужного платёжного лимита исполнена и оставила подтверждённый чистый убыток после возвратов и оспаривания; потенциальные суммы не учитываются

Связь 491 · AISAFE2-IMP-002 → AISAFE2-MIT-005

Фактическая трассировка использованных источников и явный статус подтверждения не позволяют представить неподтвержденный вывод как подтвержденное основание решения; наличие источника не гарантирует истинность.

Ограничение: Не проверяет истинность ответа и не предотвращает вред, если пользователь всё же действует по неподтверждённому выводу

Связь 492 · AISAFE2-IMP-002 → AISAFE2-MIT-039

Сценарные критерии приемки до допуска отсеивают конфигурации модели, не пригодные для решения, на котором строится решение.

Связь 493 · AISAFE2-IMP-002 → AISAFE2-MIT-045

Для заданных классов существенных решений обязательная независимая проверка исходных данных и правил до завершения решения снижает вероятность вреда от ошибочного AI-вывода.

Ограничение: Только для решений, включённых владельцем процесса в обязательный блокирующий маршрут; проверяющий может ошибиться или формально одобрить вывод.

Связь 494 · AISAFE2-IMP-003 → AISAFE2-MIT-023

Корреляция событий и признаков кампании выявляет повторяющиеся цепочки вовлечения пользователей в злонамеренную схему.

Связь 495 · AISAFE2-IMP-003 → AISAFE2-MIT-038

Связывание субъектов, правила недопустимого использования и санкции усложняют повторное и распределенное злоупотребление сервисом.

Связь 496 · AISAFE2-IMP-004 → AISAFE2-MIT-010

Фильтр на клиентской или входной границе может отклонить распознанные чувствительные данные до передачи модели или внешнему получателю.

Ограничение: Если данные уже приняты AI-системой, мера ограничивает последующую передачу, но не отменяет факт передачи потребителем; нераспознанные значения могут пройти контроль.

Связь 497 · AISAFE2-IMP-004 → AISAFE2-MIT-024

После подтвержденной утечки сценарий уведомляет затронутого пользователя, отзывает или меняет доступные системе учетные данные и проверяет прекращение доступа; раскрытие при этом не отменяется.

Ограничение: Уведомление, отзыв и локализация начинаются после утечки и не отменяют состоявшееся раскрытие.

Связь 498 · AISAFE2-IMP-005 → AISAFE2-MIT-025

Шлюз требует утверждённые и версионированные этические критерии для пользовательского сценария, результаты их проверки и решение по исключениям до выпуска, снижая вероятность допуска заведомо несоответствующего поведения.

Связь 499 · AISAFE2-IMP-005 → AISAFE2-MIT-045

Для заданных классов существенных решений проверяющий до их завершения сопоставляет результат с независимыми фактами и утверждёнными этическими критериями, снижая вероятность вреда пользователю.

Ограничение: Не охватывает прямой вредный ответ модели и решения вне обязательного блокирующего маршрута; компетентный проверяющий тоже может ошибиться.

Связь 500 · AISAFE2-IMP-006 → AISAFE2-MIT-001

Проверка интерфейсного контракта отклоняет некорректную структуру и известные недопустимые сочетания до модели или следующего компонента, предотвращая один подтвержденный класс неожидаемых ответов.

Связь 501 · AISAFE2-IMP-006 → AISAFE2-MIT-002

Контрактный разбор и контекстное экранирование не позволяют неожиданному выводу стать активным кодом, командой или API-параметром; сам неожиданный ответ остается возможен.

Ограничение: Мера блокирует активную интерпретацию вывода, но неожиданный или ложный текст остается возможен.

Связь 502 · AISAFE2-IMP-006 → AISAFE2-MIT-003

Типизированная сборка контекста не позволяет недоверенному содержимому стать инструкцией на охваченном пути внедрения вредоносных инструкций.

Связь 503 · AISAFE2-IMP-006 → AISAFE2-MIT-004

Отдельный детектор и маршрут обработки фиксируют признаки обхода ограничений или внедрения вредоносных инструкций до обычного продолжения запроса.

Связь 504 · AISAFE2-IMP-006 → AISAFE2-MIT-036

Сравнение с зафиксированным базовым профилем обнаруживает дрейф и деградацию, проявляющиеся как неожиданные ответы.

Связь 505 · AISAFE2-IMP-006 → AISAFE2-MIT-040

Карантин и допуск точной версии обучающих данных блокируют обучение на недопущенном или измененном наборе.

Связь 506 · AISAFE2-IMP-006 → AISAFE2-MIT-047

Состязательные испытания и ограничения входа снижают вероятность известного по модели угроз класса ошибочного ответа.

Связь 507 · AISAFE2-IMP-006 → AISAFE2-MIT-051

При пропуске или некорректном завершении обязательной проверки завершающий обработчик не выпускает неожиданный результат как обычный или разрешающий; иные причины неожиданного ответа остаются вне связи.

Ограничение: Контроль блокирует неожиданный результат только при неполном выполнении заранее объявленного обязательного плана анализа; иные причины неожиданного ответа остаются.

Связь 508 · AISAFE2-IMP-007 → AISAFE2-MIT-002

Проверка контракта приёмника блокирует обход через интерпретацию вывода модели как команды, кода или API-параметра, но не подтверждает корректность бизнес-логики.

Ограничение: Покрыт только обход через небезопасный разбор, команды или внедрение через параметры API, а не вся бизнес-логика ядра агента.

Связь 509 · AISAFE2-IMP-007 → AISAFE2-MIT-004

Версионируемый детектор формирует сигнал при наблюдаемых признаках внедрения вредоносных инструкций или обхода ограничений и направляет запрос по отдельному маршруту; общую логику он не проверяет.

Ограничение: Детектор охватывает только наблюдаемые валидированные классы внедрения вредоносных инструкций и обхода ограничений и не проверяет общий логический обход.

Связь 510 · AISAFE2-IMP-007 → AISAFE2-MIT-016

Разрешающий список инструментов и проверка аргументов блокируют вызов или параметры вне предусмотренной бизнес-логики.

Связь 511 · AISAFE2-IMP-007 → AISAFE2-MIT-019

Неизменяемый предпросмотр и отдельное решение человека блокируют выбранное высокорисковое действие после ошибочного плана агента.

Ограничение: Шлюз человеческого подтверждения блокирует только выбранные и настроенные действия повышенного риска, а не все переходы ядра агента.

Связь 512 · AISAFE2-IMP-007 → AISAFE2-MIT-020

Независимый путь остановки отзывает полномочия и прекращает очереди и сессии после обнаруженного обхода логики.

Связь 513 · AISAFE2-IMP-007 → AISAFE2-MIT-029

Внешняя по отношению к модели точка авторизации отклоняет действие, даже если ядро агента сформировало допустимо выглядящий, но неразрешённый план.

Связь 514 · AISAFE2-IMP-007 → AISAFE2-MIT-045

Независимая предметная проверка ограничивает последствия обхода, если его результат становится решением с существенными последствиями.

Ограничение: Независимая проверка применима, только когда результат становится решением с существенными последствиями; логику ядра агента она не исправляет.

Связь 515 · AISAFE2-IMP-007 → AISAFE2-MIT-048

Проверка отправителя, целостности, получателя, задачи, срока и повтора отклоняет поддельную, измененную, ошибочно адресованную или повторную межагентскую задачу до обработки; иной обход логики ядра агента не покрывается.

Ограничение: Покрыт только обход через поддельную, измененную, ошибочно адресованную, просроченную или повторную межагентскую задачу; новое сообщение вредоносного аутентифицированного агента не блокируется по смыслу.

Связь 516 · AISAFE2-IMP-007 → AISAFE2-MIT-051

Ошибка, отмена или отсутствие обязательной стадии политики не превращаются в разрешение и не обходят принудительный завершающий обработчик; всю бизнес-логику ядра агента мера не проверяет.

Ограничение: Контроль закрывает только обход через неполное выполнение обязательной стадии; всю бизнес-логику, семантику политики и корректность плана ядра агента он не проверяет.

Связь 517 · AISAFE2-IMP-008 → AISAFE2-MIT-003

Результат инструмента, RAG или внешнего агента сохраняет тип и происхождение, остаётся данными, не меняет политику или доступные возможности, а действие проходит внешнюю авторизацию.

Ограничение: Разделение полномочий ограничивает путь смешения контекстов, но не делает смысл результата инструмента или сам инструмент корректным.

Связь 518 · AISAFE2-IMP-008 → AISAFE2-MIT-007

Шлюз допуска и карантин не публикуют в рабочий RAG источник без подтвержденного происхождения и назначения.

Связь 519 · AISAFE2-IMP-008 → AISAFE2-MIT-016

Реестр доступных инструментов и схема аргументов блокируют непредусмотренный инструмент или форму его вызова.

Связь 520 · AISAFE2-IMP-008 → AISAFE2-MIT-021

Правила записи, происхождение, TTL и удаление снижают вероятность переноса недоверенного результата через долговременную память.

Связь 521 · AISAFE2-IMP-008 → AISAFE2-MIT-048

Внешнее агентское сообщение принимается только от проверенного отправителя для точного получателя и задачи, в срок и без повтора; вредное новое сообщение корректно аутентифицированного агента остается возможным.

Ограничение: Покрыто принятие поддельного, просроченного или повторного внешнего агентского сообщения; корректно аутентифицированный скомпрометированный агент остаётся отдельным риском.

Связь 522 · AISAFE2-IMP-009 → AISAFE2-MIT-008

Проверка прав после извлечения, ранжирования и кэширования не возвращает фрагмент субъекту без разрешения на этот фрагмент.

Связь 523 · AISAFE2-IMP-009 → AISAFE2-MIT-010

Ненужные или запрещённые чувствительные поля удаляются до передачи в контекст модели или инструмента; авторизацию извлечения и проверку исходящей передачи это не заменяет.

Ограничение: Минимизация действует до границы AI-контура, но не заменяет авторизацию извлечения и контроль исходящей передачи.

Связь 524 · AISAFE2-IMP-009 → AISAFE2-MIT-032

Классификация фактического исходящего содержимого и получателя блокирует передачу чувствительных данных через охваченный коннектор или инструмент.

Связь 525 · AISAFE2-IMP-010 → AISAFE2-MIT-004

Опциональный детектор выявляет только валидированный класс попыток извлечения через внедрение вредоносных инструкций или обход ограничений и направляет сигнал; один эвристический детектор конфиденциальность не обеспечивает.

Ограничение: Детектор даёт сигнал лишь по проверенным способам выманивания системного промпта или инструкций инструментов; нераскрытие он не гарантирует

Связь 526 · AISAFE2-IMP-010 → AISAFE2-MIT-015

Контролируемое хранение и выдача конфигурации исключают прямое чтение промпта и секретов неавторизованным субъектом; поведенческое восстановление остаётся риском.

Связь 527 · AISAFE2-IMP-011 → AISAFE2-MIT-011

Разделение арендаторов и ограничение экспорта сокращают число векторов, доступных субъекту, и масштаб реконструкции, но не устраняют атаку инверсии эмбеддингов для векторов, полученных по разрешённому пути.

Ограничение: Изоляция и контроль экспорта сокращают доступный набор векторов и масштаб последствий, но не предотвращают восстановление исходных данных по векторам, полученным разрешённым путём.

Связь 528 · AISAFE2-IMP-011 → AISAFE2-MIT-033

Блокирующий шлюз не выпускает изменение RAG, векторного хранилища или памяти без актуальных подтверждений применимых контролей доступа, происхождения, мониторинга и возврата.

Связь 529 · AISAFE2-IMP-012 → AISAFE2-MIT-043

Ограничение чтения и экспорта, шифрование и разделение ключей блокируют неавторизованное получение охваченного артефакта модели.

Связь 530 · AISAFE2-IMP-012 → AISAFE2-MIT-044

Сквозной учёт бюджета и корреляция запросов выявляют серию, похожую на функциональное извлечение через интерфейс; при обнаружении применяются ограничения, предусмотренные политикой.

Ограничение: Только для наблюдаемого извлечения через интерфейс; прямое получение весов и уже завершённое копирование мера не предотвращает.

Связь 531 · AISAFE2-IMP-013 → AISAFE2-MIT-010

Минимизация, редактирование и токенизация удаляют распознанные чувствительные данные до их передачи в обучающий AI-контур.

Связь 532 · AISAFE2-IMP-013 → AISAFE2-MIT-042

Методы защиты приватности и проверки запоминания, определения участия записи в обучении и инверсии снижают вероятность извлечения записей из модели.

Связь 533 · AISAFE2-IMP-014 → AISAFE2-MIT-015

Исключение секретов и внутренних адресов из промпта и защищаемой конфигурации уменьшает чувствительность сведений, доступных при раскрытии; архитектурный вывод не исключается.

Ограничение: Исключение секретов и внутренних адресов снижает чувствительность конфигурации, но архитектуру всё ещё можно восстановить по другим признакам.

Связь 534 · AISAFE2-IMP-015 → AISAFE2-MIT-012

Несовпадение подписи или хеша выявляет измененный артефакт модели до загрузки; сам доступ к хранилищу или изменение данных эта мера не предотвращает.

Ограничение: Проверяется целостность артефакта модели перед загрузкой; чтение или запись в хранилище общего назначения и повышение полномочий не предотвращаются.

Связь 535 · AISAFE2-IMP-015 → AISAFE2-MIT-034

Политика продвижения разрешает среде исполнения только одобренную ссылку на версию и блокирует произвольную подмену в реестре.

Ограничение: Контроль действует только на реестр продвижения версий и утверждённую рабочую ссылку, а не на произвольные хранимые данные.

Связь 536 · AISAFE2-IMP-015 → AISAFE2-MIT-049

Привязанные к контексту пространства имён и состояние среды выполнения, а также запрет повторного назначения до очистки блокируют доступ к остаточному состоянию другой границы доверия; общую авторизацию хранилища мера не заменяет.

Ограничение: Мера разделяет только объявленные пространства запросов, сессий и задач и повторно используемую среду выполнения; общую авторизацию хранилищ она не заменяет.

Связь 537 · AISAFE2-IMP-015 → AISAFE2-MIT-052

После применимого срока или вывода компонента мера прекращает новые вызовы, отзывает права и удаляет либо изолирует охваченные копии, не оставляя старый путь доступа в управляемой области; общее повышение привилегий остается вне связи.

Ограничение: Мера закрывает доступ через неотозванные объекты и полномочия в управляемом жизненном цикле, но не заменяет общую защиту от повышения привилегий и не обещает физическое стирание на всех слоях.

Связь 538 · AISAFE2-IMP-016 → AISAFE2-MIT-012

Запрет исполняемой десериализации и перехода к менее безопасному загрузчику блокирует конкретный путь запуска инфраструктурной функции через артефакт модели; иные пути не покрываются.

Ограничение: Запрет опасной десериализации предотвращает выполнение кода только на пути загрузки исполняемого артефакта модели; другие инфраструктурные функции и причины RCE остаются вне связи.

Связь 539 · AISAFE2-IMP-016 → AISAFE2-MIT-017

Изолированная среда инструментов ограничивает сеть, файлы, секреты и время процесса при компрометации инструмента.

Связь 540 · AISAFE2-IMP-016 → AISAFE2-MIT-018

Краткоживущая идентичность с минимальными правами не дает агенту вызвать функциональность за пределами делегированной области.

Связь 541 · AISAFE2-IMP-016 → AISAFE2-MIT-030

Брокер коннекторов проверяет получателя, операцию и ресурс и отклоняет делегированный вызов вне разрешенной области.

Связь 542 · AISAFE2-IMP-016 → AISAFE2-MIT-031

Изолированный браузерный профиль или виртуальная машина ограничивают доступ скомпрометированного браузерного агента или агента управления компьютером к сети, файлам и учётным данным.

Связь 543 · AISAFE2-IMP-017 → AISAFE2-MIT-006

Сквозной лимит токенов и времени прекращает обработку до превышения установленного бюджета.

Связь 544 · AISAFE2-IMP-018 → AISAFE2-MIT-006

Отдельный бюджет платных инструментов и стоимость на субъект блокируют новый вызов после достижения лимита.

Связь 545 · AISAFE2-IMP-019 → AISAFE2-MIT-006

Лимиты запросов, шагов, времени и вычислительного бюджета не позволяют одному процессу бесконтрольно потреблять охваченные ресурсы.

Связь 546 · AISAFE2-IMP-019 → AISAFE2-MIT-020

Аварийная остановка прекращает работу неуправляемого агента, очереди и связанные полномочия после обнаружения перерасхода.

Связь 547 · AISAFE2-IMP-019 → AISAFE2-MIT-023

Корреляция событий ресурса, числа шагов и длительности выявляет наблюдаемые аномальные циклы и распределенное потребление для установленного маршрута обработки.

Ограничение: Корреляция видит только наблюдаемые аномалии потребления ресурсов и не заменяет принудительное применение бюджета до первого действия.

Связь 548 · AISAFE2-IMP-020 → AISAFE2-MIT-006

Персональный бюджет и уведомление ограничивают начисления только там, где оператор AI-системы управляет пользовательским тарифом.

Ограничение: Не ограничивает счета внешних сервисов и расходы, которые оператор AI-системы не тарифицирует пользователю

Связь 549 · AISAFE2-IMP-021 → AISAFE2-MIT-006

Лимит числа шагов, повторов и времени завершает запрос до бесконечного цикла.

Связь 550 · AISAFE2-IMP-021 → AISAFE2-MIT-020

Независимый механизм аварийной остановки прекращает уже возникший цикл и отзывает полномочия агента.

Связь 551 · AISAFE2-IMP-022 → AISAFE2-MIT-006

Предварительная проверка бюджета, учёт выполненной внутренней работы и жёсткие ограничения среды выполнения снижают вероятность недоступности из-за усиления нагрузки одним принятым запросом; общая ёмкость, отказоустойчивость и внешняя перегрузка остаются вне связи.

Ограничение: Внутренний бюджет ограничивает ресурсную амплификацию среды выполнения, но не заменяет планирование мощности, отказоустойчивость, SLO и защиту от внешней нагрузки.

Связь 552 · AISAFE2-IMP-022 → AISAFE2-MIT-024

Восстановление: проверенное возвращение в рабочее состояние и разбор причины снижают длительность отказа и риск его повторения.

Связь 553 · AISAFE2-IMP-022 → AISAFE2-MIT-024

Реагирование: исполняемый сценарий локализует отказавшую зависимость и координирует действия до безопасного возврата в работу.

Связь 554 · AISAFE2-IMP-022 → AISAFE2-MIT-046

Матрица отказов переводит приложение в проверенный безопасный режим при тайм-ауте, ошибке или частичном ответе AI-зависимости.

Ограничение: Безопасная деградация ограничивает последствия отказа AI-зависимости для охваченного приложения, но не восстанавливает доступность самой зависимости, её мощность или полный сервис.

Связь 555 · AISAFE2-IMP-023 → AISAFE2-MIT-037

Маршрутизация по классу риска, контроль очереди и метрики нагрузки снижают вероятность перегрузки и усталости проверяющих.

Связь 556 · AISAFE2-IMP-024 → AISAFE2-MIT-004

На пути утечки детектор формирует сигнал при наблюдаемых признаках внедрения вредоносных инструкций или обхода ограничений; он не классифицирует чувствительные данные и не заменяет DLP-контроль.

Ограничение: Сигнал внедрения вредоносных инструкций или обхода ограничений не классифицирует данные и не заменяет DLP.

Связь 557 · AISAFE2-IMP-024 → AISAFE2-MIT-010

Классификация, минимизация и редактирование удаляют запрещенные данные до передачи во внешний или иной недоверенный AI-контур.

Связь 558 · AISAFE2-IMP-024 → AISAFE2-MIT-024

После подтвержденной утечки сценарий уведомляет затронутых субъектов, отзывает или меняет доступные системе учетные данные и локализует путь раскрытия; сам факт утечки остается.

Ограничение: Реагирование после утечки уведомляет, отзывает и локализует, но не обращает раскрытие вспять.

Связь 559 · AISAFE2-IMP-024 → AISAFE2-MIT-032

Проверка фактического получателя после перенаправлений и политика исходящих данных блокируют запрещенную передачу через коннектор.

Связь 560 · AISAFE2-IMP-024 → AISAFE2-MIT-049

Подтверждённая очистка или изоляция до повторного назначения ресурса блокирует межарендаторскую утечку через остаточное состояние общей среды инференса; иные пути утечки не покрываются.

Ограничение: Покрыта только межарендаторская утечка через остаточное состояние общей среды инференса; другие пути передачи в недоверенный контур остаются.

Связь 561 · AISAFE2-IMP-024 → AISAFE2-MIT-051

Если контроль данных объявлен обязательной стадией, его неполное выполнение блокирует передачу результата; незарегистрированные пути и данные вне ее области не покрываются.

Ограничение: Контроль блокирует передачу только на зарегистрированном пути, где проверка данных объявлена обязательной стадией; остальные пути и неохваченные классы данных остаются.

Связь 562 · AISAFE2-IMP-024 → AISAFE2-MIT-054

На каждом объявленном пути обязательный блокирующий шлюз сопоставляет фактическую цель, класс данных, место, получателя, конечную точку и срок с актуальным решением и блокирует неизвестный или запрещённый поток.

Связь 563 · AISAFE2-IMP-031 → AISAFE2-MIT-050

Разделение полномочий и независимые неизменяемые или защищенные от удаления копии не позволяют одной скомпрометированной учетной записи либо вредоносной операции уничтожить все восстановимые версии датасета.

Связь 564 · AISAFE2-IMP-032 → AISAFE2-MIT-010

Минимизация, редактирование или токенизация уменьшают содержимое, доступное при смешении контекстов, но не исправляют саму изоляцию.

Ограничение: Минимизация сокращает чувствительность ошибочно выданного артефакта, но не создает изоляцию контекстов и не предотвращает саму выдачу.

Связь 565 · AISAFE2-IMP-032 → AISAFE2-MIT-023

Версионируемое правило сопоставляет проверенный субъект и контекст запроса с владельцем ответа, объекта, пространства имён кэша или контрольной метки среды выполнения и формирует сигнал при несовпадении; неинструментированные пути остаются вне наблюдения.

Ограничение: Корреляция обнаруживает нарушение только на инструментированных путях и после появления сигнала; она не предотвращает смешение.

Связь 566 · AISAFE2-IMP-032 → AISAFE2-MIT-024

После подтвержденного смешения сценарий локализует путь, отзывает доступ, изолирует затронутые пространства имён и среды выполнения и сохраняет подтверждения без расширения утечки; состоявшееся раскрытие не обращается вспять.

Ограничение: Локализация и отзыв ограничивают развитие инцидента после обнаружения, но не обращают уже состоявшееся раскрытие вспять.

Связь 567 · AISAFE2-IMP-032 → AISAFE2-MIT-049

Привязка к субъекту, запросу или задаче, отдельные пространства имён, ключ кэша, учитывающий контекст, и очистка повторно используемой среды выполнения блокируют выдачу артефакта другому изолированному контексту на охваченном пути.

Ограничение: Мера предотвращает смешение на объявленных путях пространства имён, кэша и повторно используемой среды выполнения, но не охватывает каждый сохраняемый артефакт или внешний компонент.

Связь 568 · AISAFE2-IMP-033 → AISAFE2-MIT-009

Версионирование базы знаний и индекса с атомарным откатом сохраняет пригодный предыдущий выпуск и предотвращает невосстановимую утрату этих объектов при повреждении рабочей версии.

Ограничение: Предотвращает невосстановимую утрату базы знаний или индекса только при сохранённой пригодной версии и проверенном атомарном откате; модель, конфигурацию и уничтожение всех копий мера не покрывает

Связь 569 · AISAFE2-IMP-033 → AISAFE2-MIT-034

Управляемый реестр с сохранённым пригодным артефактом предыдущей версии позволяет откатить повреждённый выпуск модели до невосстановимой утраты этого артефакта.

Ограничение: Предотвращает невосстановимую утрату версии модели лишь при сохранённом пригодном артефакте и закрытых обходных путях его загрузки; реестр сам по себе не создаёт независимую копию

Связь 570 · AISAFE2-IMP-033 → AISAFE2-MIT-068

Изолированные неизменяемые или защищённые от удаления версии вне полномочий рабочего контура не позволяют одной скомпрометированной учётной записи либо вредоносной операции уничтожить все восстановимые копии критичного недатасетного AI-актива.

Связь 571 · AISAFE2-WEAK-001 → AISAFE2-MIT-045

Независимый предметный проверяющий принимает решение по фактам вне модели, компенсируя склонность к автоматическому доверию.

Связь 572 · AISAFE2-WEAK-002 → AISAFE2-MIT-005

Фактическая трассировка источников и явный статус подтверждения устраняют отсутствие происхождения у охваченного вывода.

Связь 573 · AISAFE2-WEAK-003 → AISAFE2-MIT-002

Контракт приёмника, разрешающий список и безопасное представление создают принудительную границу для охваченного вывода модели.

Связь 574 · AISAFE2-WEAK-003 → AISAFE2-MIT-056

Категориальная проверка создаёт обязательную границу допустимости вывода на охваченных маршрутах.

Ограничение: Мера проверяет категориальную допустимость вывода, но не проверяет структуру, контекстное кодирование и авторизацию конкретного получателя.

Связь 575 · AISAFE2-WEAK-004 → AISAFE2-MIT-002

Контекстное экранирование, разбор по типу и изолированное исполнение устраняют небезопасную обработку вывода на охваченном пути.

Связь 576 · AISAFE2-WEAK-004 → AISAFE2-MIT-017

На прикладном пути, где AI-вывод может стать кодом, командой или вызовом инструмента, отдельная изоляция или исполнитель не дают ему исполняться в доверенном процессе с его полномочиями; обычное небезопасное средство отображения и истинность содержимого остаются вне связи.

Ограничение: Изоляция действует при исполнении кода и команд инструментом; отображение в браузере и другие потребители требуют отдельного контекстного кодирования.

Связь 577 · AISAFE2-WEAK-005 → AISAFE2-MIT-005

Пользовательский интерфейс показывает происхождение, неподтверждённые части и маршрут проверки вместо неразмеченного AI-вывода.

Связь 578 · AISAFE2-WEAK-007 → AISAFE2-MIT-003

Раздельные типизированные каналы инструкций и данных не дают недоверенному контенту стать инструкцией на охваченном пути.

Связь 579 · AISAFE2-WEAK-007 → AISAFE2-MIT-066

Неизменяемая версия цели не позволяет пользовательскому, RAG-, инструментальному или межагентскому содержимому стать изменением управляющей цели на охваченном пути.

Ограничение: Мера защищает типизированную цель, политику и разрешённые возможности, но не разделяет все остальные доверенные инструкции и недоверенные фрагменты контекста.

Связь 580 · AISAFE2-WEAK-008 → AISAFE2-MIT-029

Авторизация действия вне модели не полагается на внутренний конфликт между возможностью и ограничением безопасности и блокирует неразрешённый внешний эффект.

Связь 581 · AISAFE2-WEAK-009 → AISAFE2-MIT-029

Внешняя точка принятия решения по политике проверяет полномочие независимо от того, как модель сбалансировала инструкцию и запрет.

Связь 582 · AISAFE2-WEAK-010 → AISAFE2-MIT-029

Внешняя авторизация не доверяет ограничениям, размытым контекстным обучением, и отклоняет действие по фактическому субъекту и ресурсу.

Связь 583 · AISAFE2-WEAK-011 → AISAFE2-MIT-029

Детерминированная внешняя политика ограничивает последствия поверхностной настройки безопасного поведения модели при переходе к внешнему действию.

Связь 584 · AISAFE2-WEAK-012 → AISAFE2-MIT-012

Безопасный формат, идентичность фактически открытых байтов и проверенный загрузчик устраняют непроверенный или исполняемый путь загрузки артефакта модели; остальную цепочку поставки мера не проверяет.

Ограничение: Проверяется путь загрузки артефакта модели; данные обучения, зависимости и сборка требуют отдельных проверок.

Связь 585 · AISAFE2-WEAK-012 → AISAFE2-MIT-013

Блокирующая проверка кода, образов, моделей, данных и инструментов связывает точный выпуск с проверенным составом.

Связь 586 · AISAFE2-WEAK-013 → AISAFE2-MIT-035

Манифест происхождения фиксирует точные версии данных, моделей, кода и инструментов вместо непроверяемой цепочки.

Связь 587 · AISAFE2-WEAK-013 → AISAFE2-MIT-040

Проверяемое происхождение связывает синтетические и обучающие данные с точным запуском.

Ограничение: Прослеживаются обучающие и оценочные наборы; происхождение RAG-корпуса и рабочих данных эта мера не устанавливает.

Связь 588 · AISAFE2-WEAK-013 → AISAFE2-MIT-060

Разрешённые переходы стадий и неизменяемые версии ограничивают продвижение данных.

Ограничение: Проверяются версии и переходы между стадиями, но не происхождение исходных данных и преобразования внутри стадии.

Связь 589 · AISAFE2-WEAK-014 → AISAFE2-MIT-007

Карантин и шлюз допуска не публикуют непроверенный источник или контур обратной связи в рабочий RAG.

Связь 590 · AISAFE2-WEAK-014 → AISAFE2-MIT-021

Политика записи, сведения о происхождении, TTL и удаление не позволяют непроверенной обратной связи бессрочно закрепиться в памяти.

Связь 591 · AISAFE2-WEAK-014 → AISAFE2-MIT-040

Шлюз допуска блокирует синтетический или обучающий материал без отдельного происхождения.

Ограничение: Шлюз допуска охватывает обучающие и оценочные данные, но не запись в RAG-корпус и память агента.

Связь 592 · AISAFE2-WEAK-014 → AISAFE2-MIT-060

Шлюз продвижения блокирует прямую и обратную запись корпуса в обход разрешённой стадии.

Ограничение: Мера покрывает продвижение набора данных или корпуса, но не память и не произвольные пути обратной связи.

Связь 593 · AISAFE2-WEAK-015 → AISAFE2-MIT-039

Сценарные испытания точной модели выявляют непригодность, связанную с переобучением, до допуска; само свойство модели не считается устранённым.

Связь 594 · AISAFE2-WEAK-016 → AISAFE2-MIT-018

Отдельная машинная идентичность и минимальные права устраняют широкие полномочия задания обучения.

Связь 595 · AISAFE2-WEAK-016 → AISAFE2-MIT-041

Минимальная идентичность запуска и связь результата с допущенными данными и кодом устраняют избыточный доступ в охваченном конвейере обучения.

Связь 596 · AISAFE2-WEAK-016 → AISAFE2-MIT-043

Разделение чтения, исполнения и экспорта устраняет избыточный доступ к охваченным артефактам модели.

Связь 597 · AISAFE2-WEAK-017 → AISAFE2-MIT-039

Критерии пригодности и сценарные тесты исключают неподходящую модель из допуска к целевому сценарию.

Связь 598 · AISAFE2-WEAK-018 → AISAFE2-MIT-044

Бюджет и ограничение информативности интерфейса уменьшают возможность восстановить модель по наблюдаемому соответствию входов и выходов.

Ограничение: Мера ограничивает объём запросов и необязательный диагностический сигнал, но разрешённые ответы сохраняют устойчивые признаки, по которым можно распознавать модель, настройки или подключённые возможности.

Связь 599 · AISAFE2-WEAK-018 → AISAFE2-MIT-044

Снижение точности и рандомизация там, где это допустимо, удаляют необязательный высокоточный детерминированный сигнал.

Ограничение: Снижение точности и рандомизация удаляют только необязательный высокоточный воспроизводимый сигнал; базовая наблюдаемость соответствия входов и выходов объединённой карточки сохраняется.

Связь 600 · AISAFE2-WEAK-019 → AISAFE2-MIT-044

Интерфейс исключает логиты и логарифмы вероятностей из обычного ответа.

Ограничение: Исключение логитов не исключает раскрытие полных векторов оценок или отдельного показателя уверенности.

Связь 601 · AISAFE2-WEAK-019 → AISAFE2-MIT-044

Интерфейс не возвращает полные векторы вероятностей и оценок уверенности без отдельного обоснования.

Ограничение: Минимальный ответ не содержит полных векторов оценок; логиты и отдельный показатель уверенности требуют самостоятельного ограничения.

Связь 602 · AISAFE2-WEAK-019 → AISAFE2-MIT-044

Минимальный формат ответа исключает сигнал уверенности из обычного интерфейса обращения к модели.

Ограничение: Исключение показателя уверенности не исключает раскрытие полных векторов оценок и логитов.

Связь 603 · AISAFE2-WEAK-022 → AISAFE2-MIT-008

Проверка прав на фрагмент после последнего преобразования устраняет общий доступ к данным на охваченном пути чтения RAG.

Ограничение: Контроль закрывает выбранный путь чтения RAG, но не согласует права между обучением, хранением, CI/CD и эксплуатацией и не устраняет транзитивный доступ на других стадиях.

Связь 604 · AISAFE2-WEAK-022 → AISAFE2-MIT-018

Отдельная машинная идентичность и пересмотр эффективных прав сужают доступ рабочей нагрузки.

Ограничение: Мера покрывает машинные идентичности и их эффективные права, но не политику доступа людей и не все стадии жизненного цикла.

Связь 605 · AISAFE2-WEAK-022 → AISAFE2-MIT-041

Отдельная идентичность и минимальные права задания устраняют широкий доступ в охваченном конвейере обучения.

Ограничение: Отдельная идентичность и минимальные права ограничивают только охваченный конвейер обучения; общие учётные записи и хранилища на других стадиях требуют отдельной проверки.

Связь 606 · AISAFE2-WEAK-022 → AISAFE2-MIT-043

Разделение чтения, исполнения и экспорта устраняет широкий доступ к охваченным артефактам модели.

Ограничение: Разделение прав защищает выбранные артефакты модели, но не задаёт сквозную модель субъектов, объектов и операций для остальных данных и стадий.

Связь 607 · AISAFE2-WEAK-022 → AISAFE2-MIT-060

Раздельные права записи стадий ограничивают изменение данных утверждёнными переходами.

Ограничение: Ограничены права записи при переходах данных между стадиями; чтение моделей, артефактов и инструментов эта мера не регулирует.

Связь 608 · AISAFE2-WEAK-022 → AISAFE2-MIT-061

Совместная проверка IAM, ACL, политик и сети закрывает широкую доступность хранилищ.

Ограничение: Мера ограничивает доступ к хранилищам, но не задаёт права заданий, жизненный цикл доступа людей и другие поверхности вне хранилищ.

Связь 609 · AISAFE2-WEAK-023 → AISAFE2-MIT-011

Шифрование и изоляция устраняют открытое хранение только для охваченной базы знаний или векторного хранилища.

Ограничение: Шифруются охваченные векторные хранилища и их копии; другие датасеты, память и экспортные файлы проверяются отдельно.

Связь 610 · AISAFE2-WEAK-023 → AISAFE2-MIT-043

Шифрование хранилища и контроль ключей устраняют открытое хранение только для охваченного артефакта модели.

Ограничение: Шифруются охваченные копии артефакта модели; RAG-корпус, память агента и наборы данных остаются вне этой меры.

Связь 611 · AISAFE2-WEAK-023 → AISAFE2-MIT-062

Подтверждённое шифрование каждой управляемой копии устраняет открытое хранение в заявленной области.

Связь 612 · AISAFE2-WEAK-024 → AISAFE2-MIT-043

Шифрование и закрытый маршрут получения устраняют незашифрованную передачу только для охваченного артефакта модели.

Ограничение: Защищён маршрут передачи артефакта модели; каналы инструментов, RAG и других получателей эта мера не шифрует.

Связь 613 · AISAFE2-WEAK-024 → AISAFE2-MIT-063

Защищённый протокол с проверкой удалённой стороны устраняет незашифрованную передачу на охваченном канале.

Связь 614 · AISAFE2-WEAK-025 → AISAFE2-MIT-018

Аутентифицированная доставка краткоживущего полномочия ограничивает выдаваемые машинные учётные данные.

Ограничение: Мера покрывает выдаваемые машинные учётные данные и их доставку, но не все прикладные секреты и материалы KMS.

Связь 615 · AISAFE2-WEAK-025 → AISAFE2-MIT-043

Раздельные ключи и полномочия чтения и расшифрования устраняют слабое управление ключами только для охваченного артефакта модели.

Ограничение: Разделены ключи и права для артефакта модели; секреты инструментов, поставщиков и других данных остаются вне этой меры.

Связь 616 · AISAFE2-WEAK-025 → AISAFE2-MIT-064

Инвентарь, защищённые хранение и доставка, срок, ротация и отзыв устраняют неуправляемый жизненный цикл секрета.

Ограничение: Жизненный цикл секретов управляется, но права на шифротекст, ключ и расшифрование нужно разделять отдельно.

Связь 617 · AISAFE2-WEAK-026 → AISAFE2-MIT-038

Сквозная корреляция и единое ограничение кампании компенсируют часть последствий отсутствующей квоты.

Ограничение: Сквозная корреляция и единое ограничение кампании не гарантируют жёсткий лимит на каждом интерфейсе.

Связь 618 · AISAFE2-WEAK-026 → AISAFE2-MIT-044

Сквозной бюджет запросов по субъекту и интерфейсу вводит отсутствующее ограничение частоты для извлечения модели.

Связь 619 · AISAFE2-WEAK-027 → AISAFE2-MIT-038

Корреляция, санкции и ограничение кампании повышают стоимость продолжения злоупотребления.

Ограничение: Корреляция, санкции и ограничение кампании не изменяют исходную цену запроса и экономику сервиса.

Связь 620 · AISAFE2-WEAK-027 → AISAFE2-MIT-044

Бюджет и санкции ограничивают объем дешевых запросов, не меняя саму экономику сервиса.

Связь 621 · AISAFE2-WEAK-028 → AISAFE2-MIT-022

Полные события со сквозным идентификатором корреляции создают отсутствующую наблюдаемость AI-цепочки.

Ограничение: Связанные события позволяют восстановить цепочку запросов, но признаки кампании извлечения модели надо выявлять отдельными правилами.

Связь 622 · AISAFE2-WEAK-028 → AISAFE2-MIT-023

Корреляционные правила и проверяемые сценарии обнаружения создают мониторинг запросов на признаки кампании извлечения.

Ограничение: Правила выявляют кампанию извлечения, если события полны и связаны между собой; сбор и сохранение событий требуют отдельной меры.

Связь 623 · AISAFE2-WEAK-028 → AISAFE2-MIT-036

Контроль дрейфа может сформировать сигнал для злоупотребления, проявляющегося измеримым отклонением рабочих данных или поведения.

Ограничение: Контроль дрейфа не восполняет неполноту событий, сквозной корреляции субъектов и специализированных правил атак.

Связь 624 · AISAFE2-WEAK-030 → AISAFE2-MIT-044

Если проверяемая метка внедрена до распространения и привязана к точной версии, она даёт дополнительный сигнал атрибуции подозрительной копии.

Ограничение: Бюджеты запросов сами не создают метку; без неё связь не действует. Даже внедрённая метка служит атрибуции, а не предотвращению копирования.

Связь 625 · AISAFE2-WEAK-032 → AISAFE2-MIT-043

Разделение прав на чтение и экспорт хранимых артефактов модели и управление их ключами затрудняют получение артефакта через учтённое хранилище.

Ограничение: Шифрование копий затрудняет извлечение артефакта с носителя; физический доступ к работающему хосту, памяти и GPU остаётся отдельным риском.

Связь 626 · AISAFE2-WEAK-033 → AISAFE2-MIT-042

Минимизация, дедупликация, методы защиты приватности и проверки запоминания снижают вероятность выдачи запомненной записи, не гарантируя её отсутствие.

Ограничение: Снижение и проверка запоминания не исключают воспроизведения редкой записи в ответе модели.

Связь 627 · AISAFE2-WEAK-033 → AISAFE2-MIT-042

Испытания на извлечение и определение участия записи выявляют остаточную утечку; ограничение вклада, подходящие методы обучения и блокировка выпуска при провале испытаний снижают риск её раскрытия.

Ограничение: Проверки участия записи и вывода атрибутов не проверяют дословное воспроизведение её содержимого.

Связь 628 · AISAFE2-WEAK-035 → AISAFE2-MIT-027

Назначение владельцев данных, системы, подтверждений и риска закрывает часть пробела в управлении, но не формирует полноценную функцию управления данными.

Ограничение: Владелец назначен, но правила классификации, допустимого использования и хранения данных эта мера не устанавливает.

Связь 629 · AISAFE2-WEAK-035 → AISAFE2-MIT-052

Инвентарь, назначение, владелец, срок и решение о завершении вводят отсутствующее управление жизненным циклом охваченных AI-активов, но не формируют полноценную функцию управления данными.

Ограничение: Инвентарь и срок хранения охвачены; классификацию чувствительности и решение о допустимости использования данных нужно принимать отдельно.

Связь 630 · AISAFE2-WEAK-036 → AISAFE2-MIT-010

Классификация выявляет запрещённые данные, а минимизация, очистка и маскирование исключают их до передачи в AI-контур.

Связь 631 · AISAFE2-WEAK-037 → AISAFE2-MIT-042

Дедупликация и выбранный метод защиты приватности вводятся в точный процесс обучения вместо их отсутствия.

Связь 632 · AISAFE2-WEAK-039 → AISAFE2-MIT-032

На контролируемой исходящей границе коннектора проверка фактических данных и получателя может заблокировать передачу чувствительного или избыточного результата внешнему адресату.

Ограничение: Связь охватывает только исходящую передачу через контролируемый коннектор. Выдача инструментом данных в модель, интерфейс, память или журнал требует отдельных проверок.

Связь 633 · AISAFE2-WEAK-040 → AISAFE2-MIT-010

Маскирование, редактирование или токенизация вводятся до передачи данных в AI-контур.

Связь 634 · AISAFE2-WEAK-041 → AISAFE2-MIT-015

Управляемый доступ к конфигурации промпта и секретам закрывает прямой неавторизованный путь чтения.

Связь 635 · AISAFE2-WEAK-041 → AISAFE2-MIT-064

Управление связанными секретами сокращает риск раскрытия защищённой конфигурации промптов и инструментов.

Ограничение: Мера управляет только связанными секретами и компенсирует часть последствий, но не защищает сами системные промпты и инструкции инструментов.

Связь 636 · AISAFE2-WEAK-042 → AISAFE2-MIT-039

Приемочные критерии исключают слабую для целевого сценария модель из допуска.

Связь 637 · AISAFE2-WEAK-043 → AISAFE2-MIT-029

Внешняя авторизация блокирует неразрешённое действие независимо от наличия или качества запрета в системном промпте.

Связь 638 · AISAFE2-WEAK-044 → AISAFE2-MIT-013

Проверка компонентов в CI/CD вводит отсутствующую практику безопасности для составного выпуска.

Ограничение: Проверка состава перед выпуском не охватывает проектирование, безопасную реализацию и изменения во время эксплуатации.

Связь 639 · AISAFE2-WEAK-044 → AISAFE2-MIT-025

Блокирующий шлюз выпуска не допускает выпуск без заданных проверок, владельцев и решения по исключениям.

Ограничение: Шлюз проверяет подтверждения перед выпуском, но не выполняет разработку, тестирование и сопровождение системы.

Связь 640 · AISAFE2-WEAK-044 → AISAFE2-MIT-034

Управляемые продвижение, откат и отзыв точной версии закрывают модельную ветвь небезопасного выпуска.

Ограничение: Мера покрывает модельную версию, но не код, данные, инструменты, конфигурацию и остальные компоненты составного выпуска.

Связь 641 · AISAFE2-WEAK-044 → AISAFE2-MIT-053

Архитектура, потоки данных, модель угроз и условия пересмотра поддерживаются в версионируемом виде и создают необходимые входные данные для проектирования защиты выпуска; остальные части жизненного цикла безопасной разработки эта связь не покрывает.

Ограничение: Архитектура и модель угроз задают входы для защиты; реализацию, тестирование и безопасное развёртывание они не выполняют.

Связь 642 · AISAFE2-WEAK-044 → AISAFE2-MIT-071

Независимая ручная проверка и выбранные проверки точных commit и build блокируют выпуск дефекта выше утверждённого порога и требуют исправления с повтором исходной проверки.

Ограничение: Покрыт только выпуск собственного кода и бизнес-логики; вне связи остаются модели, зависимости, эффективная конфигурация, поэтапное развёртывание, эксплуатация и откат всей AI-системы.

Связь 643 · AISAFE2-WEAK-045 → AISAFE2-MIT-055

Ролевой цикл пересмотра и отзыва удаляет первопричину устаревшего доступа: при смене обязанностей или завершении работ прежние права операторов, разработчиков и администраторов MLOps перестают действовать в установленный срок.

Связь 644 · AISAFE2-WEAK-046 → AISAFE2-MIT-027

Фиксация архитектурных и договорных границ и владельцев подтверждений закрывает часть пробела управления соответствием; набор обязательных требований остаётся внешним входом.

Ограничение: Владельцы подтверждений назначены; применимые требования и их правовую трактовку нужно определить отдельно.

Связь 645 · AISAFE2-WEAK-046 → AISAFE2-MIT-028

Реестр утверждений связывает заявление о выполнении применимого контроля с точной версией, подтверждением, владельцем, сроком и принятым остаточным риском.

Ограничение: Реестр не определяет применимые требования и не доказывает соответствие.

Связь 646 · AISAFE2-WEAK-046 → AISAFE2-MIT-054

Детерминированный шлюз исполняет уже установленное решение на охваченных потоках данных, компенсируя часть последствий пробела; он не определяет применимые обязательства, владельцев и программу обеспечения соответствия.

Ограничение: Шлюз исполняет утверждённые правила для подключённых потоков данных; применимость требований и решение о допустимости обработки определяются отдельно.

Связь 647 · AISAFE2-WEAK-047 → AISAFE2-MIT-026

Инвентарь компонентов, владелец, срок и подтверждение рабочей версии создают отсутствующий процесс управления уязвимостями.

Связь 648 · AISAFE2-WEAK-047 → AISAFE2-MIT-071

Находки ручной проверки, SAST, DAST, фаззинга, поиска секретов и целевых abuse-case тестов связываются с точными commit и build, блокируют выпуск, требуют исправления и повторной проверки.

Ограничение: Проверяется собственный код до выпуска; сторонние зависимости, уязвимости работающей версии и установка исправлений требуют отдельного процесса.

Связь 649 · AISAFE2-WEAK-048 → AISAFE2-MIT-024

Исполняемые сценарии локализации, восстановления и регулярные учения создают отсутствующий процесс реагирования и восстановления.

Связь 650 · AISAFE2-WEAK-048 → AISAFE2-MIT-065

Защищённый архив сохраняет проверяемые материалы для расследования и ограничивает их незаметное изменение, удаление, смешение между делами и раскрытие постороннему читателю.

Ограничение: Мера создаёт защищённый набор материалов, но не задаёт сценарии обнаружения, локализации, отзыва, восстановления и безопасного возврата в работу.

Связь 651 · AISAFE2-WEAK-049 → AISAFE2-MIT-029

Авторизация на стороне исполнителя блокирует неразрешённое внешнее действие, вызванное поддельным или скомпрометированным межагентским сообщением.

Ограничение: Не аутентифицирует сообщение и не предотвращает его принятие или изменение внутреннего плана; охвачены только проверяемые внешние действия.

Связь 652 · AISAFE2-WEAK-049 → AISAFE2-MIT-048

Проверяемый отправитель, защищенная целостность, привязка к получателю и задаче, срок и атомарная защита от повтора устраняют неявное доверие к неподтвержденному сообщению на охваченном A2A-пути.

Связь 653 · AISAFE2-WEAK-050 → AISAFE2-MIT-049

Очистка либо переинициализация памяти, кэшей, буферов, временных файлов и повторно используемой среды выполнения до передачи ресурса другой границе доверия устраняет доступное остаточное состояние на охваченном пути.

Связь 654 · AISAFE2-WEAK-051 → AISAFE2-MIT-051

Неизменяемый ограниченный план, ровно одно конечное состояние для каждой обязательной стадии и блокирующий завершающий обработчик устраняют выпуск обычного или разрешающего результата при неполном анализе.

Связь 655 · AISAFE2-WEAK-052 → AISAFE2-MIT-006

Предварительное резервирование бюджета, учёт выполненной работы и жёсткие пределы вводят отсутствующий внутренний бюджет до каждой ограничиваемой операции.

Связь 656 · AISAFE2-WEAK-053 → AISAFE2-MIT-052

Версионируемый инвентарь и последовательность прекращения вызовов, отзыва маршрутов, идентичностей, учетных данных, прав и возможностей и удаления, передачи или удержания данных и копий устраняют неполный вывод в пределах контролируемого жизненного цикла.

Связь 657 · AISAFE2-WEAK-054 → AISAFE2-MIT-005

Фактическая связь с использованными источниками и явный статус подтверждения дают потребителю проверяемое основание обнаружить часть недостоверных утверждений.

Ограничение: Мера помогает проверить утверждения, для которых доступны подтверждённые источники, но не гарантирует истинность всего ответа и не устраняет стохастическую генерацию.

Связь 658 · AISAFE2-WEAK-054 → AISAFE2-MIT-045

Независимый проверяющий сопоставляет существенный вывод с фактами и правилами вне модели до решения с существенными последствиями.

Ограничение: Мера действует только для выбранных решений с существенными последствиями и сохраняет риск ошибки проверяющего и общей причины ошибки модели и человека.

Связь 659 · AISAFE2-WEAK-055 → AISAFE2-MIT-029

Независимая авторизация отделяет решение модели от полномочия на конкретный внешний эффект и блокирует неизвестное или запрещённое действие.

Ограничение: Мера покрывает авторизацию внешних эффектов, но не принуждает модель соблюдать все правила вычисления, последовательность шагов, лимиты и внутренние бизнес-инварианты.

Связь 660 · AISAFE2-WEAK-055 → AISAFE2-MIT-054

Технический шлюз применяет уже утверждённые ограничения к фактическому потоку данных независимо от интерпретации модели.

Ограничение: Мера покрывает только утверждённые ограничения обработки и передачи данных с наблюдаемыми атрибутами и не исполняет остальные бизнес-правила и порядок действий.

Связь 661 · AISAFE2-WEAK-055 → AISAFE2-MIT-066

Внешний серверный объект удерживает обязательную границу цели и возможностей вне решения модели и требует нового решения авторизации для её изменения.

Ограничение: Мера не даёт модели или недоверенному источнику расширить цель, но не исполняет все правила вычисления, порядок шагов, лимиты и иные бизнес-инварианты.

Связь 662 · AISAFE2-WEAK-056 → AISAFE2-MIT-001

Проверка входных данных по утверждённой схеме отклоняет структурно недопустимые аргументы инструмента, если она принудительно применяется непосредственно перед исполнителем; она не ограничивает набор инструментов, версии и операции и не заменяет специализированный реестр.

Ограничение: Связь действует лишь при обязательной проверке аргументов непосредственно перед исполнителем инструмента.

Связь 663 · AISAFE2-WEAK-056 → AISAFE2-MIT-016

Строгая схема, разрешённые операции и проверка типов, диапазонов и сочетаний блокируют сформированные моделью недопустимые аргументы до исполнителя.

Связь 664 · AISAFE2-WEAK-057 → AISAFE2-MIT-019

Блокирующий шлюз исполняет чувствительное действие только после явного решения уполномоченного человека по неизменённым получателю, параметрам и эффекту.

Связь 665 · AISAFE2-WEAK-058 → AISAFE2-MIT-029

Независимый механизм проверяет доступные идентичности пользователя и агента, объект и операцию непосредственно перед внешним эффектом.

Ограничение: Мера компенсирует потерю контекста для охваченных внешних действий, но не пробрасывает исходную идентичность в каждый инструмент и не обеспечивает объектную авторизацию всех чтений.

Связь 666 · AISAFE2-WEAK-058 → AISAFE2-MIT-030

Суженное делегированное полномочие связывает вызов коннектора с точными получателем, операцией, ресурсом, сроком и маршрутом.

Ограничение: Мера устраняет широкий сервисный контекст только для зарегистрированных коннекторов и поддерживаемой схемы делегирования; остальные инструменты и пользовательские права остаются вне её области.

Связь 667 · AISAFE2-WEAK-059 → AISAFE2-MIT-041

Раздельные входы, задания, идентичности и выходные артефакты сохраняют границу арендатора на охваченном пути дообучения.

Ограничение: Мера разделяет охваченные задания и артефакты, но не доказывает отсутствие межарендной информации в уже созданной общей обученной модели.

Связь 668 · AISAFE2-WEAK-059 → AISAFE2-MIT-042

Ограничение вклада записей и специальные проверки приватности уменьшают и измеряют возможность извлечь сведения другого арендатора через поведение модели.

Ограничение: Мера снижает вероятность и обнаруживает часть межарендных утечек, но не создаёт строгую границу арендатора в параметрах общей дообученной модели.

Связь 669 · AISAFE2-WEAK-060 → AISAFE2-MIT-002

Проверка и безопасное представление ответа исключают необязательные технические детали из ошибки, уже попавшей в AI-вывод или прикладной результат.

Ограничение: Мера покрывает охваченный путь результата, но прямые ошибки API, инструмента, RAG-контура или внешнего адаптера могут пересечь пользовательскую границу до этой проверки.

Связь 670 · AISAFE2-WEAK-061 → AISAFE2-MIT-054

Шлюз сопоставляет фактические поля телеметрии, получателя и назначение с утверждённым решением и блокирует запрещённую внешнюю передачу.

Ограничение: Мера покрывает только инвентаризированные и классифицированные пути телеметрии с наблюдаемыми атрибутами; скрытый публичный эндпойнт или неучтённая аналитика остаются вне её области.

Связь 671 · AISAFE2-WEAK-062 → AISAFE2-MIT-006

Сквозные счётчики сохраняют остаток бюджета между запросом, сессией, агентными шагами, инструментами, средой выполнения и внешним провайдером и блокируют превышение до следующей операции.

Связь 672 · AISAFE2-WEAK-063 → AISAFE2-MIT-006

Политика показывает доступную оценку или верхнюю границу, применяет пользовательский бюджет и предупреждает либо запрашивает подтверждение до установленного порога расходов.

Ограничение: Мера устраняет пробел только для расходов, которые продукт способен заранее оценить, наблюдать и остановить; независимые счета внешнего получателя или делегированного аккаунта могут остаться вне контроля.

Связь 673 · AISAFE2-WEAK-064 → AISAFE2-MIT-046

Ограниченное ожидание, прекращение повторов, идемпотентность и заранее проверенный режим не позволяют отказу AI-зависимости вызвать повторный или частичный опасный эффект.

Ограничение: Мера ограничивает последствия отказа охваченной AI-зависимости, но не предотвращает исходный отказ и не покрывает автоматически каждую внешнюю прикладную зависимость.

Связь 674 · AISAFE2-WEAK-065 → AISAFE2-MIT-037

Маршрутизация по классу риска и безопасное групповое подтверждение сокращают лишние обращения к человеку для охваченных однородных операций.

Ограничение: Мера определяет маршрутизацию и групповую обработку, но не задаёт доказуемые критерии эквивалентности, срок действия и аннулирование для безопасного автоматического переиспользования ранее принятого решения.

Связь 675 · AISAFE2-WEAK-066 → AISAFE2-MIT-037

Пределы нагрузки, замещение, маршруты эскалации и измерение очереди уменьшают влияние дефицита проверяющих на охваченный процесс.

Ограничение: Мера перераспределяет и ограничивает нагрузку, но не создаёт необходимую численность, компетентность, график или фактический человеческий резерв.

Связь 676 · AISAFE2-WEAK-067 → AISAFE2-MIT-037

Неизменяемый предпросмотр и обязательный контекст уменьшают трудоёмкость решения и позволяют оператору проверить охваченное действие.

Ограничение: Мера требует контекст текущего решения, но не задаёт интерфейс поиска и отображения истории аналогичных задач и ранее принятых решений.

Связь 677 · AISAFE2-WEAK-068 → AISAFE2-MIT-037

Метрики нагрузки и качества, ротация, разделение ролей и независимая выборочная проверка выявляют и ограничивают формальное одобрение и деградацию решений при утомлении.

Связь 678 · AISAFE2-WEAK-069 → AISAFE2-MIT-037

Заранее заданные блокирование, замещающий маршрут, приоритетная деградация и новое подтверждение после просрочки определяют поведение при переполнении и недоступности проверяющего.

Связь 679 · AISAFE2-WEAK-070 → AISAFE2-MIT-054

Для уже утверждённых ограничений данных технический шлюз блокирует запрещённую передачу в публичный или иной недоверенный AI-сервис независимо от поведения пользователя.

Ограничение: Мера компенсирует только охваченные потоки данных с отдельным утверждённым решением и не создаёт общеорганизационную политику, обучение сотрудников или санкционированные альтернативы.

Связь 680 · AISAFE2-WEAK-071 → AISAFE2-MIT-053

Версионируемая архитектура фиксирует доверенные зоны, субъекты, классы данных, получателей и разрешённые и запрещённые потоки до выпуска.

Связь 681 · AISAFE2-WEAK-072 → AISAFE2-MIT-054

Технический шлюз сопоставляет фактический переход данных с утверждённой зоной, назначением и получателем и блокирует неизвестный или запрещённый поток.

Связь 682 · AISAFE2-WEAK-072 → AISAFE2-MIT-057

Разрешающий список и запрет по умолчанию создают обязательный рубеж для наблюдаемых атрибутов сетевого пути.

Ограничение: Сетевой рубеж ограничивает путь, но не проверяет класс данных, цель передачи и права вызывающего субъекта.

Связь 683 · AISAFE2-WEAK-073 → AISAFE2-MIT-005

Связь с фактически использованными источниками и явный статус подтверждения дают потребителю часть независимо проверяемого материала до применения результата.

Ограничение: Показываются источники и статус утверждений; отчёт анализа кода и аннотации рискованных элементов надо передавать отдельно.

Связь 684 · AISAFE2-WEAK-074 → AISAFE2-MIT-010

Минимизация и маскирование до сборки контекста уменьшают объём чувствительного содержимого, которое модель способна передать через выводной канал.

Ограничение: Мера сокращает доступное модели чувствительное содержание, но не создаёт технического разграничения между оставшимся контекстом и выводом и не гарантирует нераскрытие.

Связь 685 · AISAFE2-WEAK-075 → AISAFE2-MIT-006

Сквозные счётчики и жёсткий совокупный бюджет останавливают учитываемое выделение ресурсов и расходов до превышения заданного владельцем предела.

Ограничение: Мера ограничивает только наблюдаемые и управляемые системой ресурсы и стоимость; независимое масштабирование или тарификация провайдера могут остаться вне охваченного контура.

Связь 686 · AISAFE2-WEAK-076 → AISAFE2-MIT-006

Жёсткие бюджеты и пределы параллелизма на арендатора ограничивают долю учтённого общего ресурса, которую один субъект способен занять.

Ограничение: Мера ограничивает учитываемое потребление арендатора, но сама не гарантирует изоляцию планировщика, очередей, сети и ускорителей на всех общих ресурсах.

Связь 687 · AISAFE2-WEAK-077 → AISAFE2-MIT-006

Ограничение приёма, времени и параллелизма удерживает обслуживаемый спрос в пределах доступного ресурса и снижает последствия нехватки ёмкости.

Ограничение: Мера ограничивает нагрузку после достижения заданных пределов, но не определяет требуемую ёмкость, запас на пик и своевременность масштабирования.

Связь 688 · AISAFE2-WEAK-078 → AISAFE2-MIT-006

Лимиты запросов, времени и параллелизма прекращают приём и выполнение части работы до неограниченного исчерпания общего ресурса.

Ограничение: Мера покрывает ограничиваемые операции, но не заменяет ограниченные очереди, сброс избыточной нагрузки и приоритеты на каждом участке перегруженного контура.

Связь 689 · AISAFE2-WEAK-079 → AISAFE2-MIT-046

Проверенный ограниченный режим сохраняет безопасную часть функции при недоступности единственного внешнего AI-сервиса или зависимости.

Ограничение: Мера ограничивает последствия отказа охваченной зависимости, но не создаёт переносимость, альтернативного поставщика и подготовленное переключение для восстановления полной функции.

Связь 690 · AISAFE2-WEAK-080 → AISAFE2-MIT-002

Контракт получателя и проверка структуры принудительно ограничивают формализуемую часть вывода до его отображения, сохранения или исполнения.

Ограничение: Проверяются формат, схема и допустимые значения; смысловое нарушение внутри формально верного ответа может пройти.

Связь 691 · AISAFE2-WEAK-080 → AISAFE2-MIT-056

Версионируемый детектор и безопасная реакция блокируют известные классы недопустимого содержания.

Ограничение: Неизвестная категория или перефразированное запрещённое содержание может пройти детектор.

Связь 692 · AISAFE2-WEAK-081 → AISAFE2-MIT-058

Проверка учётных данных до маршрутизации устанавливает субъекта и закрывает эквивалентный обход.

Связь 693 · AISAFE2-WEAK-082 → AISAFE2-MIT-059

Независимая усиленная проверка отклоняет привилегированного субъекта без требуемых факторов.

Связь 694 · AISAFE2-WEAK-083 → AISAFE2-MIT-057

Разрешающий список и запрет по умолчанию закрывают прямой путь из неутверждённого источника к служебному интерфейсу.

Связь 695 · AISAFE2-WEAK-084 → AISAFE2-MIT-026

Управление уязвимостями может выявить и устранить отдельную достигнутую небезопасную конфигурацию после её квалификации как находки.

Ограничение: Мера работает с отдельными выявленными находками и рабочими версиями, но не определяет полный эталон настроек и не проверяет каждое развёртывание или дрейф.

Связь 696 · AISAFE2-WEAK-084 → AISAFE2-MIT-053

Архитектура и модель угроз определяют требуемые свойства защиты и триггеры пересмотра для охваченных компонентов.

Ограничение: Мера создаёт входы для эталона и пересмотра, но не задаёт все эффективные настройки и не принуждает соответствие фактической конфигурации.

Связь 697 · AISAFE2-WEAK-084 → AISAFE2-MIT-067

Версионируемый эталон и проверка эффективной конфигурации при развёртывании и в эксплуатации устраняют отсутствие обязательных защищённых настроек и неизвестный дрейф.

Связь 698 · AISAFE2-WEAK-085 → AISAFE2-MIT-009

Версионирование и атомарный откат закрывают ветвь опубликованной базы знаний и её индекса.

Ограничение: Мера охватывает только базу знаний и не гарантирует независимую неизменяемую копию датасета, модели или критичной конфигурации вне рабочих полномочий.

Связь 699 · AISAFE2-WEAK-085 → AISAFE2-MIT-034

Контролируемое продвижение сохраняет идентичность и откат конкретной версии модели.

Ограничение: Мера покрывает модельную версию и её продвижение, но не создаёт независимые защищённые копии всех критичных AI-активов.

Связь 700 · AISAFE2-WEAK-085 → AISAFE2-MIT-050

Неизменяемые независимые копии и проверенное восстановление закрывают ветвь обучающих, тестовых и оценочных наборов данных.

Ограничение: Мера ограничена наборами данных и не покрывает модели, базу знаний, индексы и критичную конфигурацию.

Связь 701 · AISAFE2-WEAK-085 → AISAFE2-MIT-068

Независимые неизменяемые версии вне рабочих полномочий и проверенное восстановление точной версии устраняют возможность одной операцией уничтожить все восстановимые экземпляры критичного AI-актива.

Связь 702 · AISAFE2-WEAK-086 → AISAFE2-MIT-008

Сохранение актуальных меток доступа у окончательной единицы и её обязательная авторизация после последнего преобразования не допускают запрещённый фрагмент в контекст.

Связь 703 · AISAFE2-WEAK-087 → AISAFE2-MIT-011

Изоляция, авторизация и ограничение экспорта уменьшают доступ к векторам в охваченном хранилище.

Ограничение: Мера ограничивает пути доступа и экспорта, но не назначает векторам класс чувствительности источника и не охватывает внешнее вычисление и все копии.

Связь 704 · AISAFE2-WEAK-087 → AISAFE2-MIT-054

Технический шлюз блокирует обработку или передачу вектора, если для него уже существует наблюдаемое утверждённое ограничение.

Ограничение: Мера исполняет существующую классификацию на охваченных потоках, но при утрате наследования не создаёт отсутствующий класс и может не распознать вектор как чувствительный.

Связь 705 · AISAFE2-WEAK-087 → AISAFE2-MIT-062

Шифрование защищает содержимое управляемой сохраняемой копии вектора от прямого чтения без ключа.

Ограничение: Мера действует только на инвентаризированных копиях с уже выбранным режимом защиты и не исправляет отсутствующее наследование классификации.

Связь 706 · AISAFE2-WEAK-087 → AISAFE2-MIT-063

Защищённый канал ограничивает перехват вектора между охваченными конечными точками.

Ограничение: Мера защищает транспорт уже учтённого потока, но не определяет допустимого получателя и класс обработки производного представления.

Связь 707 · AISAFE2-WEAK-087 → AISAFE2-MIT-069

Производное векторное представление наследует не менее строгий класс чувствительности источника во всех управляемых копиях и потоках.

Связь 708 · AISAFE2-WEAK-088 → AISAFE2-MIT-025

Шлюз блокирует выпуск без утверждённых версионированных этических критериев и связанных с точной версией результатов их проверки.

Ограничение: Шлюз требует и проверяет наличие результата на пути выпуска, но сам не формирует критерии, не проводит оценку и не закрывает отдельный периодический эксплуатационный пересмотр.

Связь 709 · AISAFE2-WEAK-088 → AISAFE2-MIT-045

Независимый проверяющий применяет утверждённые этические критерии к решениям с существенными последствиями и блокирует либо эскалирует несоответствие.

Ограничение: Мера действует для выбранных индивидуальных решений и не заменяет системную оценку воздействия точной версии на затрагиваемые группы до выпуска и периодически.

Связь 710 · AISAFE2-WEAK-089 → AISAFE2-MIT-016

Реестр и диспетчер предоставляют только разрешённые для текущей задачи и роли инструменты и операции до выбора вызова.

Связь 711 · AISAFE2-WEAK-090 → AISAFE2-MIT-017

Изолированный диспетчер передаёт исполнителю только разрешённые входы, секреты, сеть и ресурсы и не допускает прямое исполнение в основном процессе.

Связь 712 · AISAFE2-WEAK-090 → AISAFE2-MIT-031

Отдельный браузерный профиль ограничивает данные, сеть и действия браузерного исполнителя или исполнителя, управляющего пользовательским интерфейсом.

Ограничение: Мера закрывает только браузерный вариант и управление пользовательским интерфейсом, но не изолирует остальные интерпретаторы, парсеры, серверы инструментов и сгенерированный код.

Связь 713 · AISAFE2-WEAK-091 → AISAFE2-MIT-012

Обязательный безопасный формат и проверка загрузчиком до открытия не позволяют артефакту модели исполнить непредусмотренный код.

Связь 714 · AISAFE2-WEAK-092 → AISAFE2-MIT-030

Посредник выдаёт или выбирает краткоживущее полномочие для точных получателя, аудитории, операции и ресурса, а коннектор проверяет эти ограничения до вызова.

Связь 715 · AISAFE2-WEAK-093 → AISAFE2-MIT-006

Сквозной счётчик, сохраняющий идентичность арендатора и операции, помогает отнести измеренное потребление к фактическому инициатору; начисление требуется сверять отдельным механизмом биллинга.

Ограничение: Мера покрывает измеримые ресурсы и начисления внутри контролируемого контура, но не исправляет внешние счета без доступной связи арендатора с операцией.

Связь 716 · AISAFE2-WEAK-094 → AISAFE2-MIT-064

Защищённые хранение и доставка, короткий срок, ротация и отзыв удерживают исходное делегированное учётное значение вне модели, журналов и нецелевых инструментов.

Ограничение: Мера закрывает хранение и жизненный цикл исходного учётного значения только там, где коннектор поддерживает изолированную доставку и не требует передачи значения модели.

Связь 717 · AISAFE2-WEAK-095 → AISAFE2-MIT-030

Посредник ограничивает платёжное полномочие получателем, операцией, ресурсом и сроком, сокращая область возможного злоупотребления.

Ограничение: Сумму, частоту и совокупный бюджет платёжный коннектор должен принудительно проверять перед каждым списанием; сама эта мера их не ограничивает.

Связь 718 · AISAFE2-IMP-001 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Выполнение действий, явно предложенных AI-системой» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: После выполнения действия, явно предложенного AI-системой, наступает вред защищаемому интересу человека или его финансовому положению.

Связь 719 · AISAFE2-IMP-001 → AISAFE2-COMP-014

Компонент «Контур человеческой проверки и подтверждения» участвует в реализации воздействия «Выполнение действий, явно предложенных AI-системой» на поверхности «Правила внедрения человеческой проверки при обработке запроса»: После выполнения действия, явно предложенного AI-системой, наступает вред защищаемому интересу человека или его финансовому положению.

Связь 720 · AISAFE2-IMP-001 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Выполнение действий, явно предложенных AI-системой» на поверхности «Сервисный и прикладной слой приложения»: После выполнения действия, явно предложенного AI-системой, наступает вред защищаемому интересу человека или его финансовому положению.

Связь 721 · AISAFE2-IMP-002 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Последствия принятия решений на основе результатов AI-системы» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: После принятия и исполнения решения, для которого результат AI-системы был существенным основанием, наступает вред человеку, организации или их финансовым интересам.

Связь 722 · AISAFE2-IMP-002 → AISAFE2-COMP-014

Компонент «Контур человеческой проверки и подтверждения» участвует в реализации воздействия «Последствия принятия решений на основе результатов AI-системы» на поверхности «Правила внедрения человеческой проверки при обработке запроса»: После принятия и исполнения решения, для которого результат AI-системы был существенным основанием, наступает вред человеку, организации или их финансовым интересам.

Связь 723 · AISAFE2-IMP-002 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Последствия принятия решений на основе результатов AI-системы» на поверхности «Сервисный и прикладной слой приложения»: После принятия и исполнения решения, для которого результат AI-системы был существенным основанием, наступает вред человеку, организации или их финансовым интересам.

Связь 724 · AISAFE2-IMP-003 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Вовлечение пользователей в злонамеренные схемы» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: AI-сервис используется, чтобы склонить пользователя к участию в обманной, мошеннической или иной вредоносной схеме.

Связь 725 · AISAFE2-IMP-004 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Передача потребителем чувствительных данных» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Пользователь раскрывает секреты, персональные или иные чувствительные сведения через диалог либо подключённый инструмент AI-системы.

Связь 726 · AISAFE2-IMP-004 → AISAFE2-COMP-003

Компонент «Управление контекстом общения с моделью» участвует в реализации воздействия «Передача потребителем чувствительных данных» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Пользователь раскрывает секреты, персональные или иные чувствительные сведения через диалог либо подключённый инструмент AI-системы.

Связь 727 · AISAFE2-IMP-004 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Передача потребителем чувствительных данных» на поверхности «Коннекторы к прикладным API и данным»: Пользователь раскрывает секреты, персональные или иные чувствительные сведения через диалог либо подключённый инструмент AI-системы.

Связь 728 · AISAFE2-IMP-004 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Передача потребителем чувствительных данных» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Пользователь раскрывает секреты, персональные или иные чувствительные сведения через диалог либо подключённый инструмент AI-системы.

Связь 729 · AISAFE2-IMP-004 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Передача потребителем чувствительных данных» на поверхности «Сервисный и прикладной слой приложения»: Пользователь раскрывает секреты, персональные или иные чувствительные сведения через диалог либо подключённый инструмент AI-системы.

Связь 730 · AISAFE2-IMP-006 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Неожидаемые (для разработчика AI-системы) ответы моделей» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Непредусмотренный результат модели достигает потребителя или следующего компонента и нарушает ожидаемую целостность либо надёжность обработки.

Связь 731 · AISAFE2-IMP-006 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Неожидаемые (для разработчика AI-системы) ответы моделей» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы»: Непредусмотренный результат модели достигает потребителя или следующего компонента и нарушает ожидаемую целостность либо надёжность обработки.

Связь 732 · AISAFE2-IMP-006 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Неожидаемые (для разработчика AI-системы) ответы моделей» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Непредусмотренный результат модели достигает потребителя или следующего компонента и нарушает ожидаемую целостность либо надёжность обработки.

Связь 733 · AISAFE2-IMP-006 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Неожидаемые (для разработчика AI-системы) ответы моделей» на поверхности «Сервисный и прикладной слой приложения»: Непредусмотренный результат модели достигает потребителя или следующего компонента и нарушает ожидаемую целостность либо надёжность обработки.

Связь 734 · AISAFE2-IMP-007 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Обход ограничений или предусмотренной логики работы Agent Core» на поверхности «Планирование и последовательность шагов»: После обхода ограничений или предусмотренной логики ядра агента исполняется неутверждённое действие, вызов или переход состояния.

Связь 735 · AISAFE2-IMP-007 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Обход ограничений или предусмотренной логики работы Agent Core» на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: После обхода ограничений или предусмотренной логики ядра агента исполняется неутверждённое действие, вызов или переход состояния.

Связь 736 · AISAFE2-IMP-007 → AISAFE2-COMP-012

Компонент «Идентификация, авторизация, ключи и секреты» участвует в реализации воздействия «Обход ограничений или предусмотренной логики работы Agent Core» на поверхности «Каталоги пользователей и сервисных идентичностей»: После обхода ограничений или предусмотренной логики ядра агента исполняется неутверждённое действие, вызов или переход состояния.

Связь 737 · AISAFE2-IMP-007 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Обход ограничений или предусмотренной логики работы Agent Core» на поверхности «Сервисный и прикладной слой приложения»: После обхода ограничений или предусмотренной логики ядра агента исполняется неутверждённое действие, вызов или переход состояния.

Связь 738 · AISAFE2-IMP-007 → AISAFE2-COMP-019

Компонент «Контур исполнения политик и защиты AI-запросов» участвует в реализации воздействия «Обход ограничений или предусмотренной логики работы Agent Core» на поверхности «AI-шлюзы и прокси принудительного контроля»: После обхода ограничений или предусмотренной логики ядра агента исполняется неутверждённое действие, вызов или переход состояния.

Связь 739 · AISAFE2-IMP-007 → AISAFE2-COMP-021

Компонент «Межагентское взаимодействие и реестр агентов» участвует в реализации воздействия «Обход ограничений или предусмотренной логики работы Agent Core» на поверхности «Реестры агентов и описание их возможностей»: После обхода ограничений или предусмотренной логики ядра агента исполняется неутверждённое действие, вызов или переход состояния.

Связь 740 · AISAFE2-IMP-008 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)» на поверхности «Общие инструменты и их среды выполнения»: Инструментальный слой выдаёт результат либо выполняет действие, которое не предусмотрено контрактом и изменяет информационное состояние или работу сервиса.

Связь 741 · AISAFE2-IMP-008 → AISAFE2-COMP-006

Запросный RAG возвращает непредусмотренный результат поиска или ранжирования и тем самым изменяет поведение инструментального слоя для текущего обращения.

Связь 742 · AISAFE2-IMP-008 → AISAFE2-COMP-007

Компонент «Долговременная память» участвует в реализации воздействия «Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)» на поверхности «Извлечение данных для последующих запросов»: Инструментальный слой выдаёт результат либо выполняет действие, которое не предусмотрено контрактом и изменяет информационное состояние или работу сервиса.

Связь 743 · AISAFE2-IMP-008 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)» на поверхности «Сервисный и прикладной слой приложения»: Инструментальный слой выдаёт результат либо выполняет действие, которое не предусмотрено контрактом и изменяет информационное состояние или работу сервиса.

Связь 744 · AISAFE2-IMP-008 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)» на поверхности «Операционные реляционные и документные базы данных»: Инструментальный слой выдаёт результат либо выполняет действие, которое не предусмотрено контрактом и изменяет информационное состояние или работу сервиса.

Связь 745 · AISAFE2-IMP-008 → AISAFE2-COMP-021

Компонент «Межагентское взаимодействие и реестр агентов» участвует в реализации воздействия «Неожидаемое поведение слоя инструментов (в том числе RAG и сторонних агентов)» на поверхности «Реестры агентов и описание их возможностей»: Инструментальный слой выдаёт результат либо выполняет действие, которое не предусмотрено контрактом и изменяет информационное состояние или работу сервиса.

Связь 746 · AISAFE2-IMP-009 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 747 · AISAFE2-IMP-009 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 748 · AISAFE2-IMP-009 → AISAFE2-COMP-006

Авторизованный запрос RAG возвращает чувствительный фрагмент из подготовленной базы знаний получателю, которому это содержимое не должно быть раскрыто.

Связь 749 · AISAFE2-IMP-009 → AISAFE2-COMP-007

Компонент «Долговременная память» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Извлечение данных для последующих запросов»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 750 · AISAFE2-IMP-009 → AISAFE2-COMP-013

Компонент «Телеметрия и обнаружение» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Сборщики журналов, трассировок и метрик»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 751 · AISAFE2-IMP-009 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 752 · AISAFE2-IMP-009 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Сервисный и прикладной слой приложения»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 753 · AISAFE2-IMP-009 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Раскрытие чувствительных данных, как-либо участвующих в формировании ответов от инструментов» на поверхности «Операционные реляционные и документные базы данных»: Чувствительные данные из инструмента, RAG, базы знаний, долговременной памяти, телеметрии, журналов или иного источника попадают в ответ либо к получателю без необходимого права.

Связь 754 · AISAFE2-IMP-010 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Раскрытие содержимого системного промпта и инструкций инструментов» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Пользователь или другой неавторизованный субъект получает системную инструкцию либо инструкции, задающие работу инструментов.

Связь 755 · AISAFE2-IMP-010 → AISAFE2-COMP-003

Компонент «Управление контекстом общения с моделью» участвует в реализации воздействия «Раскрытие содержимого системного промпта и инструкций инструментов» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Пользователь или другой неавторизованный субъект получает системную инструкцию либо инструкции, задающие работу инструментов.

Связь 756 · AISAFE2-IMP-010 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Раскрытие содержимого системного промпта и инструкций инструментов» на поверхности «MCP-клиенты и MCP-серверы»: Пользователь или другой неавторизованный субъект получает системную инструкцию либо инструкции, задающие работу инструментов.

Связь 757 · AISAFE2-IMP-010 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Раскрытие содержимого системного промпта и инструкций инструментов» на поверхности «Операционные реляционные и документные базы данных»: Пользователь или другой неавторизованный субъект получает системную инструкцию либо инструкции, задающие работу инструментов.

Связь 758 · AISAFE2-IMP-011 → AISAFE2-COMP-006

Содержимое подготовленной базы знаний или свойства индекса реконструируются через последовательность разрешённых поисковых запросов и наблюдение за возвращаемыми фрагментами.

Связь 759 · AISAFE2-IMP-011 → AISAFE2-COMP-012

Компонент «Идентификация, авторизация, ключи и секреты» участвует в реализации воздействия «Компрометация баз знаний» на поверхности «Каталоги пользователей и сервисных идентичностей»: Содержимое базы знаний, RAG-корпуса или векторного индекса становится доступно неуполномоченному субъекту через чтение, экспорт либо реконструкцию.

Связь 760 · AISAFE2-IMP-011 → AISAFE2-COMP-023

Содержимое подготовленной базы знаний или индекса становится доступно неуполномоченному субъекту через прямое чтение, экспорт либо компрометацию опубликованной копии.

Связь 761 · AISAFE2-IMP-012 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Кража модели» на поверхности «Собственные и предоставленные поставщиком модели»: Атакующий получает артефакт модели либо функционально воспроизводит её ценное поведение без разрешения владельца.

Связь 762 · AISAFE2-IMP-012 → AISAFE2-COMP-010

Компонент «Хранилище и реестр артефактов модели» участвует в реализации воздействия «Кража модели» на поверхности «Веса, конфигурация и формат модели»: Атакующий получает артефакт модели либо функционально воспроизводит её ценное поведение без разрешения владельца.

Связь 763 · AISAFE2-IMP-012 → AISAFE2-COMP-011

Компонент «Вычислительная и сетевая среда выполнения» участвует в реализации воздействия «Кража модели» на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Атакующий получает артефакт модели либо функционально воспроизводит её ценное поведение без разрешения владельца.

Связь 764 · AISAFE2-IMP-012 → AISAFE2-COMP-012

Компонент «Идентификация, авторизация, ключи и секреты» участвует в реализации воздействия «Кража модели» на поверхности «Каталоги пользователей и сервисных идентичностей»: Атакующий получает артефакт модели либо функционально воспроизводит её ценное поведение без разрешения владельца.

Связь 765 · AISAFE2-IMP-012 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Кража модели» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Атакующий получает артефакт модели либо функционально воспроизводит её ценное поведение без разрешения владельца.

Связь 766 · AISAFE2-IMP-012 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Кража модели» на поверхности «Организации, проекты и административные API провайдера»: Атакующий получает артефакт модели либо функционально воспроизводит её ценное поведение без разрешения владельца.

Связь 767 · AISAFE2-IMP-013 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели» на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы»: Конфиденциальные записи или атрибуты обучающих данных становятся доступны через модель либо связанные с обучением артефакты.

Связь 768 · AISAFE2-IMP-013 → AISAFE2-COMP-008

Компонент «Данные для подсистемы обучения» участвует в реализации воздействия «Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели» на поверхности «Обучающие и валидационные выборки»: Конфиденциальные записи или атрибуты обучающих данных становятся доступны через модель либо связанные с обучением артефакты.

Связь 769 · AISAFE2-IMP-013 → AISAFE2-COMP-009

Компонент «Конвейер обучения и дообучения» участвует в реализации воздействия «Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели» на поверхности «Подготовка данных и признаков для обучения»: Конфиденциальные записи или атрибуты обучающих данных становятся доступны через модель либо связанные с обучением артефакты.

Связь 770 · AISAFE2-IMP-013 → AISAFE2-COMP-010

Компонент «Хранилище и реестр артефактов модели» участвует в реализации воздействия «Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели» на поверхности «Веса, конфигурация и формат модели»: Конфиденциальные записи или атрибуты обучающих данных становятся доступны через модель либо связанные с обучением артефакты.

Связь 771 · AISAFE2-IMP-013 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Раскрытие чувствительных данных, попавших в датасет, использующийся для обучения или дообучения модели» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Конфиденциальные записи или атрибуты обучающих данных становятся доступны через модель либо связанные с обучением артефакты.

Связь 772 · AISAFE2-IMP-014 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 773 · AISAFE2-IMP-014 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «API и протокол обращения к модели»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 774 · AISAFE2-IMP-014 → AISAFE2-COMP-003

Компонент «Управление контекстом общения с моделью» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 775 · AISAFE2-IMP-014 → AISAFE2-COMP-006

Ответы поиска, метаданные или ошибки запросного RAG раскрывают сведения об источниках, технологии индексации или алгоритме ранжирования.

Связь 776 · AISAFE2-IMP-014 → AISAFE2-COMP-010

Компонент «Хранилище и реестр артефактов модели» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «Веса, конфигурация и формат модели»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 777 · AISAFE2-IMP-014 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «Сервисный и прикладной слой приложения»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 778 · AISAFE2-IMP-014 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «Операционные реляционные и документные базы данных»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 779 · AISAFE2-IMP-014 → AISAFE2-COMP-022

Компонент «Программные артефакты и цепочка поставки» участвует в реализации воздействия «Раскрытие сведений об используемых технологиях или алгоритмах работы системы, источниках данных» на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости»: Из поведения или конфигурации системы раскрываются её архитектура, алгоритмы, используемые технологии либо происхождение и свойства данных.

Связь 780 · AISAFE2-IMP-015 → AISAFE2-COMP-007

Компонент «Долговременная память» участвует в реализации воздействия «Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)» на поверхности «Сохранённые единицы информации»: Субъект читает или изменяет сохранённые данные AI-контура за пределами предоставленных ему полномочий.

Связь 781 · AISAFE2-IMP-015 → AISAFE2-COMP-008

Компонент «Данные для подсистемы обучения» участвует в реализации воздействия «Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)» на поверхности «Обучающие и валидационные выборки»: Субъект читает или изменяет сохранённые данные AI-контура за пределами предоставленных ему полномочий.

Связь 782 · AISAFE2-IMP-015 → AISAFE2-COMP-010

Компонент «Хранилище и реестр артефактов модели» участвует в реализации воздействия «Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)» на поверхности «Веса, конфигурация и формат модели»: Субъект читает или изменяет сохранённые данные AI-контура за пределами предоставленных ему полномочий.

Связь 783 · AISAFE2-IMP-015 → AISAFE2-COMP-012

Несанкционированное чтение или изменение сохранённого секрета, ключа либо учётных данных непосредственно затрагивает контур идентификации и секретов.

Связь 784 · AISAFE2-IMP-015 → AISAFE2-COMP-013

Компонент «Телеметрия и обнаружение» участвует в реализации воздействия «Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)» на поверхности «Сборщики журналов, трассировок и метрик»: Субъект читает или изменяет сохранённые данные AI-контура за пределами предоставленных ему полномочий.

Связь 785 · AISAFE2-IMP-015 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)» на поверхности «Операционные реляционные и документные базы данных»: Субъект читает или изменяет сохранённые данные AI-контура за пределами предоставленных ему полномочий.

Связь 786 · AISAFE2-IMP-015 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Несанкционированный доступ к чтению или изменению хранимых данных (в том числе за счёт повышения привилегий)» на поверхности «Организации, проекты и административные API провайдера»: Субъект читает или изменяет сохранённые данные AI-контура за пределами предоставленных ему полномочий.

Связь 787 · AISAFE2-IMP-015 → AISAFE2-COMP-022

Несанкционированное чтение или изменение исходного кода, пакета, образа либо записи реестра непосредственно затрагивает программные артефакты AI-контура.

Связь 788 · AISAFE2-IMP-015 → AISAFE2-COMP-023

Неуполномоченный субъект читает или изменяет подготовленный корпус, поисковый индекс либо опубликованную версию базы знаний.

Связь 789 · AISAFE2-IMP-015 → AISAFE2-COMP-025

Неуполномоченный субъект читает или изменяет исходный набор данных, API, файл либо другую поверхность данных инструмента.

Связь 790 · AISAFE2-IMP-016 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Планирование и последовательность шагов»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 791 · AISAFE2-IMP-016 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 792 · AISAFE2-IMP-016 → AISAFE2-COMP-011

Компонент «Вычислительная и сетевая среда выполнения» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 793 · AISAFE2-IMP-016 → AISAFE2-COMP-012

Несанкционированное использование IAM-политики, выдачи ключа, ротации либо отзыва секрета непосредственно является доступом к инфраструктурной функции.

Связь 794 · AISAFE2-IMP-016 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 795 · AISAFE2-IMP-016 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Сервисный и прикладной слой приложения»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 796 · AISAFE2-IMP-016 → AISAFE2-COMP-019

Компонент «Контур исполнения политик и защиты AI-запросов» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «AI-шлюзы и прокси принудительного контроля»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 797 · AISAFE2-IMP-016 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Организации, проекты и административные API провайдера»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 798 · AISAFE2-IMP-016 → AISAFE2-COMP-021

Компонент «Межагентское взаимодействие и реестр агентов» участвует в реализации воздействия «Несанкционированный доступ к функциональности инфраструктурных компонентов (в том числе за счёт повышения привилегий)» на поверхности «Реестры агентов и описание их возможностей»: Субъект вызывает функцию, инструмент или инфраструктурное действие, для которого у него нет делегированных полномочий.

Связь 799 · AISAFE2-IMP-017 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Неограниченное потребление токенов при обращении к языковым моделям» на поверхности «API и протокол обращения к модели»: Токенный расход запроса, сессии или агентной цепочки растёт без установленной верхней границы.

Связь 800 · AISAFE2-IMP-017 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Неограниченное потребление токенов при обращении к языковым моделям» на поверхности «Планирование и последовательность шагов»: Токенный расход запроса, сессии или агентной цепочки растёт без установленной верхней границы.

Связь 801 · AISAFE2-IMP-017 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Неограниченное потребление токенов при обращении к языковым моделям» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Токенный расход запроса, сессии или агентной цепочки растёт без установленной верхней границы.

Связь 802 · AISAFE2-IMP-017 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Неограниченное потребление токенов при обращении к языковым моделям» на поверхности «Сервисный и прикладной слой приложения»: Токенный расход запроса, сессии или агентной цепочки растёт без установленной верхней границы.

Связь 803 · AISAFE2-IMP-017 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Неограниченное потребление токенов при обращении к языковым моделям» на поверхности «Организации, проекты и административные API провайдера»: Токенный расход запроса, сессии или агентной цепочки растёт без установленной верхней границы.

Связь 804 · AISAFE2-IMP-018 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Незапланированные расходы на использование платных инструментов» на поверхности «Планирование и последовательность шагов»: Платные вызовы инструментов создают расходы, не заложенные в бюджет владельца соответствующей учётной записи.

Связь 805 · AISAFE2-IMP-018 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Незапланированные расходы на использование платных инструментов» на поверхности «Коннекторы к прикладным API и данным»: Платные вызовы инструментов создают расходы, не заложенные в бюджет владельца соответствующей учётной записи.

Связь 806 · AISAFE2-IMP-018 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Незапланированные расходы на использование платных инструментов» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Платные вызовы инструментов создают расходы, не заложенные в бюджет владельца соответствующей учётной записи.

Связь 807 · AISAFE2-IMP-018 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Незапланированные расходы на использование платных инструментов» на поверхности «Сервисный и прикладной слой приложения»: Платные вызовы инструментов создают расходы, не заложенные в бюджет владельца соответствующей учётной записи.

Связь 808 · AISAFE2-IMP-018 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Незапланированные расходы на использование платных инструментов» на поверхности «Организации, проекты и административные API провайдера»: Платные вызовы инструментов создают расходы, не заложенные в бюджет владельца соответствующей учётной записи.

Связь 809 · AISAFE2-IMP-019 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Чрезмерное потребление AI-системой собственных вычислительных ресурсов» на поверхности «Планирование и последовательность шагов»: Выполнение AI-цепочки занимает вычислительные ресурсы среды сверх допустимого операционного бюджета.

Связь 810 · AISAFE2-IMP-019 → AISAFE2-COMP-011

Компонент «Вычислительная и сетевая среда выполнения» участвует в реализации воздействия «Чрезмерное потребление AI-системой собственных вычислительных ресурсов» на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Выполнение AI-цепочки занимает вычислительные ресурсы среды сверх допустимого операционного бюджета.

Связь 811 · AISAFE2-IMP-019 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Чрезмерное потребление AI-системой собственных вычислительных ресурсов» на поверхности «Сервисный и прикладной слой приложения»: Выполнение AI-цепочки занимает вычислительные ресурсы среды сверх допустимого операционного бюджета.

Связь 812 · AISAFE2-IMP-019 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Чрезмерное потребление AI-системой собственных вычислительных ресурсов» на поверхности «Операционные реляционные и документные базы данных»: Выполнение AI-цепочки занимает вычислительные ресурсы среды сверх допустимого операционного бюджета.

Связь 813 · AISAFE2-IMP-020 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Неожиданно высокие расходы пользователей AI-системы» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Использование AI-сервиса приводит к начислениям для пользователя выше ожидаемой им стоимости.

Связь 814 · AISAFE2-IMP-020 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Неожиданно высокие расходы пользователей AI-системы» на поверхности «Сервисный и прикладной слой приложения»: Использование AI-сервиса приводит к начислениям для пользователя выше ожидаемой им стоимости.

Связь 815 · AISAFE2-IMP-020 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Неожиданно высокие расходы пользователей AI-системы» на поверхности «Организации, проекты и административные API провайдера»: Использование AI-сервиса приводит к начислениям для пользователя выше ожидаемой им стоимости.

Связь 816 · AISAFE2-IMP-021 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Зацикливание AI-системы при обработке запросов» на поверхности «Планирование и последовательность шагов»: Обработка запроса не завершается и многократно повторяет шаги, вызовы или самокоррекцию, удерживая ресурсы системы.

Связь 817 · AISAFE2-IMP-021 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Зацикливание AI-системы при обработке запросов» на поверхности «Общие инструменты и их среды выполнения»: Обработка запроса не завершается и многократно повторяет шаги, вызовы или самокоррекцию, удерживая ресурсы системы.

Связь 818 · AISAFE2-IMP-021 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Зацикливание AI-системы при обработке запросов» на поверхности «Сервисный и прикладной слой приложения»: Обработка запроса не завершается и многократно повторяет шаги, вызовы или самокоррекцию, удерживая ресурсы системы.

Связь 819 · AISAFE2-IMP-021 → AISAFE2-COMP-021

Компонент «Межагентское взаимодействие и реестр агентов» участвует в реализации воздействия «Зацикливание AI-системы при обработке запросов» на поверхности «Реестры агентов и описание их возможностей»: Обработка запроса не завершается и многократно повторяет шаги, вызовы или самокоррекцию, удерживая ресурсы системы.

Связь 820 · AISAFE2-IMP-022 → AISAFE2-COMP-001

Недоступный пользовательский интерфейс или API непосредственно лишает потребителя доступа к AI-системе либо результату в требуемый срок.

Связь 821 · AISAFE2-IMP-022 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «API и протокол обращения к модели»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 822 · AISAFE2-IMP-022 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Общие инструменты и их среды выполнения»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 823 · AISAFE2-IMP-022 → AISAFE2-COMP-006

Недоступность запросного извлечения или превышение его временного бюджета не позволяет AI-системе вернуть результат потребителю в установленный интервал.

Связь 824 · AISAFE2-IMP-022 → AISAFE2-COMP-007

Компонент «Долговременная память» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Извлечение данных для последующих запросов»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 825 · AISAFE2-IMP-022 → AISAFE2-COMP-011

Компонент «Вычислительная и сетевая среда выполнения» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 826 · AISAFE2-IMP-022 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 827 · AISAFE2-IMP-022 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Сервисный и прикладной слой приложения»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 828 · AISAFE2-IMP-022 → AISAFE2-COMP-018

Компонент «Операционные данные, очереди и кэши» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Операционные реляционные и документные базы данных»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 829 · AISAFE2-IMP-022 → AISAFE2-COMP-019

Компонент «Контур исполнения политик и защиты AI-запросов» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «AI-шлюзы и прокси принудительного контроля»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 830 · AISAFE2-IMP-022 → AISAFE2-COMP-020

Компонент «Контур управления внешним AI-провайдером» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Организации, проекты и административные API провайдера»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 831 · AISAFE2-IMP-022 → AISAFE2-COMP-021

Компонент «Межагентское взаимодействие и реестр агентов» участвует в реализации воздействия «Недоступность AI-системы для потребителя или отсутствие результата работы в течение директивного интервала» на поверхности «Реестры агентов и описание их возможностей»: Потребитель не получает доступ к AI-системе либо результат её работы до установленного для сервиса срока.

Связь 832 · AISAFE2-IMP-023 → AISAFE2-COMP-014

Компонент «Контур человеческой проверки и подтверждения» участвует в реализации воздействия «Повышение нагрузки на человеческие ресурсы (HITL overwhelming)» на поверхности «Очереди и интерфейсы подтверждения действий»: Поток заданий на ручную проверку превышает доступную пропускную способность компетентных проверяющих.

Связь 833 · AISAFE2-IMP-024 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Утечка чувствительных данных в недоверенный контур» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Чувствительные данные пересекают утверждённую границу доверия и становятся доступны внешней модели, инструменту или иному недоверенному компоненту.

Связь 834 · AISAFE2-IMP-024 → AISAFE2-COMP-003

Компонент «Управление контекстом общения с моделью» участвует в реализации воздействия «Утечка чувствительных данных в недоверенный контур» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Чувствительные данные пересекают утверждённую границу доверия и становятся доступны внешней модели, инструменту или иному недоверенному компоненту.

Связь 835 · AISAFE2-IMP-024 → AISAFE2-COMP-005

Компонент «Инструменты, MCP-серверы и коннекторы» участвует в реализации воздействия «Утечка чувствительных данных в недоверенный контур» на поверхности «Коннекторы к прикладным API и данным»: Чувствительные данные пересекают утверждённую границу доверия и становятся доступны внешней модели, инструменту или иному недоверенному компоненту.

Связь 836 · AISAFE2-IMP-024 → AISAFE2-COMP-016

Компонент «Адаптер внешнего AI-сервиса» участвует в реализации воздействия «Утечка чувствительных данных в недоверенный контур» на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Чувствительные данные пересекают утверждённую границу доверия и становятся доступны внешней модели, инструменту или иному недоверенному компоненту.

Связь 837 · AISAFE2-IMP-024 → AISAFE2-COMP-019

Компонент «Контур исполнения политик и защиты AI-запросов» участвует в реализации воздействия «Утечка чувствительных данных в недоверенный контур» на поверхности «AI-шлюзы и прокси принудительного контроля»: Чувствительные данные пересекают утверждённую границу доверия и становятся доступны внешней модели, инструменту или иному недоверенному компоненту.

Связь 838 · AISAFE2-IMP-024 → AISAFE2-COMP-021

При обмене через A2A-шлюз или брокер чувствительные данные из сообщения могут попасть агенту или получателю, которому их передавать нельзя.

Ограничение: Только если фактическое межагентское сообщение пересекает неутверждённую границу или поступает неуполномоченному адресату; штатная передача между доверенными агентами не является утечкой.

Связь 839 · AISAFE2-IMP-031 → AISAFE2-COMP-008

Компонент «Данные для подсистемы обучения» участвует в реализации воздействия «Утрата датасета в результате атаки» на поверхности «Обучающие и валидационные выборки»: Атака приводит к уничтожению или невосстановимому повреждению датасета, необходимого для обучения, тестирования либо валидации AI-системы.

Связь 840 · AISAFE2-IMP-032 → AISAFE2-COMP-001

Компонент «Пользовательский интерфейс и API приложения» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 841 · AISAFE2-IMP-032 → AISAFE2-COMP-002

Компонент «Модель и интерфейс инференса» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Подготовка и выдача результата инференса»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 842 · AISAFE2-IMP-032 → AISAFE2-COMP-003

Компонент «Управление контекстом общения с моделью» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 843 · AISAFE2-IMP-032 → AISAFE2-COMP-004

Компонент «Ядро агента и управление шагами (Agent Core)» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Планирование и последовательность шагов»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 844 · AISAFE2-IMP-032 → AISAFE2-COMP-005

Результат инструмента или коннектора, сопоставленный с неверным запросом, задачей либо сессией, непосредственно реализует смешение изолированных контекстов.

Связь 845 · AISAFE2-IMP-032 → AISAFE2-COMP-006

Запросное извлечение возвращает фрагмент, результат ранжирования или кэшированную выдачу из другого изолированного контекста.

Связь 846 · AISAFE2-IMP-032 → AISAFE2-COMP-007

Компонент «Долговременная память» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Сохранённые единицы информации»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 847 · AISAFE2-IMP-032 → AISAFE2-COMP-011

Компонент «Вычислительная и сетевая среда выполнения» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 848 · AISAFE2-IMP-032 → AISAFE2-COMP-017

Компонент «Серверная часть приложения и бизнес-логика» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Сервисный и прикладной слой приложения»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 849 · AISAFE2-IMP-032 → AISAFE2-COMP-018

Операционная запись, состояние сессии, сообщение очереди, задание или кэш одного изолированного контекста становится доступным другому контексту без разрешения.

Связь 850 · AISAFE2-IMP-032 → AISAFE2-COMP-021

Компонент «Межагентское взаимодействие и реестр агентов» участвует в реализации воздействия «Несанкционированное смешение артефактов между изолированными контекстами» на поверхности «Реестры агентов и описание их возможностей»: Артефакт запроса, ответа, сессии, задачи или промежуточного AI-состояния одного изолированного контекста становится доступен, подмешивается либо выдается другому контексту без разрешения.

Связь 851 · AISAFE2-IMP-032 → AISAFE2-COMP-024

Сгенерированный код, изображение, документ или сохраняемый текстовый результат одного изолированного контекста подмешивается либо выдаётся другому контексту без разрешения.

Связь 852 · AISAFE2-IMP-033 → AISAFE2-COMP-010

Атака уничтожает или необратимо повреждает артефакт модели, веса либо их восстановимые версии в реестре, поэтому требуемую точную версию нельзя вернуть.

Связь 853 · AISAFE2-IMP-033 → AISAFE2-COMP-015

Атака уничтожает единственную пригодную версию критичной конфигурации, системных инструкций или связанных выпускных артефактов в управляемом пути изменений, поэтому требуемое состояние нельзя воспроизвести.

Связь 854 · AISAFE2-IMP-033 → AISAFE2-COMP-023

Атака уничтожает подготовленную базу знаний, поисковый или векторный индекс и все пригодные версии, поэтому опубликованное состояние нельзя восстановить.

Связь 855 · AISAFE2-MIT-001 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» мера «Проверка входных данных по схемам и правилам» применяется на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Контроль проверяет вход по утверждённой схеме и отклоняет недопустимую структуру до передачи модели, инструменту или бизнес-логике.

Связь 856 · AISAFE2-MIT-001 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Проверка входных данных по схемам и правилам» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Контроль проверяет вход по утверждённой схеме и отклоняет недопустимую структуру до передачи модели, инструменту или бизнес-логике.

Связь 857 · AISAFE2-MIT-001 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Проверка входных данных по схемам и правилам» применяется на поверхности «Сервисный и прикладной слой приложения»: Контроль проверяет вход по утверждённой схеме и отклоняет недопустимую структуру до передачи модели, инструменту или бизнес-логике.

Связь 858 · AISAFE2-MIT-001 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Проверка входных данных по схемам и правилам» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Контроль проверяет вход по утверждённой схеме и отклоняет недопустимую структуру до передачи модели, инструменту или бизнес-логике.

Связь 859 · AISAFE2-MIT-001 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Проверка входных данных по схемам и правилам» применяется на поверхности «Реестры агентов и описание их возможностей»: Контроль проверяет вход по утверждённой схеме и отклоняет недопустимую структуру до передачи модели, инструменту или бизнес-логике.

Связь 860 · AISAFE2-MIT-002 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» мера «Проверка и безопасная обработка AI-вывода» применяется на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Связь 861 · AISAFE2-MIT-002 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Проверка и безопасная обработка AI-вывода» применяется на поверхности «Подготовка и выдача результата инференса»: Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Ограничение: Связь охватывает проверку структуры и формата вывода на границе интерфейса модели; безопасная обработка получателем и блокирование активных эффектов исполняются в других компонентах.

Связь 862 · AISAFE2-MIT-002 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» мера «Проверка и безопасная обработка AI-вывода» применяется на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Связь 863 · AISAFE2-MIT-002 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Проверка и безопасная обработка AI-вывода» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Связь 864 · AISAFE2-MIT-002 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Проверка и безопасная обработка AI-вывода» применяется на поверхности «Сервисный и прикладной слой приложения»: Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Связь 865 · AISAFE2-MIT-002 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Проверка и безопасная обработка AI-вывода» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Контроль проверяет структуру и допустимость AI-вывода для конкретного получателя до отображения, сохранения или исполнения.

Связь 866 · AISAFE2-MIT-002 → AISAFE2-COMP-024

Контроль проверяет структуру и допустимость сгенерированного кода, документа, изображения или текста для конкретного получателя до хранения, передачи, отображения или исполнения.

Связь 867 · AISAFE2-MIT-003 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» мера «Разделение доверенных инструкций и недоверенного контента при сборке контекста» применяется на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Сборщик контекста сохраняет происхождение и уровень доверия каждого фрагмента и не позволяет недоверенному содержимому стать системной инструкцией.

Связь 868 · AISAFE2-MIT-003 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Разделение доверенных инструкций и недоверенного контента при сборке контекста» применяется на поверхности «Планирование и последовательность шагов»: Сборщик контекста сохраняет происхождение и уровень доверия каждого фрагмента и не позволяет недоверенному содержимому стать системной инструкцией.

Связь 869 · AISAFE2-MIT-004 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» мера «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» применяется на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Версионируемый детектор выявляет признаки внедрения вредоносных инструкций и обхода ограничений и передаёт результат в заданную политику реакции.

Связь 870 · AISAFE2-MIT-004 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Версионируемый детектор выявляет признаки внедрения вредоносных инструкций и обхода ограничений и передаёт результат в заданную политику реакции.

Связь 871 · AISAFE2-MIT-005 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» мера «Отображение происхождения и статуса подтверждения AI-вывода» применяется на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Интерфейс связывает существенный вывод с фактически использованными источниками и явно показывает статус независимого подтверждения.

Связь 872 · AISAFE2-MIT-005 → AISAFE2-COMP-006

Результат запросного извлечения содержит проверяемое происхождение каждого существенного фрагмента и явно передаёт статус подтверждения следующему получателю.

Связь 873 · AISAFE2-MIT-006 → AISAFE2-COMP-002

На API инференса контроль учитывает вызовы и токены модели в бюджете, переданном приложением; шаги агента, инструменты и общая стоимость ограничиваются на своих исполнительных границах.

Ограничение: Общий сервисный аккаунт провайдера и лимит отдельного вызова не обеспечивают жёсткий бюджет фактического пользователя без его идентичности и сквозного счётчика приложения.

Связь 874 · AISAFE2-MIT-006 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Планирование и последовательность шагов»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 875 · AISAFE2-MIT-006 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 876 · AISAFE2-MIT-006 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 877 · AISAFE2-MIT-006 → AISAFE2-COMP-013

Телеметрия собирает измерения запросов, токенов, шагов и стоимости и сигнализирует о превышении; отклонение следующей операции исполняют шлюз API, планировщик или обёртка инструмента.

Ограничение: Контур телеметрии измеряет потребление ресурсов и формирует сигналы превышения, но не исполняет бюджеты и не останавливает работу в среде выполнения.

Связь 878 · AISAFE2-MIT-006 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 879 · AISAFE2-MIT-006 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Сервисный и прикладной слой приложения»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 880 · AISAFE2-MIT-006 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Операционные реляционные и документные базы данных»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 881 · AISAFE2-MIT-006 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 882 · AISAFE2-MIT-006 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Организации, проекты и административные API провайдера»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 883 · AISAFE2-MIT-006 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Ограничение потребления ресурсов и стоимости» применяется на поверхности «Реестры агентов и описание их возможностей»: Сквозные счётчики и жёсткие бюджеты ограничивают запросы, токены, шаги агента, вызовы инструментов, время, параллелизм и стоимость.

Связь 884 · AISAFE2-MIT-007 → AISAFE2-COMP-023

В подготовленную базу знаний допускаются только идентифицированные источники, прошедшие установленную проверку происхождения, целостности и качества.

Связь 885 · AISAFE2-MIT-008 → AISAFE2-COMP-006

Каждый запрос к подготовленной базе знаний авторизуется на минимальной значимой гранулярности до поиска и возврата фрагмента.

Связь 886 · AISAFE2-MIT-008 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» политики доступа и решения авторизации связывают пользователя, ресурс и операцию с разрешением на извлечение конкретного фрагмента.

Связь 887 · AISAFE2-MIT-008 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Авторизация извлекаемых данных на минимальной значимой гранулярности» применяется на поверхности «Операционные реляционные и документные базы данных»: Каждый запрос к хранилищу авторизуется на минимальной значимой гранулярности до извлечения документа, фрагмента или записи.

Связь 888 · AISAFE2-MIT-009 → AISAFE2-COMP-023

Версионирование связывает опубликованную базу знаний с точным составом источников и индекса и позволяет атомарно восстановить допустимую версию.

Связь 889 · AISAFE2-MIT-010 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «API и протокол обращения к модели»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 890 · AISAFE2-MIT-010 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 891 · AISAFE2-MIT-010 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 892 · AISAFE2-MIT-010 → AISAFE2-COMP-007

В компоненте «Долговременная память» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Правила записи, обновления, истечения и удаления»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 893 · AISAFE2-MIT-010 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Обучающие и валидационные выборки»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 894 · AISAFE2-MIT-010 → AISAFE2-COMP-009

Минимизация и фильтрация применяются непосредственно до передачи обучающих записей в подготовку, запуск и выполнение задачи обучения.

Связь 895 · AISAFE2-MIT-010 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 896 · AISAFE2-MIT-010 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 897 · AISAFE2-MIT-010 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 898 · AISAFE2-MIT-010 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Сервисный и прикладной слой приложения»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 899 · AISAFE2-MIT-010 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «Операционные реляционные и документные базы данных»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 900 · AISAFE2-MIT-010 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Чувствительные данные минимизируются, маскируются или исключаются до передачи модели, RAG, памяти, обучению, инструменту или журналу.

Связь 901 · AISAFE2-MIT-010 → AISAFE2-COMP-021

Перед отправкой сообщения через A2A-шлюз чувствительные поля и вложения минимизируются по цели и правам другого агента; запись чувствительных сведений в реестр проверяется отдельно.

Ограничение: Связь охватывает межагентские сообщения и записи реестра с чувствительными данными; обычный обмен без таких данных не требует этой проверки.

Связь 902 · AISAFE2-MIT-010 → AISAFE2-COMP-023

Чувствительные данные минимизируются, маскируются или исключаются во время подготовки корпуса до индексирования и публикации базы знаний.

Связь 903 · AISAFE2-MIT-010 → AISAFE2-COMP-025

Данные минимизируются и фильтруются на границе исходного источника до раскрытия инструменту, RAG или следующему AI-компоненту.

Связь 904 · AISAFE2-MIT-011 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» роли и решения доступа ограничивают чтение и экспорт векторов по субъекту, хранилищу и операции; изоляция самого индекса обеспечивается в хранилище.

Связь 905 · AISAFE2-MIT-011 → AISAFE2-COMP-023

Изоляция, авторизация и ограничения экспорта сужают пути доступа к подготовленному корпусу, индексу и их массовому извлечению.

Связь 906 · AISAFE2-MIT-012 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Проверка целостности и безопасная загрузка артефактов модели» применяется на поверхности «Веса, конфигурация и формат модели»: Загрузчик проверяет идентичность, происхождение, целостность и безопасный формат артефакта модели до его открытия или исполнения.

Связь 907 · AISAFE2-MIT-012 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Проверка целостности и безопасная загрузка артефактов модели» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Загрузчик проверяет идентичность, происхождение, целостность и безопасный формат артефакта модели до его открытия или исполнения.

Связь 908 · AISAFE2-MIT-012 → AISAFE2-COMP-012

Хранилище ключей и доверенных корней предоставляет материал для проверки подписи артефакта; загрузчик модели в другом компоненте проверяет происхождение, целостность и формат до открытия или исполнения.

Ограничение: Компонент ключей не выполняет загрузку артефакта и не подтверждает его безопасный формат самостоятельно.

Связь 909 · AISAFE2-MIT-013 → AISAFE2-COMP-005

Перед выпуском инструмента, MCP-сервера или коннектора шлюз проверяет его версию, зависимости, образ и происхождение; рабочий компонент исполняет только допущенный экземпляр.

Ограничение: Сам инструмент или коннектор не заменяет независимый шлюз выпуска и проверку состава в контуре разработки.

Связь 910 · AISAFE2-MIT-013 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Проверка компонентов цепочки поставки перед выпуском» применяется на поверхности «Подготовка данных и признаков для обучения»: Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки.

Связь 911 · AISAFE2-MIT-013 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Проверка компонентов цепочки поставки перед выпуском» применяется на поверхности «Веса, конфигурация и формат модели»: Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки.

Связь 912 · AISAFE2-MIT-013 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Проверка компонентов цепочки поставки перед выпуском» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки.

Связь 913 · AISAFE2-MIT-013 → AISAFE2-COMP-022

В компоненте «Программные артефакты и цепочка поставки» мера «Проверка компонентов цепочки поставки перед выпуском» применяется на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости»: Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки.

Связь 914 · AISAFE2-MIT-014 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Фиксация версии и базового профиля поведения модели» применяется на поверхности «Собственные и предоставленные поставщиком модели»: Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки.

Связь 915 · AISAFE2-MIT-014 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Фиксация версии и базового профиля поведения модели» применяется на поверхности «Веса, конфигурация и формат модели»: Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки.

Связь 916 · AISAFE2-MIT-014 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Фиксация версии и базового профиля поведения модели» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки.

Связь 917 · AISAFE2-MIT-014 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Фиксация версии и базового профиля поведения модели» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки.

Связь 918 · AISAFE2-MIT-014 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Фиксация версии и базового профиля поведения модели» применяется на поверхности «Организации, проекты и административные API провайдера»: Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки.

Связь 919 · AISAFE2-MIT-015 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» мера «Защита конфигурации системных промптов и инструкций инструментов» применяется на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Системные промпты, инструкции инструментов и правила сборки контекста хранятся как защищённая версионируемая конфигурация.

Связь 920 · AISAFE2-MIT-015 → AISAFE2-COMP-005

Версия инструкций и описаний инструментов, принадлежащих MCP-серверу или коннектору, защищается от несанкционированного изменения и проверяется перед использованием агентом.

Ограничение: Системные промпты и общие правила сборки контекста не следует считать конфигурацией инструмента без отдельной архитектурной привязки.

Связь 921 · AISAFE2-MIT-015 → AISAFE2-COMP-012

Роли и решения авторизации ограничивают изменение защищённой конфигурации промптов и инструкций; секреты из неё хранятся и выдаются через предназначенное для них хранилище.

Ограничение: Каталог идентичностей сам по себе не хранит и не версионирует системные промпты.

Связь 922 · AISAFE2-MIT-015 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Защита конфигурации системных промптов и инструкций инструментов» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Системные промпты, инструкции инструментов и правила сборки контекста хранятся как защищённая версионируемая конфигурация.

Связь 923 · AISAFE2-MIT-015 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Защита конфигурации системных промптов и инструкций инструментов» применяется на поверхности «Операционные реляционные и документные базы данных»: Системные промпты, инструкции инструментов и правила сборки контекста хранятся как защищённая версионируемая конфигурация.

Связь 924 · AISAFE2-MIT-015 → AISAFE2-COMP-022

Если промпты или инструкции инструментов включены в собственный версионируемый артефакт выпуска, его состав, подпись и происхождение проверяются как часть цепочки поставки.

Ограничение: Сторонние библиотеки и пакеты сами по себе не являются хранилищем системных промптов; для внешней или динамической конфигурации нужен отдельный контроль.

Связь 925 · AISAFE2-MIT-016 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Ограничение доступных инструментов и проверка их аргументов» применяется на поверхности «Планирование и последовательность шагов»: Реестр ограничивает доступный агенту набор инструментов и проверяет версию, операцию и аргументы каждого вызова до исполнения.

Связь 926 · AISAFE2-MIT-016 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Ограничение доступных инструментов и проверка их аргументов» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Реестр ограничивает доступный агенту набор инструментов и проверяет версию, операцию и аргументы каждого вызова до исполнения.

Связь 927 · AISAFE2-MIT-016 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Ограничение доступных инструментов и проверка их аргументов» применяется на поверхности «Сервисный и прикладной слой приложения»: Реестр ограничивает доступный агенту набор инструментов и проверяет версию, операцию и аргументы каждого вызова до исполнения.

Связь 928 · AISAFE2-MIT-016 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Ограничение доступных инструментов и проверка их аргументов» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Реестр ограничивает доступный агенту набор инструментов и проверяет версию, операцию и аргументы каждого вызова до исполнения.

Связь 929 · AISAFE2-MIT-017 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Изолированное исполнение инструментов» применяется на поверхности «Общие инструменты и их среды выполнения»: Недоверенный код, файлы и инструменты исполняются в отдельной среде с ограниченными ресурсами, сетью, секретами и состоянием.

Связь 930 · AISAFE2-MIT-017 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Изолированное исполнение инструментов» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Недоверенный код, файлы и инструменты исполняются в отдельной среде с ограниченными ресурсами, сетью, секретами и состоянием.

Связь 931 · AISAFE2-MIT-018 → AISAFE2-COMP-004

Ядро агента выполняет шаги от отдельной машинной идентичности с минимальными краткоживущими полномочиями, связанными с конкретными получателем, ресурсом и операцией.

Связь 932 · AISAFE2-MIT-018 → AISAFE2-COMP-005

Инструмент, MCP-сервер или коннектор использует отдельную машинную идентичность и принимает только минимальное действующее полномочие для конкретной операции и ресурса.

Связь 933 · AISAFE2-MIT-018 → AISAFE2-COMP-009

Задание обучения использует отдельную машинную идентичность и минимальные полномочия.

Связь 934 · AISAFE2-MIT-018 → AISAFE2-COMP-011

Среда выполнения доставляет полномочие через аутентифицированный канал идентичности нагрузки.

Связь 935 · AISAFE2-MIT-018 → AISAFE2-COMP-012

Контур идентификации выдаёт известной рабочей нагрузке краткоживущее полномочие, доставляет его через аутентифицированный канал и проверяет получателя, срок и отзыв.

Связь 936 · AISAFE2-MIT-018 → AISAFE2-COMP-016

Внешний AI-клиент использует отдельную машинную идентичность с ограниченной областью полномочий.

Связь 937 · AISAFE2-MIT-018 → AISAFE2-COMP-020

Управляющая плоскость поставщика проверяет машинную идентичность и минимальные эффективные права.

Связь 938 · AISAFE2-MIT-018 → AISAFE2-COMP-021

Межагентский контур сохраняет отдельные машинные идентичности участников и не расширяет полномочия при делегировании между агентами.

Связь 939 · AISAFE2-MIT-018 → AISAFE2-COMP-023

Загрузчик RAG использует отдельную идентичность для чтения источника и записи в индекс.

Связь 940 · AISAFE2-MIT-019 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Подтверждение человеком высокорисковых действий до исполнения» применяется на поверхности «Планирование и последовательность шагов»: Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению.

Связь 941 · AISAFE2-MIT-019 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Подтверждение человеком высокорисковых действий до исполнения» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению.

Связь 942 · AISAFE2-MIT-019 → AISAFE2-COMP-014

В компоненте «Контур человеческой проверки и подтверждения» мера «Подтверждение человеком высокорисковых действий до исполнения» применяется на поверхности «Правила внедрения человеческой проверки при обработке запроса»: Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению.

Связь 943 · AISAFE2-MIT-019 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Подтверждение человеком высокорисковых действий до исполнения» применяется на поверхности «Сервисный и прикладной слой приложения»: Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению.

Связь 944 · AISAFE2-MIT-019 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Подтверждение человеком высокорисковых действий до исполнения» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению.

Связь 945 · AISAFE2-MIT-020 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Аварийная остановка и возврат агента в безопасное состояние» применяется на поверхности «Выбор действия, повтор, остановка и обработка отказа»: Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Связь 946 · AISAFE2-MIT-020 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Аварийная остановка и возврат агента в безопасное состояние» применяется на поверхности «Общие инструменты и их среды выполнения»: Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Связь 947 · AISAFE2-MIT-020 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Аварийная остановка и возврат агента в безопасное состояние» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Связь 948 · AISAFE2-MIT-020 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Аварийная остановка и возврат агента в безопасное состояние» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Связь 949 · AISAFE2-MIT-020 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Аварийная остановка и возврат агента в безопасное состояние» применяется на поверхности «Сервисный и прикладной слой приложения»: Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Связь 950 · AISAFE2-MIT-020 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Аварийная остановка и возврат агента в безопасное состояние» применяется на поверхности «Реестры агентов и описание их возможностей»: Независимый канал остановки прекращает новые действия агента, отзывает полномочия и переводит процесс в заранее определённое безопасное состояние.

Связь 951 · AISAFE2-MIT-021 → AISAFE2-COMP-007

В компоненте «Долговременная память» мера «Управление доверием и сроком жизни долговременной памяти» применяется на поверхности «Правила записи, обновления, истечения и удаления»: Долговременная память принимает только данные с известным происхождением, областью доверия, владельцем и сроком жизни.

Связь 952 · AISAFE2-MIT-022 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Регистрация и сквозное связывание событий AI-цепочки» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Сквозная корреляция связывает значимые события AI-цепочки с субъектом, запросом, версиями компонентов и фактическим результатом.

Связь 953 · AISAFE2-MIT-023 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Выявление цепочек AI-атак и злоупотреблений» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции.

Связь 954 · AISAFE2-MIT-024 → AISAFE2-COMP-002

Сценарий локализации непосредственно переключает модель или прекращает её вызовы, а восстановление проверяет новую исполняемую версию.

Связь 955 · AISAFE2-MIT-024 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Планирование и последовательность шагов»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 956 · AISAFE2-MIT-024 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Общие инструменты и их среды выполнения»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 957 · AISAFE2-MIT-024 → AISAFE2-COMP-007

Сценарий непосредственно помещает затронутую долговременную память в карантин и проверяет её состояние перед возвратом.

Связь 958 · AISAFE2-MIT-024 → AISAFE2-COMP-008

Сценарий непосредственно помещает затронутый обучающий набор в карантин и сохраняет подтверждения его области.

Связь 959 · AISAFE2-MIT-024 → AISAFE2-COMP-010

Сценарий непосредственно прекращает доступ к похищаемому артефакту модели и выбирает проверенную версию для восстановления.

Связь 960 · AISAFE2-MIT-024 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 961 · AISAFE2-MIT-024 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 962 · AISAFE2-MIT-024 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Ограничение: Контур телеметрии предоставляет сигналы и подтверждения для реагирования, но локализация и восстановление исполняются в затронутых компонентах и операционном процессе.

Связь 963 · AISAFE2-MIT-024 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 964 · AISAFE2-MIT-024 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Сервисный и прикладной слой приложения»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 965 · AISAFE2-MIT-024 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Операционные реляционные и документные базы данных»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 966 · AISAFE2-MIT-024 → AISAFE2-COMP-019

Сценарий непосредственно включает безопасный режим и блокирует исходящую передачу на контуре исполнения политик.

Связь 967 · AISAFE2-MIT-024 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Организации, проекты и административные API провайдера»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 968 · AISAFE2-MIT-024 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Локализация и восстановление после AI-инцидентов» применяется на поверхности «Реестры агентов и описание их возможностей»: Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние.

Связь 969 · AISAFE2-MIT-024 → AISAFE2-COMP-023

Сценарий реагирования изолирует скомпрометированный источник, корпус или индекс и восстанавливает проверенную опубликованную базу знаний.

Связь 970 · AISAFE2-MIT-025 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Блокирующий шлюз выпуска AI-системы» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Блокирующий шлюз разрешает выпуск только при выполнении утверждённых критериев и наличии свежих подтверждений для точного состава версии.

Связь 971 · AISAFE2-MIT-026 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Общие инструменты и их среды выполнения»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 972 · AISAFE2-MIT-026 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Подготовка данных и признаков для обучения»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 973 · AISAFE2-MIT-026 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Веса, конфигурация и формат модели»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 974 · AISAFE2-MIT-026 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 975 · AISAFE2-MIT-026 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 976 · AISAFE2-MIT-026 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Сервисный и прикладной слой приложения»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 977 · AISAFE2-MIT-026 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 978 · AISAFE2-MIT-026 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Реестры агентов и описание их возможностей»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 979 · AISAFE2-MIT-026 → AISAFE2-COMP-022

В компоненте «Программные артефакты и цепочка поставки» мера «Управление уязвимостями AI-компонентов в эксплуатации» применяется на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости»: Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией.

Связь 980 · AISAFE2-MIT-027 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Закрепление ответственности и владельцев подтверждений» применяется на поверхности «Обучающие и валидационные выборки»: Роли и владельцы закрепляются за системой, данными, рисками, эксплуатацией и каждым обязательным подтверждением.

Ограничение: Связь охватывает назначение владельца и ответственности за подтверждения по наборам данных; полное распределение ответственности за систему и внешние стороны находится за границей набора.

Связь 981 · AISAFE2-MIT-027 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Закрепление ответственности и владельцев подтверждений» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Роли и владельцы закрепляются за системой, данными, рисками, эксплуатацией и каждым обязательным подтверждением.

Связь 982 · AISAFE2-MIT-028 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Управление подтверждениями и остаточным риском утверждений» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Реестр утверждений связывает каждое заявление о мере с областью, подтверждением, владельцем, сроком и принятым остаточным риском.

Связь 983 · AISAFE2-MIT-029 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Независимая авторизация действий агента перед внешним эффектом» применяется на поверхности «Планирование и последовательность шагов»: Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта.

Связь 984 · AISAFE2-MIT-029 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Независимая авторизация действий агента перед внешним эффектом» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта.

Связь 985 · AISAFE2-MIT-029 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Независимая авторизация действий агента перед внешним эффектом» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта.

Связь 986 · AISAFE2-MIT-029 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Независимая авторизация действий агента перед внешним эффектом» применяется на поверхности «Сервисный и прикладной слой приложения»: Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта.

Связь 987 · AISAFE2-MIT-029 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Независимая авторизация действий агента перед внешним эффектом» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта.

Связь 988 · AISAFE2-MIT-030 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Ограничение делегированных полномочий коннекторов» применяется на поверхности «Коннекторы к прикладным API и данным»: Делегированное полномочие коннектора ограничивается конкретными получателем, операцией, ресурсом, сроком и маршрутом.

Связь 989 · AISAFE2-MIT-030 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Ограничение делегированных полномочий коннекторов» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Делегированное полномочие коннектора ограничивается конкретными получателем, операцией, ресурсом, сроком и маршрутом.

Связь 990 · AISAFE2-MIT-030 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Ограничение делегированных полномочий коннекторов» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Делегированное полномочие коннектора ограничивается конкретными получателем, операцией, ресурсом, сроком и маршрутом.

Связь 991 · AISAFE2-MIT-031 → AISAFE2-COMP-005

Браузерный инструмент запускается в отдельном профиле, а инструмент управления компьютером — в отдельной среде. Оба получают только разрешённые данные и действия; значимый внешний эффект проходит отдельную авторизацию.

Связь 992 · AISAFE2-MIT-031 → AISAFE2-COMP-011

Среда выполнения изолирует браузерного агента и агента управления компьютером от основной системы, ограничивает ресурсы и сеть и очищает состояние после задания.

Связь 993 · AISAFE2-MIT-031 → AISAFE2-COMP-012

Для изолированных агентов выдаются отдельные учётные данные с полномочиями только на разрешённые ресурсы и операции; доступ отзывается после завершения сессии.

Связь 994 · AISAFE2-MIT-032 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Проверка исходящих данных перед передачей через коннекторы» применяется на поверхности «Авторизованный вызов, преобразование аргументов и проверка результата»: Исходящий шлюз проверяет назначение, получателя, класс и минимально необходимый объём данных до передачи через коннектор.

Связь 995 · AISAFE2-MIT-032 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Проверка исходящих данных перед передачей через коннекторы» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Исходящий шлюз проверяет назначение, получателя, класс и минимально необходимый объём данных до передачи через коннектор.

Связь 996 · AISAFE2-MIT-032 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Проверка исходящих данных перед передачей через коннекторы» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Исходящий шлюз проверяет назначение, получателя, класс и минимально необходимый объём данных до передачи через коннектор.

Связь 997 · AISAFE2-MIT-033 → AISAFE2-COMP-007

В компоненте «Долговременная память» мера «Шлюз выпуска изменений RAG, векторных хранилищ и памяти» применяется на поверхности «Правила записи, обновления, истечения и удаления»: Специализированный шлюз блокирует выпуск изменений RAG, векторного хранилища или памяти без проверки точного состава и результатов испытаний.

Связь 998 · AISAFE2-MIT-033 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Шлюз выпуска изменений RAG, векторных хранилищ и памяти» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Специализированный шлюз блокирует выпуск изменений RAG, векторного хранилища или памяти без проверки точного состава и результатов испытаний.

Связь 999 · AISAFE2-MIT-033 → AISAFE2-COMP-023

Специализированный шлюз блокирует публикацию изменений базы знаний или индекса без проверки точного состава и результатов испытаний.

Связь 1000 · AISAFE2-MIT-034 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Контролируемое продвижение моделей через реестр» применяется на поверхности «Собственные и предоставленные поставщиком модели»: Реестр продвигает только одобренную неизменяемую версию модели между стадиями и сохраняет решение о переходе.

Связь 1001 · AISAFE2-MIT-034 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Контролируемое продвижение моделей через реестр» применяется на поверхности «Веса, конфигурация и формат модели»: Реестр продвигает только одобренную неизменяемую версию модели между стадиями и сохраняет решение о переходе.

Связь 1002 · AISAFE2-MIT-034 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Контролируемое продвижение моделей через реестр» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Реестр продвигает только одобренную неизменяемую версию модели между стадиями и сохраняет решение о переходе.

Связь 1003 · AISAFE2-MIT-035 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Ведение проверяемого состава и происхождения компонентов AI-системы» применяется на поверхности «Общие инструменты и их среды выполнения»: Проверяемый состав связывает рабочую AI/ML-версию с происхождением моделей, кода, зависимостей, данных и конфигурации.

Связь 1004 · AISAFE2-MIT-035 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Ведение проверяемого состава и происхождения компонентов AI-системы» применяется на поверхности «Обучающие и валидационные выборки»: Проверяемый состав связывает рабочую AI/ML-версию с происхождением моделей, кода, зависимостей, данных и конфигурации.

Связь 1005 · AISAFE2-MIT-035 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Ведение проверяемого состава и происхождения компонентов AI-системы» применяется на поверхности «Подготовка данных и признаков для обучения»: Проверяемый состав связывает рабочую AI/ML-версию с происхождением моделей, кода, зависимостей, данных и конфигурации.

Связь 1006 · AISAFE2-MIT-035 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Ведение проверяемого состава и происхождения компонентов AI-системы» применяется на поверхности «Веса, конфигурация и формат модели»: Проверяемый состав связывает рабочую AI/ML-версию с происхождением моделей, кода, зависимостей, данных и конфигурации.

Связь 1007 · AISAFE2-MIT-035 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Ведение проверяемого состава и происхождения компонентов AI-системы» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Проверяемый состав связывает рабочую AI/ML-версию с происхождением моделей, кода, зависимостей, данных и конфигурации.

Связь 1008 · AISAFE2-MIT-035 → AISAFE2-COMP-022

В компоненте «Программные артефакты и цепочка поставки» мера «Ведение проверяемого состава и происхождения компонентов AI-системы» применяется на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости»: Проверяемый состав связывает рабочую AI/ML-версию с происхождением моделей, кода, зависимостей, данных и конфигурации.

Связь 1009 · AISAFE2-MIT-036 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Обнаружение и обработка дрейфа, смещения данных и деградации» применяется на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы»: Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации.

Связь 1010 · AISAFE2-MIT-036 → AISAFE2-COMP-006

Наблюдение сравнивает качество запросного поиска, ранжирования и возвращаемых фрагментов с базовым профилем и запускает реакцию при деградации.

Связь 1011 · AISAFE2-MIT-036 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Обнаружение и обработка дрейфа, смещения данных и деградации» применяется на поверхности «Обучающие и валидационные выборки»: Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации.

Связь 1012 · AISAFE2-MIT-036 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Обнаружение и обработка дрейфа, смещения данных и деградации» применяется на поверхности «Подготовка данных и признаков для обучения»: Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации.

Связь 1013 · AISAFE2-MIT-036 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Обнаружение и обработка дрейфа, смещения данных и деградации» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации.

Связь 1014 · AISAFE2-MIT-036 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Обнаружение и обработка дрейфа, смещения данных и деградации» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации.

Связь 1015 · AISAFE2-MIT-036 → AISAFE2-COMP-023

Наблюдение сравнивает состав подготовленного корпуса, индекс и свойства публикации с утверждённой базовой версией и запускает реакцию при дрейфе.

Связь 1016 · AISAFE2-MIT-037 → AISAFE2-COMP-014

В компоненте «Контур человеческой проверки и подтверждения» мера «Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений» применяется на поверхности «Рабочий процесс назначения, эскалации и замещения»: Шлюз распределяет запросы на подтверждение по классу риска, контролирует нагрузку проверяющих и качество решений и при перегрузке блокирует действие либо направляет его по заранее заданному маршруту эскалации.

Связь 1017 · AISAFE2-MIT-038 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1018 · AISAFE2-MIT-038 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «API и протокол обращения к модели»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1019 · AISAFE2-MIT-038 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1020 · AISAFE2-MIT-038 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1021 · AISAFE2-MIT-038 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «Сервисный и прикладной слой приложения»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1022 · AISAFE2-MIT-038 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1023 · AISAFE2-MIT-038 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Контроль злоупотребления AI-сервисом» применяется на поверхности «Организации, проекты и административные API провайдера»: Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов.

Связь 1024 · AISAFE2-MIT-039 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Проверка пригодности модели для целевого сценария» применяется на поверхности «Собственные и предоставленные поставщиком модели»: Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения.

Связь 1025 · AISAFE2-MIT-039 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Проверка пригодности модели для целевого сценария» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения.

Связь 1026 · AISAFE2-MIT-039 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Проверка пригодности модели для целевого сценария» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения.

Связь 1027 · AISAFE2-MIT-039 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Проверка пригодности модели для целевого сценария» применяется на поверхности «Организации, проекты и административные API провайдера»: Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения.

Связь 1028 · AISAFE2-MIT-040 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Допуск и целостность обучающих данных, разметки и признаков» применяется на поверхности «Обучающие и валидационные выборки»: До обучения проверяются происхождение, права, целостность, разметка и допустимость каждого набора данных и признака.

Связь 1029 · AISAFE2-MIT-040 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Допуск и целостность обучающих данных, разметки и признаков» применяется на поверхности «Подготовка данных и признаков для обучения»: До обучения проверяются происхождение, права, целостность, разметка и допустимость каждого набора данных и признака.

Связь 1030 · AISAFE2-MIT-041 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Защита конвейера обучения и дообучения» применяется на поверхности «Обучающие и валидационные выборки»: Конвейер обучения фиксирует входы, код, среду, полномочия и результат и отделяет обучение от выпуска рабочей модели.

Связь 1031 · AISAFE2-MIT-041 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Защита конвейера обучения и дообучения» применяется на поверхности «Подготовка данных и признаков для обучения»: Конвейер обучения фиксирует входы, код, среду, полномочия и результат и отделяет обучение от выпуска рабочей модели.

Связь 1032 · AISAFE2-MIT-041 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Защита конвейера обучения и дообучения» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Конвейер обучения фиксирует входы, код, среду, полномочия и результат и отделяет обучение от выпуска рабочей модели.

Связь 1033 · AISAFE2-MIT-041 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Защита конвейера обучения и дообучения» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Конвейер обучения фиксирует входы, код, среду, полномочия и результат и отделяет обучение от выпуска рабочей модели.

Связь 1034 · AISAFE2-MIT-042 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Контроль приватности обучающих данных и утечек через модель» применяется на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы»: Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Связь 1035 · AISAFE2-MIT-042 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Контроль приватности обучающих данных и утечек через модель» применяется на поверхности «Обучающие и валидационные выборки»: Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Связь 1036 · AISAFE2-MIT-042 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Контроль приватности обучающих данных и утечек через модель» применяется на поверхности «Подготовка данных и признаков для обучения»: Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Связь 1037 · AISAFE2-MIT-042 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Контроль приватности обучающих данных и утечек через модель» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Связь 1038 · AISAFE2-MIT-042 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Контроль приватности обучающих данных и утечек через модель» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Связь 1039 · AISAFE2-MIT-042 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Контроль приватности обучающих данных и утечек через модель» применяется на поверхности «Организации, проекты и административные API провайдера»: Ограничение вклада записей и специальные испытания снижают и измеряют утечку обучающих данных через поведение модели.

Связь 1040 · AISAFE2-MIT-043 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Защита артефактов модели от несанкционированного получения» применяется на поверхности «Веса, конфигурация и формат модели»: Артефакты модели защищаются авторизацией, шифрованием, учётом копий и ограничением экспорта на всём управляемом пути.

Связь 1041 · AISAFE2-MIT-043 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Защита артефактов модели от несанкционированного получения» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Артефакты модели защищаются авторизацией, шифрованием, учётом копий и ограничением экспорта на всём управляемом пути.

Связь 1042 · AISAFE2-MIT-043 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Защита артефактов модели от несанкционированного получения» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Артефакты модели защищаются авторизацией, шифрованием, учётом копий и ограничением экспорта на всём управляемом пути.

Связь 1043 · AISAFE2-MIT-044 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Ограничение извлечения модели через интерфейс» применяется на поверхности «API и протокол обращения к модели»: Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели.

Связь 1044 · AISAFE2-MIT-044 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Ограничение извлечения модели через интерфейс» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели.

Связь 1045 · AISAFE2-MIT-044 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Ограничение извлечения модели через интерфейс» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели.

Связь 1046 · AISAFE2-MIT-044 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Ограничение извлечения модели через интерфейс» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели.

Связь 1047 · AISAFE2-MIT-044 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Ограничение извлечения модели через интерфейс» применяется на поверхности «Организации, проекты и административные API провайдера»: Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели.

Связь 1048 · AISAFE2-MIT-045 → AISAFE2-COMP-014

В компоненте «Контур человеческой проверки и подтверждения» мера «Независимая проверка решений с существенными последствиями» применяется на поверхности «Правила внедрения человеческой проверки при обработке запроса»: Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями.

Связь 1049 · AISAFE2-MIT-045 → AISAFE2-COMP-015

В контуре разработки и выпуска проверяют, что для заданных классов решений есть независимые источники, блокирующий маршрут проверки и эскалация; контрольные сценарии подтверждают работу этого маршрута до выпуска.

Ограничение: Выпускная проверка маршрута не заменяет независимое подтверждение каждого реального решения с существенными последствиями.

Связь 1050 · AISAFE2-MIT-045 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Независимая проверка решений с существенными последствиями» применяется на поверхности «Сервисный и прикладной слой приложения»: Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями.

Связь 1051 · AISAFE2-MIT-046 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «API и протокол обращения к модели»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1052 · AISAFE2-MIT-046 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Общие инструменты и их среды выполнения»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1053 · AISAFE2-MIT-046 → AISAFE2-COMP-006

При недоступности запросного RAG система переходит на заранее проверенный ограниченный маршрут без повторного или частичного опасного эффекта.

Связь 1054 · AISAFE2-MIT-046 → AISAFE2-COMP-007

В компоненте «Долговременная память» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Извлечение данных для последующих запросов»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1055 · AISAFE2-MIT-046 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1056 · AISAFE2-MIT-046 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1057 · AISAFE2-MIT-046 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Сервисный и прикладной слой приложения»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1058 · AISAFE2-MIT-046 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Операционные реляционные и документные базы данных»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1059 · AISAFE2-MIT-046 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Безопасная деградация при недоступности AI-зависимости» применяется на поверхности «Реестры агентов и описание их возможностей»: При отказе AI-зависимости система переходит в заранее проверенный ограниченный режим без повторного или частичного опасного эффекта.

Связь 1060 · AISAFE2-MIT-047 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Устойчивость к состязательным входным данным» применяется на поверхности «LLM, мультимодальные, Mixture-of-Experts и иные ML-модели AI-системы»: Подготовка, ограничения входа и испытания на состязательных примерах повышают устойчивость классической ML- или мультимодальной модели.

Связь 1061 · AISAFE2-MIT-047 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Устойчивость к состязательным входным данным» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Подготовка, ограничения входа и испытания на состязательных примерах повышают устойчивость классической ML- или мультимодальной модели.

Связь 1062 · AISAFE2-MIT-048 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Аутентификация, целостность и защита от повтора межагентских сообщений» применяется на поверхности «Планирование и последовательность шагов»: Межагентское сообщение принимается только после проверки отправителя, получателя, целостности, связи с задачей, свежести и защиты от повтора.

Связь 1063 · AISAFE2-MIT-048 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Аутентификация, целостность и защита от повтора межагентских сообщений» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Межагентское сообщение принимается только после проверки отправителя, получателя, целостности, связи с задачей, свежести и защиты от повтора.

Связь 1064 · AISAFE2-MIT-048 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Аутентификация, целостность и защита от повтора межагентских сообщений» применяется на поверхности «Реестры агентов и описание их возможностей»: Межагентское сообщение принимается только после проверки отправителя, получателя, целостности, связи с задачей, свежести и защиты от повтора.

Связь 1065 · AISAFE2-MIT-049 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Очистка и изоляция ресурсов общей среды инференса» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Изоляция и очистка общей среды инференса предотвращают перенос состояния между арендаторами, пользователями, задачами и последовательными заданиями.

Связь 1066 · AISAFE2-MIT-050 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Неизменяемое версионирование и проверяемое восстановление наборов данных» применяется на поверхности «Обучающие и валидационные выборки»: Неизменяемые версии и независимые копии позволяют проверить и восстановить точный набор данных для обучения, тестирования или валидации.

Связь 1067 · AISAFE2-MIT-051 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Контроль полноты обязательного анализа и безопасный выпуск результата» применяется на поверхности «Сервисный и прикладной слой приложения»: Неизменяемый план обязательных проверок и блокирующий финализатор не выпускают обычный или разрешающий результат при неполном анализе.

Связь 1068 · AISAFE2-MIT-051 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Контроль полноты обязательного анализа и безопасный выпуск результата» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Неизменяемый план обязательных проверок и блокирующий финализатор не выпускают обычный или разрешающий результат при неполном анализе.

Связь 1069 · AISAFE2-MIT-052 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1070 · AISAFE2-MIT-052 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Собственные и предоставленные поставщиком модели»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1071 · AISAFE2-MIT-052 → AISAFE2-COMP-004

В компоненте «Ядро агента и управление шагами (Agent Core)» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Планирование и последовательность шагов»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1072 · AISAFE2-MIT-052 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Общие инструменты и их среды выполнения»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1073 · AISAFE2-MIT-052 → AISAFE2-COMP-007

В компоненте «Долговременная память» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Правила записи, обновления, истечения и удаления»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1074 · AISAFE2-MIT-052 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Обучающие и валидационные выборки»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1075 · AISAFE2-MIT-052 → AISAFE2-COMP-009

Вывод прекращает задачи, расписания, полномочия и рабочие объекты конвейера обучения.

Связь 1076 · AISAFE2-MIT-052 → AISAFE2-COMP-010

В компоненте «Хранилище и реестр артефактов модели» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Веса, конфигурация и формат модели»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1077 · AISAFE2-MIT-052 → AISAFE2-COMP-011

Вывод прекращает маршруты и освобождает относящиеся к компоненту ресурсы среды выполнения.

Связь 1078 · AISAFE2-MIT-052 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1079 · AISAFE2-MIT-052 → AISAFE2-COMP-013

Вывод прекращает сбор, маршрутизацию и доступ к относящейся к компоненту телеметрии по принятому сроку хранения.

Связь 1080 · AISAFE2-MIT-052 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1081 · AISAFE2-MIT-052 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1082 · AISAFE2-MIT-052 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Сервисный и прикладной слой приложения»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1083 · AISAFE2-MIT-052 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Операционные реляционные и документные базы данных»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1084 · AISAFE2-MIT-052 → AISAFE2-COMP-019

Вывод прекращает маршруты и исключения, удаляет конфигурацию компонента и блокирует повторное использование его правил доступа.

Связь 1085 · AISAFE2-MIT-052 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Организации, проекты и административные API провайдера»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1086 · AISAFE2-MIT-052 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Реестры агентов и описание их возможностей»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1087 · AISAFE2-MIT-052 → AISAFE2-COMP-022

В компоненте «Программные артефакты и цепочка поставки» мера «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» применяется на поверхности «Сторонние библиотеки, пакеты и зафиксированные зависимости»: Управляемый вывод прекращает маршруты и полномочия компонента и закрывает решения по его данным, копиям, интеграциям и внешним обязательствам.

Связь 1088 · AISAFE2-MIT-052 → AISAFE2-COMP-023

Управляемый вывод прекращает публикацию и доступ к корпусу, индексу и производным копиям и закрывает решения по их хранению.

Связь 1089 · AISAFE2-MIT-053 → AISAFE2-COMP-015

В компоненте «Контур разработки, тестирования и выпуска» мера «Прослеживаемая архитектура, модель угроз и пересмотр риска» применяется на поверхности «Репозитории изменений, рецензирование и CI/CD»: Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске.

Связь 1090 · AISAFE2-MIT-053 → AISAFE2-COMP-020

На поверхности управления внешним AI-провайдером мера регистрирует применимые каналы сигналов и превращает компрометацию, отзыв, изменение условий или недоступность в повторное решение по точным моделям, конечным точкам и выпускам.

Связь 1091 · AISAFE2-MIT-053 → AISAFE2-COMP-022

На поверхности программных артефактов и цепочки поставки мера связывает сигнал поставщика с точным компонентом, версией и затронутыми выпусками и запускает решение об ограничении, замене и повторных проверках.

Связь 1092 · AISAFE2-MIT-054 → AISAFE2-COMP-001

В компоненте «Пользовательский интерфейс и API приложения» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Веб-интерфейсы, мобильные интерфейсы и консоли»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1093 · AISAFE2-MIT-054 → AISAFE2-COMP-002

В компоненте «Модель и интерфейс инференса» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «API и протокол обращения к модели»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1094 · AISAFE2-MIT-054 → AISAFE2-COMP-003

В компоненте «Управление контекстом общения с моделью» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Системные, разработческие и пользовательские инструкции текущего запроса»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1095 · AISAFE2-MIT-054 → AISAFE2-COMP-004

Утверждённые ограничения обработки, хранения и передачи применяются непосредственно к состоянию задачи агента и промежуточным результатам шагов.

Связь 1096 · AISAFE2-MIT-054 → AISAFE2-COMP-005

В компоненте «Инструменты, MCP-серверы и коннекторы» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Коннекторы к прикладным API и данным»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1097 · AISAFE2-MIT-054 → AISAFE2-COMP-006

До поиска шлюз отсекает по метаданным фрагменты, обработка которых запрещена для данного получателя и цели. После извлечения, ранжирования и расширения он повторно проверяет фактически возвращаемые фрагменты до передачи в модель или другой контур.

Связь 1098 · AISAFE2-MIT-054 → AISAFE2-COMP-007

В компоненте «Долговременная память» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Правила записи, обновления, истечения и удаления»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1099 · AISAFE2-MIT-054 → AISAFE2-COMP-008

В компоненте «Данные для подсистемы обучения» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Обучающие и валидационные выборки»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1100 · AISAFE2-MIT-054 → AISAFE2-COMP-009

В компоненте «Конвейер обучения и дообучения» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Подготовка данных и признаков для обучения»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1101 · AISAFE2-MIT-054 → AISAFE2-COMP-011

В компоненте «Вычислительная и сетевая среда выполнения» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Виртуальные машины, контейнеры и оркестраторы»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1102 · AISAFE2-MIT-054 → AISAFE2-COMP-012

В компоненте «Идентификация, авторизация, ключи и секреты» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Каталоги пользователей и сервисных идентичностей»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1103 · AISAFE2-MIT-054 → AISAFE2-COMP-013

В компоненте «Телеметрия и обнаружение» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Сборщики журналов, трассировок и метрик»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1104 · AISAFE2-MIT-054 → AISAFE2-COMP-016

В компоненте «Адаптер внешнего AI-сервиса» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Клиенты, прокси и адаптеры внешних AI API»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1105 · AISAFE2-MIT-054 → AISAFE2-COMP-017

В компоненте «Серверная часть приложения и бизнес-логика» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Сервисный и прикладной слой приложения»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1106 · AISAFE2-MIT-054 → AISAFE2-COMP-018

В компоненте «Операционные данные, очереди и кэши» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Операционные реляционные и документные базы данных»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1107 · AISAFE2-MIT-054 → AISAFE2-COMP-019

В компоненте «Контур исполнения политик и защиты AI-запросов» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «AI-шлюзы и прокси принудительного контроля»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1108 · AISAFE2-MIT-054 → AISAFE2-COMP-020

В компоненте «Контур управления внешним AI-провайдером» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Организации, проекты и административные API провайдера»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1109 · AISAFE2-MIT-054 → AISAFE2-COMP-021

В компоненте «Межагентское взаимодействие и реестр агентов» мера «Принудительное соблюдение утвержденных ограничений обработки и передачи данных» применяется на поверхности «Реестры агентов и описание их возможностей»: Технический шлюз сопоставляет фактические атрибуты обработки и передачи с утверждённым решением и блокирует неизвестный или запрещённый поток.

Связь 1110 · AISAFE2-MIT-054 → AISAFE2-COMP-023

Технический шлюз сопоставляет допуск, преобразование и публикацию данных базы знаний с утверждёнными ограничениями обработки и блокирует запрещённый поток.

Связь 1111 · AISAFE2-MIT-055 → AISAFE2-COMP-012

Каталоги пользователей и сервисы авторизации хранят ролевые назначения, исполняют выдачу и отзыв прав и предоставляют фактический перечень человеческих и административных учётных записей для периодического пересмотра.

Связь 1112 · AISAFE2-MIT-055 → AISAFE2-COMP-015

В контуре разработки, тестирования и выпуска мера связывает права операторов, разработчиков и администраторов MLOps с текущими обязанностями и отзывает доступ к репозиториям, конвейерам и средам после смены роли или завершения работ.

Связь 1113 · AISAFE2-MIT-056 → AISAFE2-COMP-001

На внешнем входе и выходе мера проверяет содержимое до передачи следующему получателю.

Связь 1114 · AISAFE2-MIT-056 → AISAFE2-COMP-019

Контур исполнения политик применяет версионируемое решение допустимости до опасной границы.

Связь 1115 · AISAFE2-MIT-057 → AISAFE2-COMP-001

На внешней границе мера ограничивает публичные и внутренние входные маршруты утверждённой топологией.

Связь 1116 · AISAFE2-MIT-057 → AISAFE2-COMP-005

На поверхности инструментов и MCP мера разрешает только утверждённые сетевые маршруты.

Связь 1117 · AISAFE2-MIT-057 → AISAFE2-COMP-011

В среде выполнения мера ограничивает входящие, исходящие и межкомпонентные соединения.

Связь 1118 · AISAFE2-MIT-057 → AISAFE2-COMP-016

Для адаптера внешнего AI API мера принуждает утверждённую конечную точку и маршрут.

Связь 1119 · AISAFE2-MIT-058 → AISAFE2-COMP-001

На внешней границе мера проверяет учётные данные до начала AI-обработки.

Связь 1120 · AISAFE2-MIT-058 → AISAFE2-COMP-012

Контур идентификации устанавливает проверенную идентичность внешнего клиента до доступа.

Связь 1121 · AISAFE2-MIT-058 → AISAFE2-COMP-017

Внутренний сервис принимает только проверенный контекст и закрывает прямой обход шлюза.

Связь 1122 · AISAFE2-MIT-059 → AISAFE2-COMP-012

Контур идентификации принудительно применяет усиленную аутентификацию привилегированных людей.

Связь 1123 · AISAFE2-MIT-059 → AISAFE2-COMP-020

Административная плоскость внешнего AI-провайдера требует усиленную аутентификацию привилегированного человека.

Связь 1124 · AISAFE2-MIT-060 → AISAFE2-COMP-008

Хранилище обучающих данных принимает новую версию только через разрешённый переход стадии.

Связь 1125 · AISAFE2-MIT-060 → AISAFE2-COMP-009

Задание обучения читает только опубликованную и прослеживаемую версию данных.

Связь 1126 · AISAFE2-MIT-060 → AISAFE2-COMP-023

Формирование корпуса создаёт новую версию через разрешённый переход без обратной записи.

Связь 1127 · AISAFE2-MIT-061 → AISAFE2-COMP-007

Мера ограничивает публичный и неутверждённый доступ к памяти AI-системы.

Связь 1128 · AISAFE2-MIT-061 → AISAFE2-COMP-008

Мера ограничивает публичный и неутверждённый доступ к обучающим данным.

Связь 1129 · AISAFE2-MIT-061 → AISAFE2-COMP-010

Мера ограничивает публичный и неутверждённый доступ к реестру моделей.

Связь 1130 · AISAFE2-MIT-061 → AISAFE2-COMP-013

Мера ограничивает публичный и неутверждённый доступ к телеметрии.

Связь 1131 · AISAFE2-MIT-061 → AISAFE2-COMP-018

Мера ограничивает публичный и неутверждённый доступ к рабочим данным.

Связь 1132 · AISAFE2-MIT-061 → AISAFE2-COMP-023

Мера ограничивает публичный и неутверждённый доступ к формируемому корпусу.

Связь 1133 · AISAFE2-MIT-061 → AISAFE2-COMP-024

Мера ограничивает публичный и неутверждённый доступ к повторно используемым AI-артефактам.

Связь 1134 · AISAFE2-MIT-062 → AISAFE2-COMP-007

Мера подтверждает шифрование сохраняемых копий памяти AI-системы.

Связь 1135 · AISAFE2-MIT-062 → AISAFE2-COMP-008

Мера подтверждает шифрование обучающих данных и их резервных копий.

Связь 1136 · AISAFE2-MIT-062 → AISAFE2-COMP-010

Мера подтверждает шифрование моделей и метаданных реестра моделей.

Связь 1137 · AISAFE2-MIT-062 → AISAFE2-COMP-011

Мера подтверждает шифрование дисков и временных копий среды выполнения.

Связь 1138 · AISAFE2-MIT-062 → AISAFE2-COMP-013

Мера подтверждает шифрование телеметрии и её резервных копий.

Связь 1139 · AISAFE2-MIT-062 → AISAFE2-COMP-018

Мера подтверждает шифрование рабочих данных и экспортных копий.

Связь 1140 · AISAFE2-MIT-062 → AISAFE2-COMP-023

Мера подтверждает шифрование корпуса и производных наборов данных.

Связь 1141 · AISAFE2-MIT-062 → AISAFE2-COMP-024

Мера подтверждает шифрование повторно используемых AI-артефактов.

Связь 1142 · AISAFE2-MIT-063 → AISAFE2-COMP-001

Мера защищает транспорт внешнего входа и выхода с проверкой удалённой стороны.

Связь 1143 · AISAFE2-MIT-063 → AISAFE2-COMP-005

Мера защищает каналы инструментов и MCP с проверкой удалённой стороны.

Связь 1144 · AISAFE2-MIT-063 → AISAFE2-COMP-006

Мера защищает передачу данных к хранилищу знаний и обратно.

Связь 1145 · AISAFE2-MIT-063 → AISAFE2-COMP-011

Мера защищает внешние и межкомпонентные каналы среды выполнения.

Связь 1146 · AISAFE2-MIT-063 → AISAFE2-COMP-016

Мера защищает транспорт адаптера внешнего AI API и проверяет получателя.

Связь 1147 · AISAFE2-MIT-063 → AISAFE2-COMP-018

Мера защищает транспорт рабочих данных между приложением и хранилищем.

Связь 1148 · AISAFE2-MIT-063 → AISAFE2-COMP-021

Мера защищает межагентный канал и проверяет идентичность удалённого агента.

Связь 1149 · AISAFE2-MIT-063 → AISAFE2-COMP-023

Мера защищает передачу корпуса и данных RAG между компонентами.

Связь 1150 · AISAFE2-MIT-064 → AISAFE2-COMP-005

Мера управляет секретами инструментов и MCP без помещения значений в конфигурацию.

Связь 1151 · AISAFE2-MIT-064 → AISAFE2-COMP-009

Мера доставляет секрет заданию обучения только через утверждённую машинную идентичность.

Связь 1152 · AISAFE2-MIT-064 → AISAFE2-COMP-011

Среда выполнения получает секрет из утверждённого хранилища без постоянного открытого значения.

Связь 1153 · AISAFE2-MIT-064 → AISAFE2-COMP-012

Контур идентификации ограничивает потребителей, области и жизненный цикл связанных секретов.

Связь 1154 · AISAFE2-MIT-064 → AISAFE2-COMP-016

Мера управляет ключами и токенами адаптера внешнего AI API.

Связь 1155 · AISAFE2-MIT-064 → AISAFE2-COMP-020

Мера управляет секретами управляющей плоскости внешнего AI-провайдера.

Связь 1156 · AISAFE2-MIT-065 → AISAFE2-COMP-013

Контур телеметрии формирует исходные события, идентификаторы и временные метки, которые мера связывает с защищённой версией материала расследования.

Связь 1157 · AISAFE2-MIT-065 → AISAFE2-COMP-018

Операционное хранилище реализует отдельное размещение, версии, сроки, изоляцию дел и проверяемое удаление материалов расследования.

Связь 1158 · AISAFE2-MIT-066 → AISAFE2-COMP-003

Сборщик контекста сохраняет различие между доверенным объектом цели и недоверенными фактами, предложениями и инструкциями из остальных источников.

Связь 1159 · AISAFE2-MIT-066 → AISAFE2-COMP-004

Ядро агента хранит точную версию цели и связывает с ней каждый шаг, не изменяя её через обычное продолжение диалога.

Связь 1160 · AISAFE2-MIT-066 → AISAFE2-COMP-017

Серверная бизнес-логика аутентифицирует инициатора, создаёт новую версию цели и проводит повторную авторизацию допустимого изменения.

Связь 1161 · AISAFE2-MIT-066 → AISAFE2-COMP-019

Контур исполнения политик отклоняет шаг с неизвестной, устаревшей или неавторизованно изменённой версией цели до внешнего действия.

Связь 1162 · AISAFE2-MIT-067 → AISAFE2-COMP-002

Эталон задаёт и проверяет эффективные настройки сервера инференса для точных версии и среды.

Связь 1163 · AISAFE2-MIT-067 → AISAFE2-COMP-004

Эталон ограничивает служебные режимы, привилегии процесса и конфигурацию агентной среды.

Связь 1164 · AISAFE2-MIT-067 → AISAFE2-COMP-005

Эталон и контроль дрейфа блокируют неизвестную или небезопасную конфигурацию инструментов и коннекторов.

Связь 1165 · AISAFE2-MIT-067 → AISAFE2-COMP-010

Эталон применяется к административным, хранилищным и служебным настройкам реестра моделей.

Связь 1166 · AISAFE2-MIT-067 → AISAFE2-COMP-011

Среда выполнения предоставляет эффективные значения для автоматической сверки и изолирует неустранённое критичное отклонение.

Связь 1167 · AISAFE2-MIT-067 → AISAFE2-COMP-013

Эталон ограничивает служебные интерфейсы, экспорт и режимы телеметрии.

Связь 1168 · AISAFE2-MIT-067 → AISAFE2-COMP-015

Проверка точной конфигурации по применимому эталону создаёт блокирующее подтверждение выпуска.

Связь 1169 · AISAFE2-MIT-067 → AISAFE2-COMP-023

Эталон применяется к подсистеме подготовки базы знаний, индексам и их административным настройкам.

Связь 1170 · AISAFE2-MIT-068 → AISAFE2-COMP-008

Мера хранит независимые неизменяемые версии обучающих и оценочных наборов и проверяет восстановление выбранной версии.

Связь 1171 · AISAFE2-MIT-068 → AISAFE2-COMP-010

Мера защищает точные версии весов, формата и метаданных модели вне полномочий рабочего реестра.

Связь 1172 · AISAFE2-MIT-068 → AISAFE2-COMP-015

Мера сохраняет защищённые версии критичной конфигурации и системных инструкций из управляемого пути изменений.

Связь 1173 · AISAFE2-MIT-068 → AISAFE2-COMP-023

Мера защищает версии корпуса и индекса базы знаний и проверяет атомарное восстановление требуемой версии.

Связь 1174 · AISAFE2-MIT-069 → AISAFE2-COMP-006

Извлечение обрабатывает вектор и производный результат не слабее класса чувствительности источника.

Связь 1175 · AISAFE2-MIT-069 → AISAFE2-COMP-016

Внешнее вычисление представлений получает только данные и режим обработки, разрешённые наследованным классом чувствительности.

Связь 1176 · AISAFE2-MIT-069 → AISAFE2-COMP-023

Конвейер формирования базы знаний переносит класс источника на каждый вектор, индекс, экспорт и резервную копию.

Связь 1177 · AISAFE2-MIT-069 → AISAFE2-COMP-025

Класс чувствительности и ссылка на источник образуют обязательный вход для классификации производного представления.

Связь 1178 · AISAFE2-MIT-070 → AISAFE2-COMP-002

Активная оценка воспроизводит атаки на поведение модели и интерфейс инференса для точных версий модели, промпта, конфигурации и выпуска.

Связь 1179 · AISAFE2-MIT-070 → AISAFE2-COMP-004

Сценарии red teaming проверяют планирование, переходы, циклы, остановку и выбор действий Agent Core, а ручная оценка разбирает найденные обходы.

Связь 1180 · AISAFE2-MIT-070 → AISAFE2-COMP-005

Сценарии проверяют границы доверия, аргументы, полномочия и результаты инструментов, MCP-серверов и коннекторов.

Связь 1181 · AISAFE2-MIT-070 → AISAFE2-COMP-006

Сценарии проверяют инъекции, отравление, происхождение, авторизацию и выдачу данных на запросном RAG-пути.

Связь 1182 · AISAFE2-MIT-070 → AISAFE2-COMP-011

Активные сценарии проверяют сетевую достижимость, изоляцию, ресурсы и фактические настройки среды выполнения точной тестируемой версии.

Связь 1183 · AISAFE2-MIT-070 → AISAFE2-COMP-012

Оценка проверяет обход идентификации и авторизации, избыточные полномочия, доступ к ключам и секретам и их использование в атакующих сценариях.

Связь 1184 · AISAFE2-MIT-070 → AISAFE2-COMP-013

Контрольные атаки проверяют, что телеметрия формируется и коррелируется, правила обнаружения срабатывают, а сигнал проходит по маршруту эскалации.

Связь 1185 · AISAFE2-MIT-070 → AISAFE2-COMP-015

Контур разработки, тестирования и выпуска фиксирует область активной оценки, точные версии, находки, решения, блокировку и результаты повторной проверки.

Связь 1186 · AISAFE2-MIT-070 → AISAFE2-COMP-017

Сценарии и ручная оценка проверяют серверную координацию, бизнес-правила и фактические эффекты AI-функции на точной версии приложения.

Связь 1187 · AISAFE2-MIT-070 → AISAFE2-COMP-019

Активные сценарии пытаются обойти или использовать не по назначению утверждённые правила защиты на рабочем пути AI-запроса.

Связь 1188 · AISAFE2-MIT-070 → AISAFE2-COMP-021

Сценарии проверяют доверие, адресацию, целостность, свежесть, повтор и границы делегирования в межагентском взаимодействии.

Связь 1189 · AISAFE2-MIT-071 → AISAFE2-COMP-001

Ручная и выбранная автоматизированная проверка охватывает собственные обработчики UI и API, валидацию входов и безопасную выдачу результата до выпуска.

Связь 1190 · AISAFE2-MIT-071 → AISAFE2-COMP-004

Независимая проверка и abuse-case тесты охватывают собственный код Agent Core, его переходы, остановку, повтор и выбор действий.

Связь 1191 · AISAFE2-MIT-071 → AISAFE2-COMP-005

Проверки охватывают собственный интеграционный код инструментов, MCP-серверов и коннекторов, не подменяя контроль сторонних зависимостей.

Связь 1192 · AISAFE2-MIT-071 → AISAFE2-COMP-015

В контуре разработки и выпуска результаты ручной проверки, SAST, DAST, фаззинга, поиска секретов и abuse-case тестов связываются с точными commit и build и блокируют продвижение выше порога.

Связь 1193 · AISAFE2-MIT-071 → AISAFE2-COMP-017

Ручная проверка и целевые тесты охватывают собственные backend и бизнес-логику, включая авторизацию, переходы состояния и обязательные правила.

Связь 1194 · AISAFE2-MIT-071 → AISAFE2-COMP-019

Проверки охватывают собственный код принудительного применения политик и попытки его обхода; эффективная конфигурация в эксплуатации остаётся за AISAFE2-MIT-067.

Связь 1195 · AISAFE2-MIT-071 → AISAFE2-COMP-021

Проверки охватывают собственный код A2A-шлюзов, реестра и обработки межагентских сообщений, включая доверие, повтор и делегирование.

Связь 1196 · AISAFE2-MIT-001 → AISAFE2-MIT-047

Проверка типа, размера и структуры входа отсекает заведомо некорректные образцы до специальных испытаний и обработки допустимых состязательных входов.

Связь 1197 · AISAFE2-MIT-002 → AISAFE2-MIT-045

Безопасное представление и обработка вывода сохраняют границу исполнения, пока независимая проверка оценивает решение с существенными последствиями.

Связь 1198 · AISAFE2-MIT-003 → AISAFE2-MIT-004

Сборщик контекста сохраняет происхождение и уровень доверия каждого фрагмента и не позволяет недоверенному содержимому стать системной инструкцией. В пересекающейся области это частично снижает для меры «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» следующий риск или пробел: Вероятностный детектор не охватывает неизвестные техники и одновременно может пропускать атаки и блокировать легитимные запросы.

Ограничение: Разделение доверенных инструкций и содержимого не распознаёт все вредоносные запросы: атака может прийти и через доверенный ввод, а детектор по-прежнему требует настройки и проверки ошибок.

Связь 1199 · AISAFE2-MIT-003 → AISAFE2-MIT-007

Сборщик контекста сохраняет происхождение и уровень доверия каждого фрагмента и не позволяет недоверенному содержимому стать системной инструкцией. В пересекающейся области это частично снижает для меры «Допуск проверенных источников в контур RAG» следующий риск или пробел: Допущенный материал может стать вредоносным после компрометации источника; семантически вредный текст проходит формальные проверки и влияет на модель.

Ограничение: Разделение инструкций не проверяет подлинность, целостность и качество источника RAG; допущенный источник может быть скомпрометирован или содержать ложные сведения.

Связь 1200 · AISAFE2-MIT-003 → AISAFE2-MIT-008

Разметка происхождения и доверия фрагмента помогает не принять внедрённую в разрешённый фрагмент инструкцию за системную; право чтения этого фрагмента по-прежнему проверяется отдельно.

Ограничение: Уровень доверия RAG-фрагмента не подтверждает право пользователя читать его содержимое. Ошибки меток и устаревшие права устраняются проверкой доступа к каждому извлекаемому фрагменту.

Связь 1201 · AISAFE2-MIT-003 → AISAFE2-MIT-047

Разделение инструкций и недоверенного содержимого создаёт специальную границу внедрения вредоносных инструкций внутри более широкого контроля состязательных входов.

Связь 1202 · AISAFE2-MIT-003 → AISAFE2-MIT-066

Сохраняемая типизация источников позволяет отличить доверенную цель от пользовательских данных, RAG-фрагментов, результатов инструментов и сообщений других агентов.

Связь 1203 · AISAFE2-MIT-004 → AISAFE2-MIT-001

Версионируемый детектор выявляет признаки внедрения вредоносных инструкций и обхода ограничений и передаёт результат в заданную политику реакции. В пересекающейся области это частично снижает для меры «Проверка входных данных по схемам и правилам» следующий риск или пробел: Семантически вредоносный запрос может полностью соответствовать схеме и пройти проверку.

Ограничение: Детектор может обнаружить часть семантически вредоносных запросов, прошедших схему, но не заменяет проверку структуры и допустимых значений; неизвестные техники он может пропустить.

Связь 1204 · AISAFE2-MIT-004 → AISAFE2-MIT-003

Вероятностный детектор дает дополнительный сигнал о манипуляции, а архитектурное разделение сохраняет независимую принудительную границу доверия.

Связь 1205 · AISAFE2-MIT-004 → AISAFE2-MIT-015

Версионируемый детектор выявляет признаки внедрения вредоносных инструкций и обхода ограничений и передаёт результат в заданную политику реакции. В пересекающейся области это частично снижает для меры «Защита конфигурации системных промптов и инструкций инструментов» следующий риск или пробел: Инструкции могут быть выведены косвенно, а одобренная конфигурация может содержать логическую ошибку — раскрытие или обход правил через поведение модели без изменения защищённой конфигурации.

Ограничение: Детектор может остановить часть попыток выманить инструкции через модель, но не защищает хранение и изменение конфигурации системного промпта и не устраняет ошибку в одобренных правилах.

Связь 1206 · AISAFE2-MIT-004 → AISAFE2-MIT-031

Версионируемый детектор выявляет признаки внедрения вредоносных инструкций и обхода ограничений и передаёт результат в заданную политику реакции. В пересекающейся области это частично снижает для меры «Изолированное выполнение браузерных агентов и агентов управления компьютером» следующий риск или пробел: Адаптивная страница может скрыть смысл действия или влиять на модель через визуальный контент; агент выполняет разрешённое действие на скомпрометированной странице, смысл которого не распознан.

Ограничение: Детектор может заметить часть инструкций на странице, но не изолирует браузер, не ограничивает полномочия агента и может не распознать вредный смысл разрешённого действия.

Связь 1207 · AISAFE2-MIT-004 → AISAFE2-MIT-047

Детектор манипулятивного содержимого даёт вероятностный сигнал для одного класса состязательных входов и дополняет специальные испытания устойчивости.

Связь 1208 · AISAFE2-MIT-004 → AISAFE2-MIT-056

Сигнал внедрения вредоносной инструкции обслуживает одну категорию общей политики допустимости.

Связь 1209 · AISAFE2-MIT-005 → AISAFE2-MIT-002

Интерфейс связывает существенный вывод с фактически использованными источниками и явно показывает статус независимого подтверждения. В пересекающейся области это частично снижает для меры «Проверка и безопасная обработка AI-вывода» следующий риск или пробел: Формально валидный вывод может быть ложным, вредным или манипулятивным, и пользователь может принять некорректный, но корректно оформленный ответ.

Ограничение: Ссылка на источник и отметка о проверке помогают оценить ответ, но не обезвреживают активное содержимое для получателя и не подтверждают истинность самого источника.

Связь 1210 · AISAFE2-MIT-005 → AISAFE2-MIT-007

Интерфейс связывает существенный вывод с фактически использованными источниками и явно показывает статус независимого подтверждения. В пересекающейся области это частично снижает для меры «Допуск проверенных источников в контур RAG» следующий риск или пробел: Допуск источника не гарантирует истинность каждого утверждения.

Ограничение: Видимая ссылка и статус проверки не допускают источник в RAG и не подтверждают истинность каждого его утверждения; эти проверки выполняются отдельно.

Связь 1211 · AISAFE2-MIT-005 → AISAFE2-MIT-019

Интерфейс связывает существенный вывод с фактически использованными источниками и явно показывает статус независимого подтверждения. В пересекающейся области это частично снижает для меры «Подтверждение человеком высокорисковых действий до исполнения» следующий риск или пробел: Человек может ошибиться, не понять контекст или поддаться социальной инженерии, подтвердив вредное действие, выглядящее допустимым.

Ограничение: Происхождение вывода помогает человеку при подтверждении действия, но не гарантирует, что он поймёт последствия или отвергнет правдоподобное вредное действие.

Связь 1212 · AISAFE2-MIT-005 → AISAFE2-MIT-045

Отображение происхождения и статуса подтверждения дает проверяющему трассируемые входы, но окончательное решение остается за независимой проверкой.

Связь 1213 · AISAFE2-MIT-006 → AISAFE2-MIT-044

Общий лимит сокращает доступный объем запросов, а специализированный контроль дополнительно ограничивает извлекаемый из ответов сигнал модели.

Связь 1214 · AISAFE2-MIT-007 → AISAFE2-MIT-005

В рабочий RAG-контур допускаются только идентифицированные источники, прошедшие установленную проверку происхождения, целостности и качества. В пересекающейся области это частично снижает для меры «Отображение происхождения и статуса подтверждения AI-вывода» следующий риск или пробел: Корректная ссылка подтверждает происхождение фрагмента, но не истинность, актуальность или достаточность источника.

Ограничение: Допуск источника подтверждает его происхождение на момент проверки, но не делает каждое утверждение истинным или актуальным и не заменяет показ статуса подтверждения пользователю.

Связь 1215 · AISAFE2-MIT-007 → AISAFE2-MIT-009

Допуск источника и содержимого фиксирует доверенный состав базы знаний, который затем можно однозначно версионировать и безопасно откатывать.

Связь 1216 · AISAFE2-MIT-007 → AISAFE2-MIT-033

Результат проверки допуска источников и содержимого служит одним из входов блокирующего решения о выпуске изменения RAG.

Связь 1217 · AISAFE2-MIT-008 → AISAFE2-MIT-033

Результаты позитивных и негативных проверок гранулярной авторизации входят в специализированное решение о выпуске RAG-изменения.

Связь 1218 · AISAFE2-MIT-009 → AISAFE2-MIT-005

Версионирование связывает рабочую базу знаний с точным составом и позволяет атомарно восстановить последнюю допустимую версию. В пересекающейся области это частично снижает для меры «Отображение происхождения и статуса подтверждения AI-вывода» следующий риск или пробел: Корректно процитированный источник может быть ошибочным, устаревшим или вредоносным, и пользователь может принять убедительный, но неверный вывод даже при видимой атрибуции.

Ограничение: Версия базы знаний позволяет установить, какой источник использовался, но не подтверждает его истинность и не показывает пользователю статус независимой проверки ответа.

Связь 1219 · AISAFE2-MIT-009 → AISAFE2-MIT-007

Версионирование связывает рабочую базу знаний с точным составом и позволяет атомарно восстановить последнюю допустимую версию. В пересекающейся области это частично снижает для меры «Допуск проверенных источников в контур RAG» следующий риск или пробел: Не предотвращает последующее изменение данных, индексатора или политики доступа после проверки.

Ограничение: Откат возвращает допущенную версию базы знаний, но не предотвращает компрометацию источника до допуска и не проверяет права доступа или актуальность содержания после восстановления.

Связь 1220 · AISAFE2-MIT-009 → AISAFE2-MIT-024

Проверенный откат базы знаний дает сценарию реагирования конкретное действие восстановления после компрометации или ошибочного обновления.

Связь 1221 · AISAFE2-MIT-009 → AISAFE2-MIT-033

Связанная с версией возможность отката служит обязательным эксплуатационным подтверждением перед выпуском существенного RAG-изменения.

Связь 1222 · AISAFE2-MIT-009 → AISAFE2-MIT-046

Если резервный режим использует базу знаний, версионирование обеспечивает одну точную допущенную версию и очистку несовместимых кэшей при переключении или откате. Это предотвращает смешение и произвольное состояние версии, но не подтверждает свежесть или правильность данных и не покрывает сочетания отказов.

Ограничение: Точная версия базы знаний предотвращает смешение данных при переключении на резервный режим, но не подтверждает их свежесть или правильность и не покрывает отказ других зависимостей.

Связь 1223 · AISAFE2-MIT-010 → AISAFE2-MIT-008

Минимизация уменьшает объем чувствительных данных, которые вообще могут попасть в авторизованный результат извлечения.

Связь 1224 · AISAFE2-MIT-010 → AISAFE2-MIT-032

Ранняя минимизация сокращает множество данных, которое затем должен проверить окончательный исходящий контроль коннектора.

Связь 1225 · AISAFE2-MIT-010 → AISAFE2-MIT-033

Подтверждение минимизации чувствительных данных входит в набор актуальных результатов специализированного шлюза выпуска RAG.

Связь 1226 · AISAFE2-MIT-010 → AISAFE2-MIT-042

Минимизация до передачи уменьшает чувствительный обучающий материал, а испытания приватности отдельно проверяют утечки через поведение модели.

Связь 1227 · AISAFE2-MIT-011 → AISAFE2-MIT-008

Ограничение доступа и экспорта хранилища защищает сам индекс, а гранулярная авторизация отдельно фильтрует результат конкретного извлечения.

Связь 1228 · AISAFE2-MIT-011 → AISAFE2-MIT-033

Проверки доступа, экспорта и изоляции векторного хранилища дают шлюзу выпуска подтверждения по инфраструктурной границе RAG.

Связь 1229 · AISAFE2-MIT-012 → AISAFE2-MIT-034

Проверка целостности и безопасная загрузка подтверждают в охваченной цепочке, что на продвижение подан ожидаемый артефакт, а не подмененная версия.

Связь 1230 · AISAFE2-MIT-012 → AISAFE2-MIT-041

Проверка артефакта на границе загрузки дополняет управляемую передачу результата защищенного конвейера в хранилище или реестр.

Связь 1231 · AISAFE2-MIT-013 → AISAFE2-MIT-012

Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки. В пересекающейся области это частично снижает для меры «Проверка целостности и безопасная загрузка артефактов модели» следующий риск или пробел: Подписанный артефакт может содержать скрытую закладку или быть получен из скомпрометированного обучения — доверенный поставщик подписывает вредоносную либо ошибочную модель, целостность которой формально подтверждена.

Ограничение: Проверка цепочки поставки перед выпуском может выявить известный риск артефакта, но не гарантирует отсутствие скрытой закладки и не заменяет проверку целостности при загрузке.

Связь 1232 · AISAFE2-MIT-013 → AISAFE2-MIT-026

Состав и результаты предрелизных проверок создают исходный инвентарь, по которому после выпуска отслеживаются новые уязвимости.

Связь 1233 · AISAFE2-MIT-013 → AISAFE2-MIT-035

Шлюз выпуска проверяет модели, программные зависимости, образы, конфигурации и внешние компоненты по утверждённой политике цепочки поставки. В пересекающейся области это частично снижает для меры «Ведение проверяемого состава и происхождения компонентов AI-системы» следующий риск или пробел: Состав подтверждает прослеживаемость известных объектов, но не безопасность или отсутствие скрытой зависимости.

Ограничение: Проверка известных компонентов перед выпуском не обеспечивает полноту перечня зависимостей и непрерывное ведение их происхождения; скрытый компонент может остаться вне проверки.

Связь 1234 · AISAFE2-MIT-014 → AISAFE2-MIT-012

Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки. В пересекающейся области это частично снижает для меры «Проверка целостности и безопасная загрузка артефактов модели» следующий риск или пробел: Подписанный артефакт может содержать скрытую закладку или быть получен из скомпрометированного обучения — доверенный поставщик подписывает вредоносную либо ошибочную модель, целостность которой формально подтверждена.

Ограничение: Базовый профиль помогает заметить отклонение поведения модели, но не подтверждает целостность файла при загрузке и может не выявить закладку, присутствовавшую уже в исходной версии.

Связь 1235 · AISAFE2-MIT-014 → AISAFE2-MIT-013

Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки. В пересекающейся области это частично снижает для меры «Проверка компонентов цепочки поставки перед выпуском» следующий риск или пробел: Сканеры пропускают неизвестные закладки: вредоносный или логически небезопасный компонент не распознан известными проверками и проходит в подписанный выпуск.

Ограничение: Профиль поведения помогает заметить действие скрытой закладки в проверенной версии модели, но не сканирует другие компоненты и может не выявить закладку, не проявившуюся на оценочной выборке.

Связь 1236 · AISAFE2-MIT-014 → AISAFE2-MIT-034

Связанный с точной версией базовый профиль позволяет проверить поведение кандидата до его продвижения и отличить версии после выпуска.

Связь 1237 · AISAFE2-MIT-014 → AISAFE2-MIT-035

Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки. В пересекающейся области это частично снижает для меры «Ведение проверяемого состава и происхождения компонентов AI-системы» следующий риск или пробел: Инвентарь может содержать неверные сведения поставщика и не показывает поведение компонента: вредоносный компонент корректно учтён и подписан, но не распознан как опасный.

Ограничение: Профиль поведения помогает заметить отклонение модели, но не подтверждает полноту состава компонентов или достоверность сведений поставщика; опасное поведение может войти уже в исходный профиль.

Связь 1238 · AISAFE2-MIT-014 → AISAFE2-MIT-036

Фиксация версии и воспроизводимого базового профиля дает контрольную точку, без которой нельзя достоверно обнаруживать и разбирать последующие отклонения.

Связь 1239 · AISAFE2-MIT-014 → AISAFE2-MIT-039

Воспроизводимый профиль точной версии дает проверке пригодности идентифицированную исходную точку и сопоставимые результаты.

Связь 1240 · AISAFE2-MIT-014 → AISAFE2-MIT-040

Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки. В пересекающейся области это частично снижает для меры «Допуск и целостность обучающих данных, разметки и признаков» следующий риск или пробел: Разрешённый источник может быть содержательно ошибочным или уже отравленным, а сложная скрытая закладка может не проявиться в тестах.

Ограничение: Профиль может показать эффект отравления после обучения, но не проверяет допустимость и целостность самих обучающих данных; скрытая закладка может не проявиться в тестах.

Связь 1241 · AISAFE2-MIT-014 → AISAFE2-MIT-041

Базовый профиль связывает ожидаемое поведение с точными версиями модели, конфигурации, промпта и оценочной выборки. В пересекающейся области это частично снижает для меры «Защита конвейера обучения и дообучения» следующий риск или пробел: Авторизованный участник, скомпрометированная управляющая плоскость или аппаратная среда могут повлиять на обучение, внеся формально разрешённое, но вредоносное изменение.

Ограничение: Изменение поведения может указать на компрометацию обучения, но базовый профиль не защищает конвейер, доступы и вычислительную среду; вредное изменение может не проявиться на оценочной выборке.

Связь 1242 · AISAFE2-MIT-016 → AISAFE2-MIT-003

Реестр ограничивает доступный агенту набор инструментов и проверяет версию, операцию и аргументы каждого вызова до исполнения. В пересекающейся области это частично снижает для меры «Разделение доверенных инструкций и недоверенного контента при сборке контекста» следующий риск или пробел: Модель может семантически интерпретировать корректно размеченные данные как инструкцию даже при сохранённой разметке доверия; переход адаптивного текста в значимое действие должен блокироваться независимой политикой.

Ограничение: Ограничение инструментов сдерживает выполнение действия по инструкции из недоверенного текста, но не мешает модели принять этот текст за инструкцию или исказить ответ без вызова инструмента.

Связь 1243 · AISAFE2-MIT-016 → AISAFE2-MIT-017

Ограниченный список инструментов и проверенные аргументы уменьшают поверхность исполняемого кода, которую затем содержит изолированная среда.

Связь 1244 · AISAFE2-MIT-016 → AISAFE2-MIT-029

Проверка выбранного инструмента и аргументов формирует однозначный неизменяемый запрос, который внешний авторизатор может оценить непосредственно перед исполнением.

Связь 1245 · AISAFE2-MIT-017 → AISAFE2-MIT-004

Недоверенный код, файлы и инструменты исполняются в отдельной среде с ограниченными ресурсами, сетью, секретами и состоянием. В пересекающейся области это частично снижает для меры «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» следующий риск или пробел: Вероятностный детектор не охватывает неизвестные техники и одновременно может пропускать атаки и блокировать легитимные запросы.

Ограничение: Изоляция ограничивает последствия вредоносной инструкции при исполнении инструмента, но не распознаёт саму инструкцию и не устраняет ошибки детектора.

Связь 1246 · AISAFE2-MIT-017 → AISAFE2-MIT-016

Недоверенный код, файлы и инструменты исполняются в отдельной среде с ограниченными ресурсами, сетью, секретами и состоянием. В пересекающейся области это частично снижает для меры «Ограничение доступных инструментов и проверка их аргументов» следующий риск или пробел: Мера не определяет допустимость бизнес-цели внутри формально разрешённого действия и не изолирует само исполнение.

Ограничение: Изоляция сдерживает исполнение инструмента, но не определяет, был ли он разрешён агенту, допустимы ли аргументы и соответствует ли действие цели пользователя.

Связь 1247 · AISAFE2-MIT-017 → AISAFE2-MIT-026

Недоверенный код, файлы и инструменты исполняются в отдельной среде с ограниченными ресурсами, сетью, секретами и состоянием. В пересекающейся области это частично снижает для меры «Управление уязвимостями AI-компонентов в эксплуатации» следующий риск или пробел: Поставщик может не выпустить патч; эксплуатируемая неизвестная либо принятая по исключению уязвимость остаётся доступной до компенсирующего действия.

Ограничение: Изоляция снижает последствия эксплуатации уязвимости внутри неё, но не устраняет сам дефект; выход из среды или разрешённый внешний вызов остаются отдельными путями ущерба.

Связь 1248 · AISAFE2-MIT-017 → AISAFE2-MIT-031

Общая изоляция исполнения задаёт нижнюю границу, поверх которой браузерный профиль отдельно ограничивает файлы cookie, DOM, навигацию и действия в пользовательском интерфейсе.

Связь 1249 · AISAFE2-MIT-018 → AISAFE2-MIT-017

Минимальные полномочия процесса ограничивают доступные ресурсы даже внутри выбранной изолированной среды исполнения инструментов.

Связь 1250 · AISAFE2-MIT-018 → AISAFE2-MIT-030

Базовая машинная идентичность и минимальные полномочия задают верхнюю границу для еще более узкого делегированного полномочия коннектора.

Связь 1251 · AISAFE2-MIT-018 → AISAFE2-MIT-048

Каждая AI-нагрузка получает отдельную машинную идентичность и минимальные краткоживущие полномочия для конкретных получателей, ресурсов и операций. В пересекающейся области это частично снижает для меры «Аутентификация, целостность и защита от повтора межагентских сообщений» следующий риск или пробел: Скомпрометированный ключ или вредоносный аутентифицированный агент может отправить новое формально корректное сообщение с вредным смыслом.

Ограничение: Минимальные полномочия сужают ущерб от скомпрометированного агента, но не подтверждают целостность сообщений и не предотвращают повтор или вредное сообщение от доверенного отправителя.

Связь 1252 · AISAFE2-MIT-018 → AISAFE2-MIT-064

Машинная идентичность ограничивает получение секрета утверждённой рабочей нагрузкой.

Связь 1253 · AISAFE2-MIT-019 → AISAFE2-MIT-002

Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению. В пересекающейся области это частично снижает для меры «Проверка и безопасная обработка AI-вывода» следующий риск или пробел: Формально валидный вывод может быть ложным, вредным или манипулятивным, и пользователь может принять некорректный, но корректно оформленный ответ.

Ограничение: Подтверждение человеком касается только назначенных высокорисковых действий и не обезвреживает AI-вывод для других получателей; человек может одобрить убедительное ошибочное предложение.

Связь 1254 · AISAFE2-MIT-019 → AISAFE2-MIT-016

Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению. В пересекающейся области это частично снижает для меры «Ограничение доступных инструментов и проверка их аргументов» следующий риск или пробел: Формально допустимый инструмент с валидными аргументами может быть использован для нежелательной цели.

Ограничение: Человек может остановить нежелательную цель высокорискового вызова, но не заменяет список разрешённых инструментов и проверку аргументов; вредное действие может быть ошибочно одобрено.

Связь 1255 · AISAFE2-MIT-019 → AISAFE2-MIT-020

Действия установленного класса риска исполняются только после явного решения уполномоченного человека по неизменённому предложению. В пересекающейся области это частично снижает для меры «Аварийная остановка и возврат агента в безопасное состояние» следующий риск или пробел: Остановка не предотвращает первое действие: внешний эффект может уже возникнуть либо неучтённый компонент продолжает работу.

Ограничение: Предварительное подтверждение может предотвратить первое высокорисковое действие, но не останавливает уже запущенного агента и не возвращает систему в безопасное состояние.

Связь 1256 · AISAFE2-MIT-019 → AISAFE2-MIT-029

Решение уполномоченного человека по точным параметрам может выявить вредное действие, пропущенное формальной политикой; исполнение по-прежнему требует независимой авторизации.

Ограничение: Решение человека добавляет проверку для выбранных высокорисковых действий, но не заменяет независимую авторизацию каждого внешнего эффекта и может опираться на неполный контекст.

Связь 1257 · AISAFE2-MIT-019 → AISAFE2-MIT-031

Подтверждение человеком блокирует выбранные высокорисковые действия в пользовательском интерфейсе, тогда как браузерная изоляция ограничивает саму среду и доступные пути.

Связь 1258 · AISAFE2-MIT-019 → AISAFE2-MIT-037

Действующий технический шлюз подтверждения дает шаблону управления очередями принудительную точку остановки и неизменяемый предмет решения.

Связь 1259 · AISAFE2-MIT-020 → AISAFE2-MIT-024

Проверенная аварийная остановка дает сценарию реагирования немедленное действие локализации и перевод агента в безопасное состояние.

Связь 1260 · AISAFE2-MIT-021 → AISAFE2-MIT-033

Результаты проверок устойчивости к отравлению RAG и памяти входят в доказательства, которые специализированный шлюз оценивает перед выпуском.

Связь 1261 · AISAFE2-MIT-022 → AISAFE2-MIT-008

Сквозные события с идентичностью, объектом, решением и фактическим получателем позволяют проверить, что извлечение и производная передача прошли охваченную точку авторизации, и выявить наблюдаемый обход через кэш или производный объект. Журналирование поддерживает только проверку покрытия и расследование; оно не авторизует данные и не предотвращает утечку.

Ограничение: Журналы помогают обнаружить наблюдаемый обход авторизации при извлечении или передаче данных, но не принимают решение о доступе и не предотвращают первую утечку.

Связь 1262 · AISAFE2-MIT-022 → AISAFE2-MIT-023

Сквозно связанные события дают правилам обнаружения последовательность действий, идентичности и контекст, необходимые для корреляции атаки.

Связь 1263 · AISAFE2-MIT-022 → AISAFE2-MIT-024

Связанная телеметрия сохраняет временную линию и контекст, на которых строятся расследование, локализация и подтверждение восстановления.

Связь 1264 · AISAFE2-MIT-022 → AISAFE2-MIT-053

Наблюдаемые события с компонентом, субъектом, объектом и фактическим путём сопоставляются с версионируемой архитектурой; расхождение служит входом для пересмотра и блокирующего решения AISAFE2-MIT-053. Телеметрия не доказывает полноту архитектуры и не обнаруживает неинструментированный компонент или поток.

Ограничение: Сверка событий с архитектурой охватывает только инструментированные потоки; отсутствующий в телеметрии компонент или путь остаётся незамеченным.

Связь 1265 · AISAFE2-MIT-022 → AISAFE2-MIT-054

Сквозные события о субъекте, решении, объекте, получателе и фактическом эффекте позволяют проверить применение шлюза AISAFE2-MIT-054 на охваченных потоках и выявить наблюдаемый обход для расследования и исправления. Журналирование не принимает решение о допустимости, не блокирует обработку и не расширяет охват шлюза.

Ограничение: События показывают применение шлюза на охваченных потоках, но не принимают решение о допустимости, не блокируют обработку и не обнаруживают ненаблюдаемый обход.

Связь 1266 · AISAFE2-MIT-022 → AISAFE2-MIT-055

События с идентичностью оператора, действием, объектом и результатом дают проверяемое основание для внеочередного пересмотра или отзыва действующего доступа при наблюдаемом злоупотреблении. Журналирование не предотвращает первое действие и не заменяет выдачу, ограничение или отзыв прав AISAFE2-MIT-055.

Ограничение: Журнал даёт основание пересмотреть доступ после наблюдаемого злоупотребления, но не предотвращает первое действие и сам не отзывает полномочия.

Связь 1267 · AISAFE2-MIT-022 → AISAFE2-MIT-065

Сквозно связанные события дают архиву проверяемые идентификаторы источника, запроса, субъекта и фактического результата без обязательного сохранения полного содержимого.

Связь 1268 · AISAFE2-MIT-023 → AISAFE2-MIT-004

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» следующий риск или пробел: Детектор может быть обойдён новой формой атаки, оставляя риск нераспознанной инструкции.

Ограничение: Корреляция событий выявляет лишь наблюдаемые цепочки атаки; она не гарантирует обнаружение новой инструкции в отдельном запросе и не исправляет ошибки входного детектора.

Связь 1269 · AISAFE2-MIT-023 → AISAFE2-MIT-011

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Изоляция и ограничение экспорта векторных хранилищ» следующий риск или пробел: Разрешённый пользователь может выполнять смысловое извлечение малыми запросами; векторные представления сохраняют свойства исходных данных.

Ограничение: Поиск цепочек запросов может заметить часть постепенного извлечения данных, но не изолирует векторное хранилище и может пропустить медленное или распределённое извлечение.

Связь 1270 · AISAFE2-MIT-023 → AISAFE2-MIT-018

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Управление машинными идентичностями и минимальными полномочиями» следующий риск или пробел: Ошибка в классификации задачи может выдать избыточное право; злоупотребление корректно выданным правом в пределах разрешённой операции.

Ограничение: Сигнал о злоупотреблении корректно выданным правом появляется после наблюдаемой активности; он не сужает полномочия и не предотвращает первое действие.

Связь 1271 · AISAFE2-MIT-023 → AISAFE2-MIT-024

Коррелированное предупреждение с тяжестью и объяснением служит одним из проверяемых входов в сценарий реагирования на AI-инцидент.

Связь 1272 · AISAFE2-MIT-023 → AISAFE2-MIT-026

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Управление уязвимостями AI-компонентов в эксплуатации» следующий риск или пробел: Поставщик может не выпустить патч; эксплуатируемая неизвестная либо принятая по исключению уязвимость остаётся доступной до компенсирующего действия.

Ограничение: Сигнал об эксплуатации помогает начать исправление известного сценария, но не устраняет уязвимость и может не возникнуть при неизвестной или скрытой атаке.

Связь 1273 · AISAFE2-MIT-023 → AISAFE2-MIT-030

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Ограничение делегированных полномочий коннекторов» следующий риск или пробел: Скомпрометированный зарегистрированный коннектор злоупотребляет корректно суженным, но всё ещё достаточным полномочием.

Ограничение: Детектор может заметить часть злоупотреблений коннектором, но не сужает его делегированные права и может пропустить вредное действие в пределах формально разрешённой операции.

Связь 1274 · AISAFE2-MIT-023 → AISAFE2-MIT-032

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Проверка исходящих данных перед передачей через коннекторы» следующий риск или пробел: Семантически чувствительные данные могут быть выражены без известных признаков или распределены по нескольким разрешённым сообщениям, минуя классификатор.

Ограничение: Цепочка событий может показать последовательный вывод данных, но не проверяет содержание каждого исходящего сообщения и может пропустить распределённую или семантически скрытую утечку.

Связь 1275 · AISAFE2-MIT-023 → AISAFE2-MIT-038

Сигналы о последовательностях злоупотребления дают прямому ограничителю сервиса основания для блокировки, расследования и корректировки порогов.

Связь 1276 · AISAFE2-MIT-023 → AISAFE2-MIT-044

Корреляция повторных запросов и поведения субъекта дает специализированному контролю сигнал о вероятной попытке извлечения модели.

Связь 1277 · AISAFE2-MIT-023 → AISAFE2-MIT-047

Версионируемые правила выявляют наблюдаемые цепочки AI-атак и злоупотреблений и формируют проверяемый сигнал для реакции. В пересекающейся области это частично снижает для меры «Устойчивость к состязательным входным данным» следующий риск или пробел: Универсальной устойчивости к состязательным входным данным не существует: защита против одного класса атак может не работать против другого, сохраняя риск адаптивной или физической атаки вне проверенного диапазона.

Ограничение: Обнаружение знакомой цепочки атаки не делает модель устойчивой к новым, адаптивным или физическим воздействиям вне проверенного диапазона.

Связь 1278 · AISAFE2-MIT-023 → AISAFE2-MIT-053

Подтверждённая наблюдаемая цепочка атаки передаётся как конкретный сигнал для пересмотра затронутых путей, предпосылок и мер в модели угроз AISAFE2-MIT-053. Связь действует только для обнаруженных сценариев и не подтверждает полноту модели угроз или отсутствие иных путей воздействия.

Ограничение: В модель угроз попадают лишь обнаруженные и подтверждённые цепочки; отсутствие сигнала не означает, что другие пути атаки отсутствуют.

Связь 1279 · AISAFE2-MIT-024 → AISAFE2-MIT-020

Сценарий реагирования локализует затронутые AI-пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние. В пересекающейся области это частично снижает для меры «Аварийная остановка и возврат агента в безопасное состояние» следующий риск или пробел: Остановка не предотвращает первое действие: внешний эффект может уже возникнуть либо неучтённый компонент продолжает работу.

Ограничение: Реагирование локализует последствия после инцидента, но не останавливает первое действие агента и не гарантирует прекращение работы неучтённого компонента.

Связь 1280 · AISAFE2-MIT-024 → AISAFE2-MIT-046

Если непредусмотренное сочетание отказов приводит к инциденту несмотря на режим деградации, сценарий реагирования локализует затронутые пути, отзывает доступ, сохраняет подтверждения и восстанавливает проверенное состояние. Это поддержка реагирования и восстановления после инцидента, а не выбор резервного режима или предотвращение решения на устаревших данных.

Ограничение: Локализация и восстановление начинаются после инцидента; они не выбирают безопасный резервный режим и не предотвращают решение на устаревших данных.

Связь 1281 · AISAFE2-MIT-025 → AISAFE2-MIT-033

Общий процесс блокирующего решения обеспечивает точку принудительного контроля, работу с исключениями и привязку состава выпуска для специализированного профиля RAG.

Связь 1282 · AISAFE2-MIT-026 → AISAFE2-MIT-013

После выпуска процесс управления уязвимостями сопоставляет ставший известным дефект с точным рабочим компонентом, вводит исправление или компенсирующую меру и подтверждает результат. Связь действует только для известных и достижимых после выпуска дефектов; неизвестную закладку и классы, не охваченные сигналами, она не обнаруживает.

Ограничение: После выпуска обрабатываются известные и достижимые дефекты; это не расширяет проверку компонентов перед выпуском и не выявляет скрытую закладку без отдельного сигнала.

Связь 1283 · AISAFE2-MIT-026 → AISAFE2-MIT-017

Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией. В пересекающейся области это частично снижает для меры «Изолированное исполнение инструментов» следующий риск или пробел: Уязвимость изоляции или инфраструктуры может позволить выход из среды, а разрешённое внешнее действие остаётся реальным.

Ограничение: Исправление известной уязвимости среды снижает риск выхода из неё, но не ограничивает разрешённые внешние действия и не доказывает отсутствие неизвестных путей выхода.

Связь 1284 · AISAFE2-MIT-026 → AISAFE2-MIT-025

Процесс управления уязвимостями связывает находку с достижимым AI-компонентом, исправлением, повторной проверкой и фактической рабочей версией. В пересекающейся области это частично снижает для меры «Блокирующий шлюз выпуска AI-системы» следующий риск или пробел: Проверки охватывают конечный набор сценариев и могут устареть между решением и эксплуатацией; неизвестный или неохваченный тестами способ воздействия проявляется после корректно пройденного выпуска.

Ограничение: Исправление обнаруженного после выпуска дефекта не заменяет блокирующие проверки до выпуска; неизвестная или неохваченная тестами ошибка может попасть в эксплуатацию.

Связь 1285 · AISAFE2-MIT-026 → AISAFE2-MIT-049

Если межсессионное сохранение данных связано с известной и достижимой уязвимостью платформенной очистки или изоляции, процесс управления уязвимостями связывает находку с рабочей версией, вводит исправление либо компенсирующую меру и повторяет проверку. Он не перечисляет скрытое состояние ускорителя и не подтверждает очистку неподдерживаемых ресурсов.

Ограничение: Исправление известного дефекта очистки не перечисляет скрытое состояние ускорителя и не подтверждает очистку неподдерживаемых ресурсов.

Связь 1286 · AISAFE2-MIT-026 → AISAFE2-MIT-053

Управление уязвимостями принимает уведомление, устанавливает достижимость и связывает дефект с точной рабочей версией; этот результат становится проверяемым триггером и входом повторной оценки архитектурного риска, не подменяя остальные сигналы поставщика.

Связь 1287 · AISAFE2-MIT-027 → AISAFE2-MIT-025

Назначенные владельцы результатов, исключений и решения обеспечивают проверяемое право принять либо заблокировать выпуск.

Связь 1288 · AISAFE2-MIT-027 → AISAFE2-MIT-028

Закрепленный владелец позволяет однозначно выдать, ограничить, пересмотреть или отозвать утверждение и принятие остаточного риска.

Связь 1289 · AISAFE2-MIT-028 → AISAFE2-MIT-025

Ограниченная область и срок действия утверждений не позволяют шлюзу использовать просроченное либо более широкое подтверждение выпуска.

Связь 1290 · AISAFE2-MIT-028 → AISAFE2-MIT-027

Реестр утверждений связывает каждое заявление о мере с областью, подтверждением, владельцем, сроком и принятым остаточным риском. В пересекающейся области это частично снижает для меры «Закрепление ответственности и владельцев подтверждений» следующий риск или пробел: Назначение владельца не подтверждает выполнение его обязательства.

Ограничение: Проверка подтверждений показывает невыполненное обязательство владельца, но не исполняет его и не заменяет назначение ответственного.

Связь 1291 · AISAFE2-MIT-028 → AISAFE2-MIT-054

Проверка области, срока и подтверждений зарегистрированного решения до его использования не позволяет шлюзу опираться на просроченное или расширенное относительно подтверждений утверждение. Связь не исправляет ошибочность содержания решения, не обнаруживает скрытые пути и не усиливает техническое исполнение AISAFE2-MIT-054.

Ограничение: Проверка срока и области решения не исправляет его ошибочное содержание, не обнаруживает скрытый путь и не усиливает техническое исполнение шлюза.

Связь 1292 · AISAFE2-MIT-029 → AISAFE2-MIT-001

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Проверка входных данных по схемам и правилам» следующий риск или пробел: Мера не подтверждает полномочия субъекта и допустимость цели операции.

Ограничение: Авторизация действия проверяет полномочия и допустимую цель внешнего эффекта, но не проверяет структуру исходного запроса и смысл всех его значений.

Связь 1293 · AISAFE2-MIT-029 → AISAFE2-MIT-003

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Разделение доверенных инструкций и недоверенного контента при сборке контекста» следующий риск или пробел: Модель может семантически интерпретировать корректно размеченные данные как инструкцию даже при сохранённой разметке доверия; переход адаптивного текста в значимое действие должен блокироваться независимой политикой.

Ограничение: Независимый шлюз может остановить вредное действие по инструкции из недоверенного текста, но не мешает модели принять этот текст за инструкцию и исказить ответ без внешнего эффекта.

Связь 1294 · AISAFE2-MIT-029 → AISAFE2-MIT-004

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Выявление и обработка внедрения вредоносных инструкций и обхода ограничений» следующий риск или пробел: Сигнал детектора не заменяет авторизацию действий.

Ограничение: Шлюз ограничивает внешнее действие при пропуске вредоносной инструкции детектором, но не распознаёт саму инструкцию и не предотвращает манипуляцию ответом без действия.

Связь 1295 · AISAFE2-MIT-029 → AISAFE2-MIT-016

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Ограничение доступных инструментов и проверка их аргументов» следующий риск или пробел: Мера не определяет допустимость бизнес-цели внутри формально разрешённого действия и не изолирует само исполнение.

Ограничение: Проверка цели действия не заменяет реестр разрешённых инструментов и валидацию аргументов; формально допустимая цель может быть исполнена через опасный инструмент.

Связь 1296 · AISAFE2-MIT-029 → AISAFE2-MIT-017

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Изолированное исполнение инструментов» следующий риск или пробел: Не контролирует семантический эффект законного вызова внешнего сервиса.

Ограничение: Авторизация внешнего действия проверяет его допустимость, но не изолирует исполнение инструмента; вредоносный код может повредить среду до внешнего вызова.

Связь 1297 · AISAFE2-MIT-029 → AISAFE2-MIT-018

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Управление машинными идентичностями и минимальными полномочиями» следующий риск или пробел: Ошибка в классификации задачи может выдать избыточное право; злоупотребление корректно выданным правом в пределах разрешённой операции.

Ограничение: Шлюз может остановить отдельное действие с избыточным правом, но не исправляет выдачу машинной идентичности и не ограничивает все операции, совершаемые вне шлюза.

Связь 1298 · AISAFE2-MIT-029 → AISAFE2-MIT-030

Авторизация внешнего эффекта проверяет допустимость операции, а суженное полномочие коннектора ограничивает технически доступного получателя и ресурс.

Связь 1299 · AISAFE2-MIT-029 → AISAFE2-MIT-031

Внешняя авторизация проверяет допустимость существенного эффекта в пользовательском интерфейсе после того, как браузерная среда ограничила технически доступные действия.

Связь 1300 · AISAFE2-MIT-029 → AISAFE2-MIT-037

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений» следующий риск или пробел: Шлюз не заменяет компетентность, разделение обязанностей и техническую авторизацию.

Ограничение: Техническая авторизация не подтверждает компетентность проверяющего, качество предпросмотра и устойчивость процесса к перегрузке; ошибочно одобренное действие остаётся возможным.

Связь 1301 · AISAFE2-MIT-029 → AISAFE2-MIT-048

Независимый механизм авторизации проверяет предложенное агентом действие, получателя, ресурс и контекст до внешнего эффекта. В пересекающейся области это частично снижает для меры «Аутентификация, целостность и защита от повтора межагентских сообщений» следующий риск или пробел: Проверка подлинности не разрешает само действие; корректно подписанный злонамеренный отправитель остаётся допустимым источником сообщения.

Ограничение: Авторизация полученного действия не подтверждает подлинность и целостность межагентского сообщения и не предотвращает его повтор до шлюза.

Связь 1302 · AISAFE2-MIT-030 → AISAFE2-MIT-018

Коннекторы получают только полномочия, необходимые для заданного внешнего вызова, без доступа к исходным пользовательским и внутренним токенам. В пересекающейся области это частично снижает для меры «Управление машинными идентичностями и минимальными полномочиями» следующий риск или пробел: Сужение теряется, если последующий компонент игнорирует идентичность, подменяет её общей учётной записью или допускает выпуск нового широкого полномочия вне охваченного пути.

Ограничение: Суженное право коннектора действует только на охваченном маршруте; общий токен или выпуск нового широкого права в обход посредника сохраняют избыточные полномочия.

Связь 1303 · AISAFE2-MIT-030 → AISAFE2-MIT-032

Ограниченный получатель, маршрут и протокол коннектора задают проверяемый контекст для контроля окончательного исходящего объекта.

Связь 1304 · AISAFE2-MIT-031 → AISAFE2-MIT-017

Браузерный агент и агент управления компьютером работают в отдельном профиле с ограниченными данными, сетью и действиями интерфейса. В пересекающейся области это частично снижает для меры «Изолированное исполнение инструментов» следующий риск или пробел: Изоляция ограничена фактически охваченными файловыми, процессными, сетевыми и ресурсными границами и не охватывает состояние браузерного интерфейса без отдельной границы.

Ограничение: Изоляция браузерного агента охватывает его профиль и действия интерфейса, но не защищает исполнение других инструментов и внешние эффекты, разрешённые политикой браузера.

Связь 1305 · AISAFE2-MIT-032 → AISAFE2-MIT-010

Исходящий шлюз проверяет назначение, получателя, класс и минимально необходимый объём данных до передачи через коннектор. В пересекающейся области это частично снижает для меры «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» следующий риск или пробел: Классификатор и преобразование могут пропустить контекстное, закодированное или косвенно выводимое чувствительное значение.

Ограничение: Исходящий шлюз проверяет только передачу через охваченные коннекторы; он не устраняет чувствительные данные в памяти, ответе модели или другом маршруте и может пропустить косвенную утечку.

Связь 1306 · AISAFE2-MIT-032 → AISAFE2-MIT-030

Исходящий шлюз проверяет назначение, получателя, класс и минимально необходимый объём данных до передачи через коннектор. В пересекающейся области это частично снижает для меры «Ограничение делегированных полномочий коннекторов» следующий риск или пробел: Скомпрометированный зарегистрированный коннектор злоупотребляет корректно суженным, но всё ещё достаточным полномочием.

Ограничение: Проверка исходящих данных не сужает права коннектора; скомпрометированный коннектор может злоупотребить разрешённой операцией без передачи запрещённого содержимого.

Связь 1307 · AISAFE2-MIT-033 → AISAFE2-MIT-007

Специализированный шлюз блокирует выпуск изменений RAG, векторного хранилища или памяти без проверки точного состава и результатов испытаний. В пересекающейся области это частично снижает для меры «Допуск проверенных источников в контур RAG» следующий риск или пробел: Не предотвращает последующее изменение данных, индексатора или политики доступа после проверки.

Ограничение: Шлюз выпуска проверяет точную версию до публикации, но не предотвращает изменение источника, индексатора или политики доступа после допуска.

Связь 1308 · AISAFE2-MIT-033 → AISAFE2-MIT-021

Перед выпуском новой реализации или конфигурации памяти шлюз требует актуальные результаты проверки правил записи, чтения и срока жизни. Он не предотвращает постепенное отравление разрешёнными записями во время работы.

Ограничение: Проверка версии до выпуска не останавливает ложные или вредоносные записи, добавляемые в память во время работы; они требуют отдельного контроля записи и последующего обнаружения.

Связь 1309 · AISAFE2-MIT-034 → AISAFE2-MIT-052

Полный отзыв рабочей версии предоставляет процессу вывода действие прекращения модельных маршрутов.

Связь 1310 · AISAFE2-MIT-035 → AISAFE2-MIT-013

Проверяемый состав связывает рабочую версию AI-системы с происхождением моделей, кода, зависимостей, данных и конфигурации. В пересекающейся области это частично снижает для меры «Проверка компонентов цепочки поставки перед выпуском» следующий риск или пробел: Состав и проверки отражают только известные компоненты и выбранные классы дефектов.

Ограничение: Манифест показывает происхождение учтённых компонентов, но не выявляет скрытую зависимость и не доказывает, что проверенный компонент безопасен.

Связь 1311 · AISAFE2-MIT-035 → AISAFE2-MIT-025

Проверяемый состав связывает рабочую версию AI-системы с происхождением моделей, кода, зависимостей, данных и конфигурации. В пересекающейся области это частично снижает для меры «Блокирующий шлюз выпуска AI-системы» следующий риск или пробел: Успешное решение шлюза не доказывает фактическое применение контроля в среде выполнения.

Ограничение: Точный состав связывает решение о выпуске с артефактом, но не доказывает, что все проверки выполнены или что в рабочей среде применяются заявленные контроли.

Связь 1312 · AISAFE2-MIT-035 → AISAFE2-MIT-026

Инвентарь точных версий и расположения AI-компонентов задает эксплуатационный охват для поиска новых уязвимостей и определения затронутых развертываний.

Связь 1313 · AISAFE2-MIT-035 → AISAFE2-MIT-028

Проверяемый состав связывает рабочую версию AI-системы с происхождением моделей, кода, зависимостей, данных и конфигурации. В пересекающейся области это частично снижает для меры «Управление подтверждениями и остаточным риском утверждений» следующий риск или пробел: Качественная запись не усиливает слабый технический контроль: актуальное подтверждение верно для выборки, но фактическая рабочая конфигурация отклоняется после проверки без обнаружения.

Ограничение: Прослеживаемый состав помогает привязать подтверждение к версии, но не делает слабый контроль эффективным и не замечает изменение рабочей конфигурации без новой сверки.

Связь 1314 · AISAFE2-MIT-035 → AISAFE2-MIT-052

Проверяемый состав связывает рабочую версию AI-системы с происхождением моделей, кода, зависимостей, данных и конфигурации. В пересекающейся области это частично снижает для меры «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» следующий риск или пробел: Неактуальный инвентарь оставляет активные полномочия и маршруты вне процедуры вывода.

Ограничение: Процедура вывода охватит лишь учтённые компоненты и маршруты; теневой экземпляр или неактуальный манифест оставит доступ действующим.

Связь 1315 · AISAFE2-MIT-035 → AISAFE2-MIT-053

Сверка фактического состава с архитектурной записью может выявить наблюдаемый, но не описанный компонент или версию и запустить пересмотр модели угроз.

Ограничение: Проверка состава не обнаруживает неучтённый поток данных без отдельного наблюдения потоков и точек записи.

Связь 1316 · AISAFE2-MIT-036 → AISAFE2-MIT-009

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Версионирование и откат базы знаний» следующий риск или пробел: Откат сокращает длительность воздействия, но не предотвращает использование плохого содержимого до обнаружения: корректно версионированный, но содержательно вредный выпуск остаётся активным до сигнала мониторинга.

Ограничение: Сигнал дрейфа может запустить откат, но не восстанавливает версию сам и не предотвращает использование вредного содержимого до обнаружения.

Связь 1317 · AISAFE2-MIT-036 → AISAFE2-MIT-013

Если пропущенный при выпуске дефект проявляется как наблюдаемый дрейф или деградация по выбранным сигналам, мониторинг формирует сигнал и запускает ограничение, возврат версии либо исправление. Связь не обнаруживает скрытый компонент без наблюдаемого эффекта и не расширяет покрытие проверок цепочки поставки.

Ограничение: Мониторинг замечает лишь измеримый эффект дефекта после выпуска; он не проверяет скрытый компонент до выпуска и может пропустить вред без выбранного сигнала.

Связь 1318 · AISAFE2-MIT-036 → AISAFE2-MIT-014

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Фиксация версии и базового профиля поведения модели» следующий риск или пробел: Базовый профиль обнаруживает только измеряемые различия на выбранном наборе и не заменяет наблюдение эксплуатационного дрейфа или оценку неизвестного поведения.

Ограничение: Наблюдение эксплуатации зависит от выбранных метрик и доступных меток; оно не расширяет исходный профиль на неизвестное поведение и не обнаруживает незаметный для этих сигналов дефект.

Связь 1319 · AISAFE2-MIT-036 → AISAFE2-MIT-021

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Управление доверием и сроком жизни долговременной памяти» следующий риск или пробел: Повторные легитимные сообщения могут постепенно сформировать вредное состояние; разрешённая, но ложная запись влияет на будущие ответы до обнаружения и исправления.

Ограничение: Дрейф может указать на накопление ложных записей в памяти лишь после наблюдаемого эффекта; отдельную вредную запись до этого должен остановить контроль записи.

Связь 1320 · AISAFE2-MIT-036 → AISAFE2-MIT-025

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Блокирующий шлюз выпуска AI-системы» следующий риск или пробел: Проверки охватывают конечный набор сценариев и могут устареть между решением и эксплуатацией; неизвестный или неохваченный тестами способ воздействия проявляется после корректно пройденного выпуска.

Ограничение: Эксплуатационный мониторинг обнаруживает только выбранные последствия после выпуска и не заменяет блокирующие проверки до него; скрытый дефект может не дать измеримого сигнала.

Связь 1321 · AISAFE2-MIT-036 → AISAFE2-MIT-028

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Управление подтверждениями и остаточным риском утверждений» следующий риск или пробел: Реестр зависит от полноты источников и обнаружения изменений, может не заметить устаревание подтверждения.

Ограничение: Сигнал дрейфа помогает пересмотреть подтверждение, но не обновляет его автоматически и может не возникнуть при незаметном изменении конфигурации или области применения.

Связь 1322 · AISAFE2-MIT-036 → AISAFE2-MIT-033

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Шлюз выпуска изменений RAG, векторных хранилищ и памяти» следующий риск или пробел: Все известные условия могут быть выполнены, но содержательно новая атака или ошибка данных не входит в тестовый набор.

Ограничение: Наблюдение после выпуска может выявить пропущенную проверками ошибку, но не делает выпускной шлюз полнее и не предотвращает первый вредный ответ новой версии.

Связь 1323 · AISAFE2-MIT-036 → AISAFE2-MIT-039

Сигналы дрейфа и деградации инициируют повторную проверку пригодности модели для изменившегося сценария и данных.

Связь 1324 · AISAFE2-MIT-036 → AISAFE2-MIT-041

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Защита конвейера обучения и дообучения» следующий риск или пробел: Авторизованный участник, скомпрометированная управляющая плоскость или аппаратная среда могут повлиять на обучение, внеся формально разрешённое, но вредоносное изменение.

Ограничение: Поведенческий дрейф может указать на вредное изменение после обучения, но не защищает доступы и целостность конвейера; скрытая закладка может не проявиться в наблюдаемых метриках.

Связь 1325 · AISAFE2-MIT-036 → AISAFE2-MIT-047

Наблюдение сравнивает рабочие данные и поведение с базовым профилем и запускает заданную реакцию при значимом дрейфе или деградации. В пересекающейся области это частично снижает для меры «Устойчивость к состязательным входным данным» следующий риск или пробел: Устойчивость доказана только для выбранной модели угроз, версии и бюджетов атаки; адаптация противника или новый канал может обойти защиту.

Ограничение: Мониторинг может обнаружить последствия части новых атак, но не делает модель устойчивой к ним и не видит атаки без измеримого дрейфа.

Связь 1326 · AISAFE2-MIT-037 → AISAFE2-MIT-019

Шлюз распределяет запросы на подтверждение по классу риска, контролирует нагрузку проверяющих и качество решений и при перегрузке блокирует действие либо направляет его по заранее заданному маршруту эскалации. В пересекающейся области это частично снижает для меры «Подтверждение человеком высокорисковых действий до исполнения» следующий риск или пробел: Мера плохо масштабируется на частые либо срочные операции.

Ограничение: Маршрутизация и контроль нагрузки повышают шанс содержательной проверки, но не исключают ошибочное подтверждение человеком и не распространяют его на действия вне заданного класса риска.

Связь 1327 · AISAFE2-MIT-037 → AISAFE2-MIT-045

Шлюз распределяет запросы на подтверждение по классу риска, контролирует нагрузку проверяющих и качество решений и при перегрузке блокирует действие либо направляет его по заранее заданному маршруту эскалации. В пересекающейся области это частично снижает для меры «Независимая проверка решений с существенными последствиями» следующий риск или пробел: Независимость проверки зависит от реальных источников и полномочий проверяющего; ручной процесс остаётся подвержен ошибкам и давлению времени.

Ограничение: Разгрузка проверяющих не создаёт независимых источников сведений и не гарантирует компетентность или полномочия для оценки результата.

Связь 1328 · AISAFE2-MIT-038 → AISAFE2-MIT-001

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Проверка входных данных по схемам и правилам» следующий риск или пробел: Семантически вредоносный запрос может полностью соответствовать схеме и пройти проверку.

Ограничение: Политика злоупотреблений может остановить часть семантически вредных последовательностей, но не проверяет структуру каждого входа и может пропустить новый одиночный запрос.

Связь 1329 · AISAFE2-MIT-038 → AISAFE2-MIT-006

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Ограничение потребления ресурсов и стоимости» следующий риск или пробел: Локальный бюджет не ограничивает работу в неучтённом компоненте или распределённую нагрузку через разные идентичности.

Ограничение: Общая политика может ограничить распределённую нагрузку на охваченных интерфейсах, но не задаёт бюджет каждого шага, токена и инструмента и не видит неучтённый маршрут.

Связь 1330 · AISAFE2-MIT-038 → AISAFE2-MIT-010

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Минимизация и фильтрация чувствительных данных до передачи в AI-контур» следующий риск или пробел: Комбинация разрешённых признаков может раскрыть чувствительный факт — смысловая утечка без дословного секрета через разрешённый ответ или инструмент.

Ограничение: Ограничение массовых запросов затрудняет постепенное извлечение сведений, но не удаляет чувствительные данные из отдельного допустимого ответа и не видит распределённую атаку без общей идентичности.

Связь 1331 · AISAFE2-MIT-038 → AISAFE2-MIT-011

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Изоляция и ограничение экспорта векторных хранилищ» следующий риск или пробел: Изоляция и ограничение экспорта не предотвращают смысловое восстановление данных по векторам, которые субъект вправе искать и получать.

Ограничение: Обнаружение серии запросов не запрещает доступ к векторам и не исключает смысловое восстановление данных из немногих разрешённых ответов.

Связь 1332 · AISAFE2-MIT-038 → AISAFE2-MIT-023

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Выявление цепочек AI-атак и злоупотреблений» следующий риск или пробел: Правило видит только наблюдаемые события и известные последовательности, поэтому низкоскоростная, распределённая или новая атака может остаться ниже порога.

Ограничение: Единая политика злоупотреблений видит только охваченные запросы и знакомые сценарии; новую, медленную или распределённую цепочку атаки она может пропустить.

Связь 1333 · AISAFE2-MIT-038 → AISAFE2-MIT-032

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Проверка исходящих данных перед передачей через коннекторы» следующий риск или пробел: Семантически чувствительные данные могут быть выражены без известных признаков или распределены по нескольким разрешённым сообщениям, минуя классификатор.

Ограничение: Корреляция исходящих передач может заметить последовательную утечку, но не проверяет каждое сообщение и может пропустить чувствительный смысл без известных признаков.

Связь 1334 · AISAFE2-MIT-038 → AISAFE2-MIT-044

Политика противодействия злоупотреблениям ограничивает массовые и вредоносные сценарии использования AI-сервиса с учётом обхода локальных лимитов. В пересекающейся области это частично снижает для меры «Ограничение извлечения модели через интерфейс» следующий риск или пробел: Невозможно гарантировать предотвращение функционального копирования публично доступной модели: атакующий может распределять запросы между субъектами и источниками, обучая замену на допустимых ответах.

Ограничение: Лимиты запросов затрудняют массовое копирование поведения модели, но не предотвращают обучение замены на публично доступных ответах, особенно при распределении запросов между субъектами.

Связь 1335 · AISAFE2-MIT-039 → AISAFE2-MIT-014

Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения. В пересекающейся области это частично снижает для меры «Фиксация версии и базового профиля поведения модели» следующий риск или пробел: Базовый профиль обнаруживает только измеряемые различия на выбранном наборе и не заменяет наблюдение эксплуатационного дрейфа или оценку неизвестного поведения.

Ограничение: Оценка пригодности касается проверенной версии и сценариев до допуска; она не обнаруживает последующий дрейф или поведение вне тестовой выборки.

Связь 1336 · AISAFE2-MIT-039 → AISAFE2-MIT-015

Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения. В пересекающейся области это частично снижает для меры «Защита конфигурации системных промптов и инструкций инструментов» следующий риск или пробел: Инструкции могут быть выведены косвенно, а одобренная конфигурация может содержать логическую ошибку — раскрытие или обход правил через поведение модели без изменения защищённой конфигурации.

Ограничение: Тест пригодности может выявить утечку через поведение модели, но не защищает хранение и изменение системных инструкций и может не охватить новый способ их косвенного извлечения.

Связь 1337 · AISAFE2-MIT-039 → AISAFE2-MIT-036

Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения. В пересекающейся области это частично снижает для меры «Обнаружение и обработка дрейфа, смещения данных и деградации» следующий риск или пробел: Плохой базовый набор показателей и редкие события могут скрыть деградацию: опасное поведение возникает в малой подгруппе и не меняет агрегированные метрики до значительного воздействия.

Ограничение: Предварительная оценка помогает выбрать метрики для наблюдения, но редкий вред в малой группе может не изменить агрегированные показатели в эксплуатации.

Связь 1338 · AISAFE2-MIT-039 → AISAFE2-MIT-040

Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения. В пересекающейся области это частично снижает для меры «Допуск и целостность обучающих данных, разметки и признаков» следующий риск или пробел: Разрешённый источник может быть содержательно ошибочным или уже отравленным, а сложная скрытая закладка может не проявиться в тестах.

Ограничение: Тест пригодности может обнаружить проявившееся отравление, но не проверяет происхождение и целостность обучающих данных; скрытая закладка может не сработать на выборке.

Связь 1339 · AISAFE2-MIT-039 → AISAFE2-MIT-041

Оценка пригодности проверяет модель на репрезентативных сценариях и порогах до выбора версии или расширения области применения. В пересекающейся области это частично снижает для меры «Защита конвейера обучения и дообучения» следующий риск или пробел: Изоляция конвейера не подтверждает качество входных данных, пригодность модели и безопасность последующей загрузки артефакта.

Ограничение: Оценка результата обучения не защищает доступы и среду конвейера; формально разрешённое вредное изменение может не проявиться в тестовых сценариях.

Связь 1340 · AISAFE2-MIT-039 → AISAFE2-MIT-047

Общая проверка пригодности задаёт сценарий, метрики и приёмочные границы, которые специальные состязательные испытания расширяют атакующими входами.

Связь 1341 · AISAFE2-MIT-039 → AISAFE2-MIT-051

Если обязательная стадия анализа использует оцениваемую модель, испытание её точной версии на сценариях и порогах может выявить слабый детектор до выпуска.

Ограничение: Проверка пригодности модели не исправляет слабую политику, неполную выборку и немодельные стадии обязательного рабочего анализа.

Связь 1342 · AISAFE2-MIT-040 → AISAFE2-MIT-041

Допущенные и однозначно идентифицированные данные, разметка и признаки образуют проверяемые входы защищенного запуска обучения.

Связь 1343 · AISAFE2-MIT-040 → AISAFE2-MIT-042

Управляемое происхождение и допуск обучающих данных определяют точный набор, для которого затем проверяется приватность и утечка через модель.

Связь 1344 · AISAFE2-MIT-040 → AISAFE2-MIT-050

До обучения проверяются происхождение, права, целостность, разметка и допустимость каждого набора данных и признака. В пересекающейся области это частично снижает для меры «Неизменяемое версионирование и проверяемое восстановление наборов данных» следующий риск или пробел: Повреждение или нежелательное содержимое могли попасть во все копии до фиксации версии; восстановление возвращает сохранённую версию, но не доказывает корректность исходного набора данных.

Ограничение: Допуск данных снижает риск сохранить повреждённый набор, но не подтверждает независимость восстановительной копии и не исключает ошибку, уже присутствующую во всех версиях.

Связь 1345 · AISAFE2-MIT-040 → AISAFE2-MIT-060

Решение о допуске данных становится обязательным входом контролируемого продвижения между стадиями.

Связь 1346 · AISAFE2-MIT-043 → AISAFE2-MIT-012

Артефакты модели защищаются авторизацией, шифрованием, учётом копий и ограничением экспорта на всём управляемом пути. В пересекающейся области это частично снижает для меры «Проверка целостности и безопасная загрузка артефактов модели» следующий риск или пробел: Проверка целостности доступного артефакта не ограничивает его чтение или копирование.

Ограничение: Ограничение чтения артефакта уменьшает риск его копирования, но не подтверждает целостность версии при загрузке и не выявляет вредный код в разрешённом артефакте.

Связь 1347 · AISAFE2-MIT-043 → AISAFE2-MIT-044

Артефакты модели защищаются авторизацией, шифрованием, учётом копий и ограничением экспорта на всём управляемом пути. В пересекающейся области это частично снижает для меры «Ограничение извлечения модели через интерфейс» следующий риск или пробел: Ограничение интерфейса не защищает доступный артефакт.

Ограничение: Защита файла модели ограничивает прямое копирование весов, но не мешает воспроизводить поведение модели через разрешённый интерфейс запросов.

Связь 1348 · AISAFE2-MIT-044 → AISAFE2-MIT-038

Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели. В пересекающейся области это частично снижает для меры «Контроль злоупотребления AI-сервисом» следующий риск или пробел: Злоумышленник может распределять действия между сервисами и субъектами, сохраняя риск вредоносного использования допустимых возможностей без наблюдаемого шаблона.

Ограничение: Лимиты и сигналы извлечения модели охватывают этот сценарий злоупотребления, но не выявляют другие вредные действия AI-сервиса и могут пропустить распределённые запросы.

Связь 1349 · AISAFE2-MIT-044 → AISAFE2-MIT-042

Сквозной бюджет и ограничение информативности ответов затрудняют массовое извлечение поведения или защищаемых свойств модели. В пересекающейся области это частично снижает для меры «Контроль приватности обучающих данных и утечек через модель» следующий риск или пробел: Проверка покрывает только выбранные атаки, версию модели, интерфейс и базовую частоту; новый способ извлечения может пройти.

Ограничение: Ограничение массовых запросов снижает доступный объём ответов, но не проверяет, восстанавливают ли отдельные ответы обучающие данные; новая техника может пройти в пределах лимита.

Связь 1350 · AISAFE2-MIT-045 → AISAFE2-MIT-002

Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями. В пересекающейся области это частично снижает для меры «Проверка и безопасная обработка AI-вывода» следующий риск или пробел: Формально валидный вывод может быть ложным, вредным или манипулятивным, и пользователь может принять некорректный, но корректно оформленный ответ.

Ограничение: Независимая проверка касается решений с существенными последствиями и не обезвреживает AI-вывод для активного отображения или исполнения; проверяющий тоже может ошибиться.

Связь 1351 · AISAFE2-MIT-045 → AISAFE2-MIT-005

Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями. В пересекающейся области это частично снижает для меры «Отображение происхождения и статуса подтверждения AI-вывода» следующий риск или пробел: Корректно процитированный источник может быть ошибочным, устаревшим или вредоносным, и пользователь может принять убедительный, но неверный вывод даже при видимой атрибуции.

Ограничение: Проверка решения может выявить ложный источник, но не восстанавливает происхождение ответа без исходных ссылок и не охватывает каждый вывод, который видит пользователь.

Связь 1352 · AISAFE2-MIT-045 → AISAFE2-MIT-019

Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями. В пересекающейся области это частично снижает для меры «Подтверждение человеком высокорисковых действий до исполнения» следующий риск или пробел: Не доказывает правильность человеческой оценки.

Ограничение: Независимая проверка помогает человеку оценить действие, но не гарантирует правильность его подтверждения и не заменяет обязательную остановку до исполнения.

Связь 1353 · AISAFE2-MIT-045 → AISAFE2-MIT-037

Независимый проверяющий подтверждает исходные данные, применённые правила и вывод до решения с существенными последствиями. В пересекающейся области это частично снижает для меры «Управление маршрутизацией, нагрузкой и качеством человеческих подтверждений» следующий риск или пробел: Проверяющий может ошибиться или быть введён в заблуждение даже при приемлемой нагрузке: скоординированное либо ошибочное подтверждение вредного действия, формально соответствующее процессу.

Ограничение: Дополнительная оценка содержания не управляет очередью и нагрузкой проверяющих; оба проверяющих могут разделить ошибку или опереться на один ложный источник.

Связь 1354 · AISAFE2-MIT-048 → AISAFE2-MIT-003

Межагентское сообщение принимается только после проверки отправителя, получателя, целостности, связи с задачей, свежести и защиты от повтора. В пересекающейся области это частично снижает для меры «Разделение доверенных инструкций и недоверенного контента при сборке контекста» следующий риск или пробел: Граница зависит от полного охвата всех сборщиков, преобразований и межагентских передач.

Ограничение: Подлинность межагентского сообщения не задаёт уровень доверия вложенного текста; подписанное сообщение может содержать недоверенную инструкцию, которую сборщик контекста обязан отделить.

Связь 1355 · AISAFE2-MIT-048 → AISAFE2-MIT-022

Межагентское сообщение принимается только после проверки отправителя, получателя, целостности, связи с задачей, свежести и защиты от повтора. В пересекающейся области это частично снижает для меры «Регистрация и сквозное связывание событий AI-цепочки» следующий риск или пробел: Неинструментированный или скомпрометированный компонент выполняет действие без достоверного события; нарушение целостности источника способно сформировать убедительную, но ложную трассировку.

Ограничение: Проверка межагентских сообщений подтверждает их передачу на охваченном пути, но не создаёт достоверные события обо всех действиях и не видит неинструментированный компонент.

Связь 1356 · AISAFE2-MIT-048 → AISAFE2-MIT-029

Проверенное происхождение, целостность и свежесть межагентского сообщения формируют неизменяемый запрос, который внешний авторизатор оценивает перед эффектом.

Связь 1357 · AISAFE2-MIT-049 → AISAFE2-MIT-043

Изоляция и очистка общей среды инференса предотвращают перенос состояния между арендаторами, пользователями, задачами и последовательными заданиями. В пересекающейся области это частично снижает для меры «Защита артефактов модели от несанкционированного получения» следующий риск или пробел: Исполняемую модель нельзя считать некопируемой: привилегированный инсайдер, компрометация управляющей плоскости, доступ к памяти GPU или аппаратная атака могут привести к получению модели.

Ограничение: Очистка общей среды снижает риск получить остатки модели через память, но не защищает доступ к сохранённому артефакту и не исключает чтение во время работы привилегированным субъектом.

Связь 1358 · AISAFE2-MIT-053 → AISAFE2-MIT-018

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Управление машинными идентичностями и минимальными полномочиями» следующий риск или пробел: Сужение теряется, если последующий компонент игнорирует идентичность, подменяет её общей учётной записью или допускает выпуск нового широкого полномочия вне охваченного пути.

Ограничение: Модель архитектуры помогает обнаружить общую учётную запись и широкий путь полномочий, но сама не выдаёт и не отзывает права; неучтённый маршрут останется вне проверки.

Связь 1359 · AISAFE2-MIT-053 → AISAFE2-MIT-020

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Аварийная остановка и возврат агента в безопасное состояние» следующий риск или пробел: Остановка действует только на известные управляемые компоненты, полномочия и соединения; недоступный дочерний процесс может завершиться позже.

Ограничение: Описание компонентов помогает задать охват остановки, но не гарантирует отзыв полномочий и завершение каждого дочернего процесса; неучтённый путь может продолжить работу.

Связь 1360 · AISAFE2-MIT-053 → AISAFE2-MIT-022

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Регистрация и сквозное связывание событий AI-цепочки» следующий риск или пробел: Корреляция ограничена компонентами, транспортами и полями, которые действительно создают события.

Ограничение: Архитектура помогает перечислить точки событий, но не создаёт телеметрию в неинструментированном компоненте и не доказывает полноту записанных полей.

Связь 1361 · AISAFE2-MIT-053 → AISAFE2-MIT-025

Версионируемая архитектура и модель угроз задают проверяемую область, состав применимых проверок и триггеры повторного решения шлюза выпуска.

Связь 1362 · AISAFE2-MIT-053 → AISAFE2-MIT-032

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Проверка исходящих данных перед передачей через коннекторы» следующий риск или пробел: Контроль зависит от полного охвата фактических точек отправки; закодированные или переданные через неинструментированный маршрут данные могут обойти проверку.

Ограничение: Перечень путей помогает расставить исходящие шлюзы, но не перехватывает неучтённую передачу и не распознаёт чувствительное содержание, скрытое преобразованием.

Связь 1363 · AISAFE2-MIT-053 → AISAFE2-MIT-034

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Контролируемое продвижение моделей через реестр» следующий риск или пробел: Реестр не предотвращает обход через неучтённую среду или прямую смену конфигурации среды выполнения.

Ограничение: Архитектура помогает найти пути продвижения модели, но не запрещает прямую смену рабочей конфигурации или выпуск через неучтённую среду.

Связь 1364 · AISAFE2-MIT-053 → AISAFE2-MIT-046

Версионируемая архитектура и модель угроз перечисляют зависимости, состояния и значимые сочетания отказов и делают неизвестный путь блокирующим расхождением до выпуска или продолжения использования. Связь поддерживает полноту матрицы отказов, но не подтверждает свежесть резервных данных и не исполняет безопасное переключение.

Ограничение: Матрица зависимостей не подтверждает свежесть резервных данных и не выполняет безопасное переключение при отказе.

Связь 1365 · AISAFE2-MIT-053 → AISAFE2-MIT-050

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Неизменяемое версионирование и проверяемое восстановление наборов данных» следующий риск или пробел: Независимая по имени копия может разделять домен отказа, учётную запись или ошибку удаления с основной.

Ограничение: Описание доменов отказа помогает оценить независимость копий, но не делает её фактической: обе копии могут зависеть от одной учётной записи, платформы или ошибки удаления.

Связь 1366 · AISAFE2-MIT-053 → AISAFE2-MIT-051

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Контроль полноты обязательного анализа и безопасный выпуск результата» следующий риск или пробел: Контроль не обнаруживает обязательство, ошибочно исключённое из объявленного профиля.

Ограничение: Модель угроз помогает выявить пропущенный путь проверки, но не обнаружит обязательство, ошибочно исключённое из принятого профиля, и не выполняет сами проверки.

Связь 1367 · AISAFE2-MIT-053 → AISAFE2-MIT-052

Архитектура, потоки данных и модель угроз версионируются, пересматриваются при каждом существенном изменении и служат входом для решения о выпуске. В пересекающейся области это частично снижает для меры «Управляемый жизненный цикл и полный вывод AI-компонента из эксплуатации» следующий риск или пробел: Неучтённая копия, общий административный контур, физический носитель или внешняя сторона могут сохранить данные или доступ после логического вывода.

Ограничение: Архитектура помогает перечислить ресурсы для вывода, но не подтверждает удаление неучтённых копий, физических носителей или доступа внешней стороны.

Связь 1368 · AISAFE2-MIT-053 → AISAFE2-MIT-054

Версионируемая архитектура перечисляет классы данных, получателей, конечные точки и значимые пути и делает неизвестный компонент или поток блокирующим расхождением. Это поддерживает выявление скрытого пути и невидимого атрибута, но не определяет правильность решения владельца и не исполняет технический шлюз AISAFE2-MIT-054.

Ограничение: Перечень путей и классов данных не определяет правильность решения владельца о допустимости обработки и не исполняет технический шлюз; неизвестный поток остаётся вне охвата до обнаружения.

Связь 1369 · AISAFE2-MIT-053 → AISAFE2-MIT-057

Версионируемая архитектурная модель предоставляет утверждённый инвентарь необходимых сетевых соединений.

Связь 1370 · AISAFE2-MIT-053 → AISAFE2-MIT-070

Версионируемые архитектура и модель угроз задают границы активной оценки, значимые поверхности атаки и обязательные сценарии для точных версий системы.

Связь 1371 · AISAFE2-MIT-053 → AISAFE2-MIT-071

Модель угроз выделяет критичные пути собственного кода и бизнес-логики и определяет обязательные техники ручной и автоматизированной проверки перед выпуском.

Связь 1372 · AISAFE2-MIT-055 → AISAFE2-MIT-011

Плановый пересмотр и отзыв исключают просроченные, общие и не обусловленные ролью административные права к охваченному векторному хранилищу. Это сокращает только окно и объём доступа через лишнюю или утратившую основание учётную запись; действующую скомпрометированную роль и неинвентаризированный путь мера не блокирует.

Ограничение: Пересмотр известных административных прав не предотвращает компрометацию действующей роли и не усиливает техническую изоляцию векторного хранилища.

Связь 1373 · AISAFE2-MIT-055 → AISAFE2-MIT-034

Ролевой процесс пересматривает и отзывает просроченные, лишние и утратившие основание права продвижения моделей через охваченный реестр. Он уменьшает риск продвижения через ненужное или устаревшее полномочие, но не блокирует вредоносное действие действующего авторизованного участника и не защищает скомпрометированную управляющую плоскость.

Ограничение: Отзыв устаревших прав продвижения не останавливает вредное действие текущего уполномоченного участника и не защищает скомпрометированную управляющую плоскость.

Связь 1374 · AISAFE2-MIT-055 → AISAFE2-MIT-041

Выдача, пересмотр и отзыв прав исключают лишние, просроченные и утратившие основание полномочия человека на изменение охваченного конвейера обучения. Это уменьшает только риск изменения через ненужный доступ; действующего злонамеренного участника, компрометацию управляющей плоскости и аппаратной среды мера не блокирует.

Ограничение: Пересмотр человеческих прав не предотвращает вредное изменение действующим уполномоченным участником, компрометацию управляющей плоскости или аппаратной среды.

Связь 1375 · AISAFE2-MIT-055 → AISAFE2-MIT-043

Сверка ролей и фактических назначений удаляет просроченные, прямые и не обусловленные ролью административные права к артефактам модели на охваченных путях. Она сокращает поверхность привилегированного доступа, но не контролирует действия текущего уполномоченного администратора и не обнаруживает путь вне инвентаря.

Ограничение: Сверка известных назначений не контролирует действия текущего администратора и не обнаруживает путь к артефакту вне инвентаря.

Связь 1376 · AISAFE2-MIT-055 → AISAFE2-MIT-049

Пересмотр и отзыв прав удаляют лишний, просроченный или утративший основание доступ оператора к охваченной общей среде инференса. Это ограничивает только административный срез риска; мера не очищает буферы и микропрограмму, не закрывает побочные каналы и не блокирует текущего уполномоченного злоумышленника.

Ограничение: Отзыв лишнего доступа оператора не очищает буферы или микропрограмму, не закрывает побочные каналы и не останавливает текущего уполномоченного злоумышленника.

Связь 1377 · AISAFE2-MIT-055 → AISAFE2-MIT-050

Сверка ролей и фактических назначений исключает общие, лишние и просроченные административные права, способные изменить или удалить охваченные основную и восстановительную копии. Это поддерживает разделение доступа, но не создаёт независимый домен отказа и не предотвращает авторизованную ошибку платформы или удаления.

Ограничение: Разделение административных прав не создаёт независимый домен отказа и не предотвращает ошибку удаления, выполненного с действующим полномочием.

Связь 1378 · AISAFE2-MIT-055 → AISAFE2-MIT-052

При выводе компонента смена обязанностей или завершение работ запускают отзыв прав операторов, разработчиков и администраторов во всех охваченных MLOps-системах. Связь покрывает только удаление инвентаризированного человеческого доступа; она не обнаруживает неучтённые копии, физические носители и доступ внешней стороны.

Ограничение: Отзыв известных человеческих прав не обнаруживает копии, физические носители и доступ внешних сторон и не подтверждает полный вывод компонента.

Связь 1379 · AISAFE2-MIT-055 → AISAFE2-MIT-059

Актуальный перечень привилегированных человеческих аккаунтов определяет охват политики стойкой аутентификации.

Связь 1380 · AISAFE2-MIT-057 → AISAFE2-MIT-054

Обязательный сетевой маршрут может передать исходящий поток отдельному шлюзу политики данных до отправки получателю.

Ограничение: Связь действует только для потоков, которые нельзя обойти и по которым шлюз видит класс данных, получателя и цель обработки; один сетевой список адресов этого не проверяет.

Связь 1381 · AISAFE2-MIT-060 → AISAFE2-MIT-041

Опубликованная версия после контролируемого продвижения становится защищённым входом задания обучения.

Связь 1382 · AISAFE2-MIT-064 → AISAFE2-MIT-015

Ссылки на секреты не допускают открытые значения в конфигурацию промптов и инструментов.

Связь 1383 · AISAFE2-MIT-065 → AISAFE2-MIT-024

Проверяемая сохранность материалов позволяет сценарию реагирования установить затронутые объекты, подтвердить локализацию и доказательно проверить восстановленное состояние.

Связь 1384 · AISAFE2-MIT-066 → AISAFE2-MIT-029

Точная версия цели, возможностей и делегирования даёт внешнему авторизатору проверяемую границу для решения по конкретным ресурсу, операции и получателю.

Связь 1385 · AISAFE2-MIT-067 → AISAFE2-MIT-025

Результат проверки точной эффективной конфигурации по применимому эталону становится обязательным подтверждением блокирующего решения о выпуске.

Связь 1386 · AISAFE2-MIT-068 → AISAFE2-MIT-024

Проверенное восстановление точной версии критичного AI-актива предоставляет сценарию реагирования конкретное действие возврата в доверенное состояние.

Связь 1387 · AISAFE2-MIT-069 → AISAFE2-MIT-054

Наблюдаемый наследованный класс предоставляет техническому шлюзу атрибут для решения об обработке, хранении и передаче вектора.

Связь 1388 · AISAFE2-MIT-069 → AISAFE2-MIT-062

Наследованный класс чувствительности определяет обязательный режим защиты каждой сохраняемой копии векторного представления.

Связь 1389 · AISAFE2-MIT-069 → AISAFE2-MIT-063

Наследованный класс чувствительности определяет требования к каналу и получателю при передаче векторного представления.

Связь 1390 · AISAFE2-MIT-070 → AISAFE2-MIT-022

Контрольные атаки проверяют формирование и сквозную корреляцию обязательной телеметрии на охваченных путях AI-системы.

Ограничение: Контрольная атака проверяет телеметрию лишь на выбранных путях и в момент оценки; она не создаёт инструментацию и постоянную доставку событий.

Связь 1391 · AISAFE2-MIT-070 → AISAFE2-MIT-023

Активные сценарии воспроизводят утверждённые цепочки атак и проверяют срабатывание и маршрут соответствующих правил обнаружения.

Ограничение: Активная проверка охватывает выбранные цепочки атак; она не создаёт правила и непрерывную работу обнаружения.

Связь 1392 · AISAFE2-MIT-070 → AISAFE2-MIT-024

Учения в рамках активной оценки проверяют эскалацию, локализацию, восстановление и участие людей на утверждённых сценариях.

Ограничение: Учение проверяет выбранный сценарий, но не создаёт дежурство, полномочия и технические возможности реагирования на реальный инцидент.

Связь 1393 · AISAFE2-MIT-070 → AISAFE2-MIT-025

Свежие результаты активных атак и ручной экспертной оценки, связанные с точными версиями системы и конфигурации, становятся обязательным подтверждением блокирующего решения о выпуске.

Связь 1394 · AISAFE2-MIT-070 → AISAFE2-MIT-026

Подтверждённые находки в тестируемой рабочей версии получают владельца, решение, срок и повтор исходного сценария после исправления.

Ограничение: Находки активной проверки относятся к оцененной версии и сценариям; она не обеспечивает полный инвентарь уязвимостей, уведомления и развёртывание исправлений.

Связь 1395 · AISAFE2-MIT-070 → AISAFE2-MIT-067

Активные сценарии и ручная оценка выявляют достижимые небезопасные настройки и обходы в точной тестируемой конфигурации.

Ограничение: Оценка даёт снимок выбранных настроек; она не определяет постоянный эталон и не блокирует дрейф эффективной конфигурации после проверки.

Связь 1396 · AISAFE2-MIT-070 → AISAFE2-MIT-071

Red teaming и ручная оценка могут выявить дефекты собственного кода и бизнес-логики и передать их в выпускной контроль.

Ограничение: Ручная оценка может найти дефект в проверенной версии, но не заменяет систематический анализ кода, тесты разных классов и блокирующий контроль выпуска.

Связь 1397 · AISAFE2-MIT-071 → AISAFE2-MIT-025

Результаты ручной проверки и выбранных проверок безопасности, связанные с точными commit и build, становятся обязательным подтверждением блокирующего решения о выпуске.