Зрелость внедрения ИИ и зрелость информационной безопасности ИИ

Две зрелости – внедрения ИИ и его защиты – почти никогда не совпадают, и именно зазор между ними даёт большинство инцидентов и претензий регулятора.

В статье приведём обе оси к единой шкале CMMI, покажем, как организации проходят узнаваемые состояния (от отрицания до бесконтрольных агентов), и свяжем всё это с Приказом ФСТЭК № 117 и российскими стандартами.

1. Зачем измерять две зрелости

Любая организация живёт сразу в двух измерениях ИИ. Одно – насколько глубоко ИИ встроен в её процессы и продукты. Другое – насколько она вообще управляет тем, что с этим ИИ происходит с точки зрения безопасности. По опыту проектов эти измерения почти никогда не совпадают, и большую часть инцидентов и претензий регулятора даёт именно зазор между ними, а не сам факт использования ИИ.

Сложность в том, что отрасль описывает эти измерения на разных языках. Бизнес меряет внедрение по моделям вроде Gartner, безопасники – по OWASP AIMA или SAIF, регулятор оперирует пунктами Приказа. Свести их напрямую нельзя. Поэтому мы приводим обе оси к одной шкале – CMMI, пятиуровневой модели зрелости процессов, из которой в итоге выросли и SAMM, и построенная на нём AIMA. Дальше и внедрение, и безопасность ИИ описываются уровнями L1–L5, и разрыв между ними становится виден буквально как разница в номерах.

> **Суть в одной фразе:**
> **Зрелость защиты ИИ должна догонять зрелость его внедрения, а лучше – слегка опережать**. Организация на L3 по использованию ИИ и на L1 по его защите – это не «почти всё хорошо», а открытая дверь: и для утечек через теневые сервисы, и для прямого несоответствия Приказу № 117.

2. Единая шкала зрелости (CMMI)

CMMI описывает не продукт и не технологию, а зрелость процесса – насколько предсказуемо и управляемо организация что-то делает. Пять уровней читаются одинаково для любой дисциплины, поэтому модель и удобна как общий знаменатель для двух осей.

2.1. Пять уровней CMMI

- **L1, Начальный.** Процессы непредсказуемы и реактивны, всё держится на отдельных людях.

- **L2, Управляемый.** Процессы выстроены на уровне отдельных проектов, но не унифицированы по организации.
- **L3, Определённый.** Появляются единые для организации стандарты и процессы, работа становится проактивной.
- **L4, Количественно управляемый.** Процессы измеряются и управляются по метрикам.
- **L5, Оптимизирующий.** Непрерывное улучшение на основе данных и обратной связи.

2.2. Сопоставление осей на шкале CMMI

Ниже обе оси и шкала OWASP AIMA приведены к общим уровням. AIMA пользуется компактной трёхуровневой шкалой, которая ложится в верхнюю часть CMMI (L2–L5); всё, что ниже, – это уже отсутствие управляемой практики.

Уровень CMMI	Внедрение ИИ	Зрелость ИБ для ИИ	OWASP AIMA
L1 · Начальный	Осведомлённость: разговоры без стратегии	Стихийная: ИИ вне контроля, Shadow AI	ниже шкалы AIMA
L2 · Управляемый	Пилоты и PoC	Базовый контроль на уровне проектов	уровень 1
L3 · Определённый	Промышленная эксплуатация	Единые runtime-меры, моделирование угроз	уровень 2
L4 · Кол-во управляемый	Системное использование	Метрики, мониторинг, агентный контур	уровень 2–3
L5 · Оптимизирующий	ИИ в «ДНК» бизнеса	Адаптивная, самообучающаяся защита	уровень 3

Рис. 1. Лестница зрелости: внедрение ИИ и ИБ для ИИ на шкале CMMI с типичной траекторией организации.

3. Внедрение ИИ на шкале CMMI

Внедрение ИИ принято описывать пятиуровневой шкалой Gartner: осведомлённость, пилоты, эксплуатация, системное использование, трансформация. Содержательно она ложится на CMMI почти один в один, поэтому отдельную нумерацию мы не вводим – пользуемся уровнями L1–L5. Процессным якорем здесь служит ISO/IEC 42001:2023 (система менеджмента ИИ): это не шкала зрелости, но удобный источник ролей, политик и управления жизненным циклом.

Самая показательная точка – переход с L2 на L3. Именно здесь пилот превращается в боевую систему, и вместе с ним «включаются» обязательные требования к ИИ. По данным Gartner, на этом переходе застревает большинство организаций; добавим от себя, что ровно здесь чаще всего и расходятся две оси: использование уже промышленное, а защита всё ещё пилотная.

4. Безопасность ИИ на шкале CMMI

С безопасностью ИИ ситуация обратная: фреймворков много, а настоящая лестница зрелости среди них одна – OWASP AIMA (версия 1.0 вышла в августе 2025 года), наследник SAMM. AIMA раскладывает защиту ИИ на восемь доменов: Responsible AI, Governance, Data Management, Privacy, Design, Implementation, Verification, Operations – и оценивает каждый по трём уровням. Этих трёх уровней хватает, чтобы занять верхнюю часть CMMI.

Остальные рамки – не лестницы, а наполнение для неё.

Google SAIF даёт риск-карту по четырём областям (Data, Infrastructure, Model, Application) и инструмент самооценки; версия SAIF 2.0 добавляет то, без чего сегодня уже нельзя, – безопасность агентов: разрешения на инструменты, отравление памяти, многоагентные сценарии. NIST AI RMF задаёт четыре функции управления рисками (Govern, Map, Measure, Manage).

Слой проверки закрывают каталоги атак MITRE ATLAS и OWASP Top-10 для LLM-приложений. Управленческий контур – ISO/IEC 42001 и 23894, матрица контролей CSA AICM и открытая риск-карта CoSAI-RM, в которую Google передал данные SAIF.

Фреймворк	Назначение	Структура
OWASP AIMA (2025)	Единственная полноценная модель зрелости ИБ/доверия ИИ	8 доменов × 3 уровня; основа – OWASP SAMM
Google SAIF / SAIF 2.0	Риск-карта и самооценка; v2 – агентные системы	4 области (Data, Infrastructure, Model, Application) + 6 принципов
NIST AI RMF (+ GenAI Profile)	Риск-менеджмент жизненного цикла	Govern · Map · Measure · Manage
CoSAI-RM / CSA AICM	Открытая риск-карта и матрица контролей	Каталоги рисков и контролей
MITRE ATLAS / OWASP LLM Top-10	Каталоги атак и угроз (слой проверки)	Тактики, техники, типовые уязвимости
ISO/IEC 42001 · 23894	Менеджмент ИИ и управление рисками	Система менеджмента (не шкала зрелости)

5. Эволюция проблемы на практике

Шкала CMMI описывает, как должно быть. На практике организации проходят довольно предсказуемую цепочку болезненных состояний, и узнавать каждое из них в лицо полезнее, чем знать любую теорию.

5.1. Отрицание: «у нас ИИ нет»

Самый обманчивый ответ. Политики нет, инвентаризации нет – потому что «нечего инвентаризировать». А по факту сотрудники уже носят рабочие данные в публичные чат-боты и копайлоты. Угроза здесь не столько техническая, сколько управленческая: организация не видит собственного периметра. Утечка по 152-ФЗ или режиму ограниченного доступа уже возможна – просто её некому заметить. На шкале это твёрдый L1.

5.2. Неподконтрольный Shadow AI

Отрицание сменяется признанием, но не контролем. ИИ обнаруживается повсюду – в SaaS-сервисах, плагинах, ассистентах, – а единых правил, журналирования и DLP под него по-прежнему нет. Это худшее из устойчивых состояний: пользы уже много, видимости почти никакой. Большинство будущих нарушений пунктов 60–61 Приказа №117 родом именно отсюда. По внедрению организация уже на L2, по защите – всё ещё на L1.

5.3. LLM под контролем, агенты – вне контроля

Дальше наводят порядок в очевидном – в чат-интерфейсах и LLM. Появляется единый шлюз, фильтрация запросов и ответов, guardrails, проверки достоверности. По LLM это честный L3. Но пока строился этот контур, рядом вырос новый – агенты. Они не просто отвечают, а действуют: вызывают инструменты, ходят по API, через MCP подключаются к внутренним системам, общаются между собой. Меры, заточенные под «вопрос-ответ», их не покрывают.

Возникает характерный перекос: по LLM зрелость L3, по агентам – снова L1. Это сегодняшняя передовая большинства зрелых команд, и именно сюда смотрит SAIF 2.0. Перекос опасен тем, что субъективно ощущается как «у нас всё под контролем» – ведь самый заметный, чат-интерфейсный, риск действительно закрыт.

5.4. Управляемый агентный контур

Закрытие этого перекоса и есть переход на L4: разрешения инструментов по принципу минимума прав, изоляция и идентичность агентов, мониторинг их действий, метрики.

L5 добавляет то, что превращает защиту в живой процесс, – регулярный red-teaming и обновление мер под новые классы атак быстрее, чем те становятся массовыми. До этого уровня сегодня доходят единицы, и это нормально: важно осознанно двигаться, а не имитировать L5 на фоне неуправляемого L1 в соседнем контуре.

6. Российский регуляторный слой

Зарубежные модели молчат о главном для российского заказчика – об обязательных требованиях. А они за 2025–2026 годы изменились сильно, и игнорировать их при оценке зрелости нельзя: этот слой задаёт не лестницу, а планку, ниже которой опускаться запрещено.

6.1. Приказ ФСТЭК России № 117

Приказ ФСТЭК № 117 (от 11 апреля 2025 года) заменил Приказ № 17 и с 1 марта 2026 года действует в полную силу. Впервые требования к ИИ закреплены нормативно – в пунктах 60 и 61. Важная деталь: ИИ-систему регулятор не выделяет в отдельный тип объекта. Это обычная информационная система по 149-ФЗ, внутри которой работает ИИ; требования действуют и на собственную разработку, и на чужой сервис, вызываемый по API.

Что конкретно требуется:

- использовать только доверенные технологии ИИ (по подп. «ц» п. 5 Национальной стратегии развития ИИ);
- не передавать информацию ограниченного доступа разработчику модели; не применять облачные ИИ-сервисы для гостайны и сведений ограниченного доступа;
- контролировать наборы данных; задавать допустимые форматы запросов и ответов (правила для шаблонных операций и допустимую тематику для свободного текста);
- использовать статистические критерии достоверности ответов и ограничивать решения на их основе; запрещать нерегламентированное влияние ИИ на собственные параметры модели и на работу ИС.

Сопутствующее: метрики защищённости КЗИ/ПЗИ с отчётностью раз в полгода, разработка по ГОСТ Р 56939-2024, мониторинг по ГОСТ Р 59547-2021 в связке с ГосСОПКА и почти кадровое требование – не меньше 30% подразделения ИБ с профильным образованием.

> **Где срабатывает требование**
>
> Львиная доля требований к ИИ включается не при разработке, а при переводе системы в боевую эксплуатацию (для ГИС – обязательно с 1 марта 2026 года). То есть **переход с L2 на L3 не просто технический: он переводит пункты 60–61 из «когда-нибудь» в «прямо сейчас»**.

6.2. Доверие и безопасная разработка

Базовый якорь доверия – ГОСТ Р 59276-2020 «Способы обеспечения доверия», на который опираются и методические рекомендации Банка России по доверенному ИИ на финрынке. Безопасную разработку задаёт ГОСТ Р 56939-2024. Отдельный стандарт ФСТЭК по безопасной разработке ИИ (MLSecOps) пока готовится.

6.3. ИИ в критической инфраструктуре

1 июля 2025 года ТК 164 опубликовал проект ГОСТ Р «Искусственный интеллект в критической информационной инфраструктуре. Общие положения» (разработчик – ФСТЭК) – первый комплексный документ по безопасности ИИ-систем в ключевых отраслях.

30 января 2026 года Росстандарт выпустил приказ № 2-ПНСТ, где утвердил предварительный национальный стандарт (ПНСТ) 1046-2026 с тем же названием: «Искусственный интеллект в критической информационной инфраструктуре. Общие положения». То есть проект трансформировался в официальный нормативный документ, просто в формате ПНСТ, а не сразу в ГОСТ Р.

Документ распространяется на системы ИИ, которые используют в составе программно-аппаратных комплексов объектов КИИ. Он задаёт общие положения для всего жизненного цикла таких систем. Документ гармонизирован с ISO/IEC 27001, 22989, 42001 и 187-ФЗ и разделяет два жизненных цикла ИИ-системы. Для значимых объектов КИИ это отдельный, более строгий контур.

6.4. Сводка регуляторных якорей

Документ	Что регулирует / ключевые положения для ИИ
Приказ ФСТЭК № 117 (2025, в силе с 01.03.2026)	Пп. 60–61: доверенные технологии ИИ, контроль данных, форматы взаимодействия, статкритерии достоверности, запрет передачи ИОД разработчику; метрики КЗИ/ПЗИ.
Методический документ ФСТЭК (2025/2026)	18 групп технических мер; защита информации при использовании ИИ; два жизненных цикла ИИ-системы.
ГОСТ Р 59276-2020	Способы обеспечения доверия к системам ИИ; базовый якорь (используется и Банком России).
ГОСТ Р 56939-2024	Безопасная разработка ПО (РБПО) – обязательна при собственной разработке.
ГОСТ Р 59547-2021	Мониторинг ИБ; основа требований к мониторингу и связке с ГосСОПКА.
Проект ГОСТ «ИИ в КИИ» (ТК 164, 2025) или ПНСТ 1046-2026	Безопасность ИИ на объектах КИИ; гармонизация с ISO 27001/22989/42001 и 187-ФЗ.
БДУ ФСТЭК	Каталог угроз, специфичных для ИИ (обход средств защиты, модификация и отказ модели, манипуляция поведением).
Нацстратегия развития ИИ	Понятие доверенных технологий ИИ (подп. «ц» п. 5), на которое ссылается № 117.

Чего пока нет в требованиях

Качество самих моделей – объяснимость решений и борьбу с галлюцинациями – регулятор сознательно выносит за скобки: контролируется лишь достоверность ответов, и то статистически. Это и осторожность, и подсказка: зрелые команды закрывают этот участок сами, не дожидаясь отдельного норматива.

7. Разрыв между осями

Сложим разделы 2 и 6 – и сразу виден системный пробел. Модели внедрения меряют бизнес и молчат об ИБ. Модели ИБ меряют защиту, но не привязаны ни к уровню внедрения, ни к российским требованиям. Регуляторика задаёт планку, но не лестницу. В итоге никто не отвечает на главный практический вопрос: **если по использованию ИИ мы на L3, на каком уровне обязана быть наша защита?**

Ответ: на том же. Опасна именно диагональ: высокое внедрение при низкой ИБ. Организация на L3–L4 по использованию и на L1–L2 по защите одновременно копит неуправляемый риск и нарушает пункты 60–61 Приказа №117. Все три состояния из раздела 5 – отрицание, Shadow AI, неподконтрольные агенты – это и есть точки на этой диагонали.

8. Модель сопряжения: внедрение → ИБ → требования РФ

Модель связывает на одной шкале CMMI три вещи: уровень внедрения, минимально необходимый уровень зрелости ИБ для ИИ и конкретные якоря российского комплаенса. Правило простое: достигнутый уровень внедрения задаёт нижнюю планку для защиты на том же уровне.

8.1. Матрица сопряжения

Уровень (CMMI / внедрение)	Что требуется от ИБ для ИИ	Якоря РФ-комплаенса
L1 · Начальный **Осведомлённость**	– Инвентаризация ИИ-активов и сценариев – Политика допустимого использования – Осведомлённость о рисках (OWASP LLM Top-10)	– Пока в ИС нет работающего ИИ – вне обязательной части № 117 – Главный риск – утечки через теневые сервисы (152-ФЗ, ИОД)
L2 · Управляемый **Пилоты**	– Моделирование угроз пилота (домен Design AIMA) – Контроль обучающих и тестовых данных (SAIF: Data) – Первые guardrails; РБПО на этапе разработки	– ГОСТ Р 56939-2024 при собственной разработке – При выводе пилота в эксплуатацию включаются пп. 60–61
L3 · Определённый **Эксплуатация**	– Единый шлюз и runtime-защита LLM: фильтрация запросов/ответов, контроль форматов – Статистические критерии достоверности – Адверсариальное тестирование (MITRE ATLAS, домен Verification)	– Прямое соответствие пп. 60–61 Приказа № 117 – Только доверенные технологии ИИ – Запрет передачи ИОД разработчику; запрет облачных ИИ для гостайны/ИОД
L4 · Кол-во управляемый **Системный**	– Централизованный AI-governance – Непрерывный мониторинг ввода/вывода, интеграция в SOC – Контроль агентного контура (минимум прав, изоляция, SAIF 2.0); метрики, MLSecOps	– Метрики КЗИ/ПЗИ, отчётность во ФСТЭК (раз в 6 мес.) – Мониторинг по ГОСТ Р 59547-2021; ГосСОПКА – ≥30% персонала ИБ с профильным образованием – Для КИИ – проект ГОСТ «ИИ в КИИ» (187-ФЗ)
L5 · Оптимизирующий **Трансформация**	– Адаптивная защита и регулярный red-teaming – Зрелая защита агентных и многоагентных систем – Полная интеграция ИБ в жизненный цикл ИИ	– Готовность к стандарту ФСТЭК по безопасной разработке ИИ (MLSecOps) – Готовность к будущим требованиям по доверию и объяснимости (ГОСТ Р 59276)

Сокращения: ИОД – информация ограниченного доступа; РБПО – разработка безопасного ПО; КЗИ/ПЗИ – показатели защищённости информации.

8.2. Как читать матрицу

Матрицу читают в обе стороны. Сверху вниз – как обязательство: дошли по внедрению до L3, значит и защита обязана быть L3. Снизу вверх – как тормоз: не тяните внедрение выше того уровня, который реально закрываете по ИБ. Любой разрыв – это либо риск, либо несоответствие, а чаще и то и другое сразу.

9. Как применять модель

Применяется в три шага. Сначала честно определяем уровень по внедрению. Затем – по защите, по доменам AIMA, наполненным контролями SAIF и MITRE и сверенным с требованиями ФСТЭК. После этого смотрим разрыв и строим дорожную карту от ближайшего обязательного уровня, а не от идеала.

Где здесь технические средства. Большая часть обязательных мер уровня L3 – это runtime-слой: единый шлюз, фильтрация запросов и ответов, контроль форматов, проверки достоверности, guardrails. Его и закрывают специализированные средства защиты ИИ-приложений (LLM Firewall), что делает их прямым способом дотянуть защиту до L3 и выполнить пункты 60–61. На L4–L5 к ним добавляются непрерывный мониторинг, контроль агентного контура и MLSecOps.

Привязка к продуктам

В терминах этой шкалы средства класса AI.Firewall закрывают runtime-уровень, обязательный начиная с L3, а инструменты безопасной разработки – слой РБПО, нужный уже на пилотах (L2). Оценка зрелости становится точкой входа в разговор с заказчиком, а конкретные продукты – понятным маршрутом закрытия разрыва.

10. Выводы

- Внедрение ИИ и его защита – две разные зрелости. Мерить нужно обе и смотреть на разрыв между ними, а не на каждую по отдельности.
- CMMI даёт общий язык: и Gartner-уровни внедрения, и домены AIMA, и требования ФСТЭК ложатся на одну шкалу L1–L5.
- На практике организации проходят узнаваемые состояния – отрицание, неподконтрольный Shadow AI, контроль LLM при бесконтрольных агентах. Каждое из них – точка на опасной диагонали.
- Российские требования (Приказ № 117, ГОСТы, проект по КИИ) задают обязательный минимум и привязывают нагрузку к моменту перевода ИИ в эксплуатацию (переход L2 → L3).
- Runtime-средства защиты – прямой способ дотянуть ИБ до уровня внедрения начиная с L3. Агентный контур – следующий рубеж, L4 и выше.

Источники

Шкала и международные рамки

- CMMI (модель зрелости процессов; ISO/IEC 33001 и CMMI Institute) – общая шкала уровней L1–L5.
- OWASP AI Maturity Assessment (AIMA), v1.0, август 2025; OWASP SAMM; OWASP Top-10 для LLM-приложений.
- Google Secure AI Framework (SAIF) и SAIF 2.0; CoSAI Risk Map (CoSAI-RM).
- NIST AI Risk Management Framework и профиль для генеративного ИИ; MITRE ATLAS; CSA AI Controls Matrix (AICM).
- ISO/IEC 42001:2023; ISO/IEC 23894; ISO/IEC 22989. Модель зрелости внедрения ИИ (пятиуровневая, Gartner); модель MIT CISR.

Российские документы

- Приказ ФСТЭК России № 117 от 11.04.2025 (в силе с 01.03.2026) и сопровождающий методический документ.
- ГОСТ Р 59276-2020; ГОСТ Р 56939-2024; ГОСТ Р 59547-2021.
- Проект ГОСТ Р «Искусственный интеллект в критической информационной инфраструктуре. Общие положения» (ТК 164, 01.07.2025).
- БДУ ФСТЭК; Национальная стратегия развития ИИ; 187-ФЗ; 152-ФЗ; методические рекомендации Банка России по доверенному ИИ.

Модели зрелости Gartner и ряд стандартов ISO распространяются на условиях правообладателей; здесь используются их общедоступные структурные описания.