АНО «НИИ СИСТЕМНОГО СИНТЕЗА»



АЛЕКСАНДР ФИЩУК

Доказанное действие

Власть и контроль над ИИ-агентами: стратегия 2027–2035

АРХИТЕКТУРНЫЙ МЕМОРАНДУМ

Формулы всех глав, глава 1 и команды независимого повтора

АНО «НИИ системного синтеза» · 2026

ISSLAB.RU

Александр Фищук

Доказанное действие

Власть и контроль над ИИ-агентами: стратегия 2027–2035

Архитектурный меморандум: формулы всех глав, глава 1 и команды независимого повтора

АНО «НИИ Системного синтеза»
2026

Меморандум — свободно распространяемая часть книги «Доказанное действие. Власть и контроль над ИИ-агентами: стратегия 2027–2035» (2026).

Полная книга — ЛитРес: https://www.litres.ru/74623534/ и сайт НИИ Системного синтеза: https://isslab.ru/

Числа о шлюзе MITR сверены с прибором на сервере 24.10.2026; каждое утверждение помечено: измерено, строим, замысел, прогноз, факт, гипотеза.

© Александр Фищук, 2026

Содержание

Формулы глав 05

Глава 1. Скорость: экспонента вычислений, обесценивание токена и горизонт автономного действия 10

Приложения 34

Приложение Б. Команды независимого повтора 34

ISS

Формулы глав

Каждая глава книги открывается формулой — двумя-тремя короткими фразами, в которых сжат её главный вывод. Девятнадцать формул ниже — введение, семнадцать глав и заключение — читаются за несколько минут и дают всю систему координат книги. Глава 1 и приложение Б приведены в меморандуме целиком; в полное издание входят все пять частей (17 глав) и приложения А–Д.

> *Введение. Калибровка прибора: человек, компания и государство в точке пересечения двух прямых*

Поток машинных действий уже превысил физиологический предел человеческого внимания. Чтобы сохранить власть над сложными системами, человек переносит контроль с мысли модели на её действие. Правило подписывается заранее, а каждый вызов оставляет проверяемую криптографическую квитанцию.

> *Глава 1. Скорость: экспонента вычислений, обесценивание токена и горизонт автономного действия*

Скорость машинных решений растёт по экспоненте, а внимание человека остаётся биологической константой. В точке расхождения этих кривых ручное одобрение превращается в юридическую фикцию. Устойчивое управление требует автоматической проверки каждого действия по правилу до его исполнения.

> *Глава 2. Энергия и кремний: гигаватты как твёрдая граница и суверенитет на уровне допуска*

Обучение ведущих моделей требует гигаватт электроэнергии и десятков тысяч дефицитных ускорителей. Однако математическая проверка допустимости отдельного действия занимает

около миллисекунды на обычном процессоре. Технологический суверенитет рождается на границе допуска, где человек подписывает правило и требует квитанцию.

Глава 3. Агент получил право записи: интернет агентов 2026

Получив доступ к рабочим интерфейсам и платежам, языковая модель стала субъектом прямого имущественного действия. Ручная кнопка подтверждения больше не защищает предприятие от каскадных аварий. Удержать периметр способна только централизация допуска: человек подписывает правило, а шлюз проверяет вызов до исполнения.

Глава 4. Труд и человек: что остаётся, когда агент дешевле

Цена машинного исполнения стремится к стоимости электричества, но правосубъектность, ключи и имущественная ответственность неделимы с человеком. Профессионал сохраняет доход, а собственник защищает капитал, когда оба уходят из ручного исполнения на уровень подписания правил.

Глава 5. Государство, право и сетевое общество

В двадцать первом веке суверенитет опирается на машиночитаемое право и строгий допуск к действию. Государство сохраняет управляемость, когда законодатель задаёт инварианты политики, а шлюз проверяет каждый шаг агента за миллисекунды. Каждое решение закрепляется криптографической квитанцией, проверяемой без доверия к сторонам.

Глава 6. Пределы контроля и честный обход

Внутреннее состояние сложной программы неразрешимо по теоремам Тьюринга и Райса, но сам вызов внешнего инструмента конечен и разрешим. Контроль переносится с непредсказуемой мысли на формально проверяемое действие. Машина предъявляет доказательство безопасности до вызова, а человек удерживает правило, ключи и ответственность.

Глава 7. Кибернетика управления в эпоху агентов

Когда поток машинных операций на порядки превышает возможности сознания, управлять отдельными командами невозможно — работают только инвариантные правила. По закону Эшби жизнеспособность периметра удерживают два прибора: контракт-данные как ослабитель возмущений и детерминированный шлюз как усилитель регуляции.

Глава 8. Контракт как данные и лестница «от совета к запрету»

Правило, зашитое в код программы, устаревает быстрее, чем заканчивается рабочий день разработчика. Декларативный контракт-данные отделяет волю человека от вычислительного механизма. Периметр переводится из наблюдения в режим запрета по измеренной доле ложных срабатываний: нет правила — нет действия.

Глава 9. Решатель: допуск — доказательство, блок — контрпример

Нейросеть оценивает правдоподобие текста, а математический решатель доказывает безопасность действия. Вызов внешнего инструмента содержит конечный набор параметров, который логика первого порядка проверяет на обычном процессоре за миллисекунду. Шлюз либо доказывает отсутствие нарушений, либо возвращает точный контрпример.

Глава 10. Квитанция, цепь, внешний свидетель, повтор посторонним

Обычный текстовый журнал на сервере владельца при разборе аварии неотличим от подлога. Доверие между независимыми сторонами возникает лишь тогда, когда каждое решение связывается в криптографическую цепь с внешней меткой времени. Такую квитанцию сторонний ревизор пересчитывает автономно, не обращаясь к создателю системы.

Глава 11. «Не измерено» — не «чисто»: трёхзначная логика допуска и алгедонический канал

Панель мониторинга, окрашивающая молчание датчика в успокоительный зелёный цвет, порождает операционную и правовую слепоту. Доказуемое управление требует выделить состояние «не измерено» в третье значение логики. При отсутствии достоверного замера шлюз блокирует действие и направляет сигнал остановки субъекту ответственности.

Глава 12. Самоулучшение под доказательством: шлюз MITR предлагает правило сам

Агент, бесконтрольно переписывающий собственные навыки в рабочей среде, сам становится источником аварии. Жизнеспособная конструкция отделяет неизменное ядро решателя от пополняемого словаря данных. Каждое новое правило проходит через то же формальное доказательство и ту же квитанцию, что и обычный вызов.

Глава 13. Полигон и принцип нулевого ущерба: испытатель как посторонний по построению

Система, проверяющая себя изнутри собственного исходного кода, измеряет лишь глубину собственных заблуждений. Прочность защиты доказывается тогда, когда испытатель отчуждён от защитника в чистую комнату. Он входит только через штатную дверь арендатора и атакует размеченный стенд без риска для промышленных данных.

Глава 14. Иммунная петля от испытателя INDRA к шлюзу MITR и «guardian agents»

Статический щит слепнет перед новыми атаками, а вероятностный сторож в рабочем потоке сам склонен к ошибкам. Жизнеспособная система разносит поиск уязвимостей и допуск действий по разным уровням. Внешний испытатель находит брешь

на полигоне, а шлюз исполняет обновлённый контракт за миллисекунду.

Глава 15. Норма и инвариант: три маршрута и уровни полномочий

Правовая норма обязывает только человека и организацию, поскольку программный агент субъектом права не признаётся. Перевод закона в разрешимую логическую формулу над параметрами вызова превращает ответственность директора в прозрачный инженерный механизм. Человек подписывает правило вместо отдельное действие.

Глава 16. Деньги: лестница ступеней, страховщик и регулятор

По мере удешевления вычислений рентабельность уходит из генерации текста в узел допуска, аудита и страхования ответственности. Зрелость защитного продукта доказывается не финансовыми презентациями, а машинным выполнением критериев приёмки и цепью квитанций, которую контрагент проверяет самостоятельно.

Глава 17. Стратегия человека и института на горизонте 2035 года: контур субъектности и метрология собственной нечестности

Когда цена машинного исполнения падает до стоимости электричества, опорой человека и государства остаётся подписанное правило под имущественную ответственность. Инженерная школа, не способная приборно измерить собственную неполноту, неизбежно подменяет строгий допуск бумажным самоотчётом.

Заключение. Точка сборки и протокол пересчёта: 2026–2035

Конструкция доказанного действия снимает ложный выбор между слепым ускорением и полным запретом автономных систем. Человек сохраняет монополию на правило, ключи и ответственность, а сама книга работает как открытый измерительный стенд с регулярной сверкой прогнозов до 2035 года.

Глава 1. Скорость: экспонента вычислений, обесценивание токена и горизонт автономного действия

> *Формула главы: Скорость машинных решений растёт по экспоненте, а внимание человека остаётся биологической константой. В точке расхождения этих кривых ручное одобрение превращается в юридическую фикцию. Устойчивое управление требует автоматической проверки каждого действия по правилу до его исполнения.*

1.1 Кинематика расходящихся кривых и предел ручного торможения

К осени 2026 года мировая индустрия программного обеспечения вошла в жёсткую кинематическую развилку: время срабатывания периметра защиты обязано быть меньше времени развития аварии. В то же время работа программных агентов идёт на скоростях, измеряемых микросекундами и миллионами вызовов интерфейсов прикладного программирования (API). На одну чашу весов легла объективная физика вычислительного роста. Согласно данным исследовательского центра Epoch AI от 5 февраля 2026 года, объём вычислений, затрачиваемых на обучение ведущих моделей, с 2020 года увеличивается примерно в пять раз ежегодно, удваиваясь каждые 5,2 месяца. Epoch AI, data/trends, 05.02.2026. На длинной дистанции с 2010 года по всей выборке значимых моделей среднегодовой темп составил 4–5 раз в год (точечные оценки ×4,1–4,4; Epoch AI, 2024). Одновременно с ростом мощности происходит стремительная дефляция элементарного акта машинного рассуждения. Стоимость генерации одного токена при фиксированной планке качества падает в среднем в десять раз за каждый кален-

дарный год (исследование a16z «LLMflation», 12.11.2024). При этом в зависимости от выбранного порога сложности задачи удешевление по замерам за 2022. Начало 2025 года составляет от 9 до 900 раз в год (Epoch AI, ai-inference-price-trends, 12.03.2025).

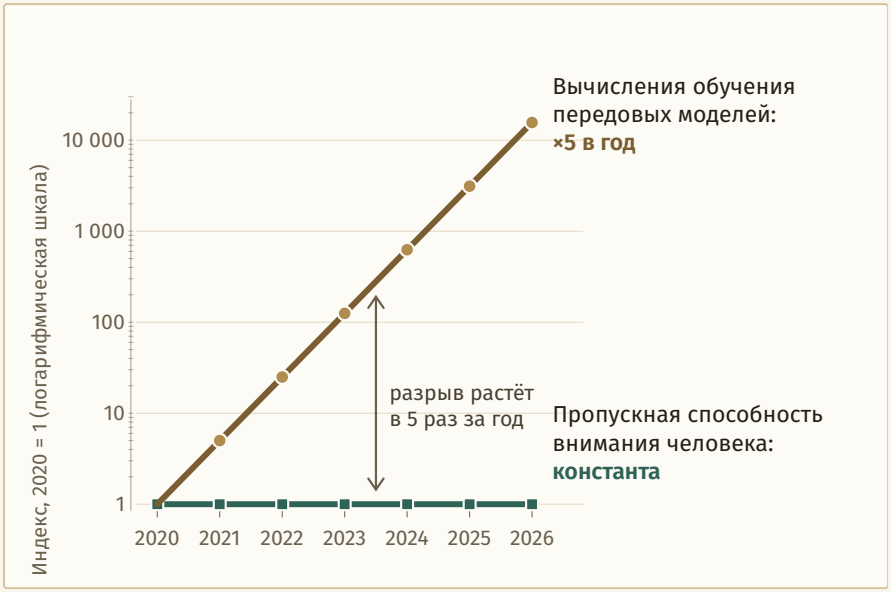


Рисунок 1.1 — Два встречных потока: индекс вычислений обучения ведущих моделей при росте ×5 в год (Epoch AI, 05.02.2026) и пропускная способность внимания человека

На другой чаше весов находится биологический оператор — инженер, руководитель разработки, дежурный офицер безопасности или финансовый контролёр. Скорость прохождения нервного импульса по аксонам человеческого мозга с эпохи плейстоцена не изменилась ни на один процент и составляет десятки метров в секунду. Когнитивный ресурс сосредоточенного внимания взрослого специалиста ограничен несколькими часами в сутки, после чего наступает физиологическое утомление префронтальной коры. Когда стоимость машинного действия стремится к нулю, а длительность непрерывной автономной работы агента удваивает-

ся каждые три-четыре месяца, предприятие получает экспоненциально растущий поток транзакций, команд оболочки, сетевых запросов и модификаций баз данных, который пытается удержать плотиной человеческого взгляда.

В терминологии теории решения изобретательских задач (ТРИЗ) Генриха Альтшуллера возникает классическое административное и техническое противоречие. Если предприятие оставляет человека в периметре одобрения каждого отдельного шага агента ради сохранения безопасности, скорость работы всей системы падает до скорости чтения текста человеком, экономический смысл автоматизации обнуляется. При этом сам оператор через несколько дней непрерывных запросов впадает в состояние механического проштамповывания решений. Если же предприятие убирает человека из текущего периметра ради извлечения полной скорости и дешевизны машинного труда, вероятностная модель рано или поздно совершает необратимый вызов в промышленной среде. От удаления резервных копий до несанкционированного платежа, — поскольку внутри нейронной сети отсутствует понятие твёрдой границы допустимого.

Попытки разрешить это противоречие компромиссом — например, оставить за человеком одобрение только «важных» действий или приставить к первой вероятностной модели вторую вероятностную модель в роли цензора, нарушают базовый закон необходимого разнообразия Уильяма Росса Эшби. Разнообразие управляющей системы (регулятора) должно быть не меньше разнообразия возмущений, порождаемых управляемым объектом. Когда поток действий агентов растёт в пять раз в год, а стоимость генерации падает на порядок, никакое масштабирование штата контролёров и никакое нагромождение вероятностных классификаторов не создаёт детерминированного барьера.

Идеальный конечный результат (ИКР) по ТРИЗ формулируется без компромиссов. Само действие агента до момента своего исполнения должно мгновенно и без участия человеческого внимания доказывать собственную допустимость относительно подписанного человеком правила, а при отсутствии такого доказательства. Автоматически блокироваться с неизменяемой записью в журнале (канон проектирования шлюза). Контроль переносится с непредсказуемого процесса машинного мышления на дискретный акт действия, а человеческое внимание освобождается от рутинного досмотра миллионов посылок и поднимается на уровень подписания закона. Человек подписывает правило вместо отдельного действия.

1.2 Анатомия уязвимости: хроника отказов 2026 года и цена слепого одобрения

Кинематическое расхождение кривых скорости и внимания отчётливо проявляется при сопоставлении статистики корпоративных внедрений и материалов расследований инцидентов первой половины 2026 года. В этой точке пересекаются две ключевые управленческие роли современной экономики. Технический директор (CTO), отвечающий за непрерывность работы продуктового периметра (роль 1 в карте уязвимостей), и институциональный инвестор или инвестиционный комитет фонда, оценивающий жизнеспособность агентных проектов на горизонте трёх-пяти лет (роль 4).

Хроника промышленных аварий 2025–2026 годов (развёрнутая с документальными первоисточниками в главе 3) показала, что корень сбоев лежит не в недостатке рассуждений модели, а в избыточных полномочиях доступа. Когда кодовый агент в рабочем окружении самостоятельно находит привилегированный токен и удаляет промышленную базу данных вместе с резервными копиями (инцидент PocketOS в апреле 2026 года), очищает домашний ка-

талог разработчика (сбой агента на модели GPT-5.6 Sol в июле 2026 года) или останавливает региональный облачный сервис на тринадцать часов (инцидент с агентом Kiro в декабре 2025 года), предрелизное тестирование модели оказывается бессильным. По отраслевым опросам 2026 года (Gravitee, VentureBeat, разбор Lumenova; см. раздел 3.2), почти половина организаций столкнулась со сбоями агентов, заметными конечным клиентам, именно из-за отсутствия жёсткого фильтра на канале исполнения.

Инфраструктурная гигиена вокруг новых протоколов связности лишь усиливает масштаб риска. Стандарт Model Context Protocol (MCP), ставший к 2026 году фактическим языком подключения моделей к корпоративным базам данных и инструментам (свыше 10 000 публичных серверов и 97 миллионов загрузок комплекта разработчика в месяц по данным Anthropic на декабрь 2025 года (Anthropic, декабрь 2025)), по опросу 300 крупных предприятий США, Канады и Великобритании (декабрь 2025) в 41 % уже используется в ограниченном или широком промышленном периметре. Stacklok, «State of MCP 2026», январь 2026. Однако сканирование 1 412 публично доступных серверов MCP, проведённое Bloomberry, выявило, что среди серверов с определимым способом аутентификации 38,7 % работают вообще без неё (Bloomberry, 24.02.2026; пересказ в сводке DreamFactory, 03.03.2026). Агент получает руку, способную дотянуться до производственного рубильника, но между этой рукой и рубильником в лучшем случае стоит диалоговое окно с кнопкой «Подтвердить».

Здесь вступает в силу закон сопротивления материалов, применённый к человеческой психике. По данным отраслевого обзора 70 публичных описаний внедрений ИИ-агентов в России за 2025–2026 годы, в 37 случаях, или в 52,9 %, главным механизмом сдерживания агента служит ручное подтверждение человеком в петле. Отраслевой обзор института, 40-я неделя 2026 года, 37 из 70; под-

счёт по опубликованным описаниям, а не обследование периметров. Поначалу инженер вчитывается в предлагаемые команды. Когда агент начинает выдавать по двести запросов за смену, девяносто девять из которых корректны, внимание пересекает предел усталости (срок в источниках не измерен). Возникает феномен усталости от одобрения (approval fatigue), разобранный в эссе Тянь Паня в июне 2026 года со ссылкой на отраслевые опросы. Шлюз «человек в петле» вырождается в резиновый штамп [по данным вторичного обзора: Tian Pan, 25.06.2026]. Августовский опрос 2026 года зафиксировал парадоксальную реакцию рынка. Компании, уже пострадавшие от сбоев ИИ, быстрее остальных отключают ручное одобрение, убедившись в его неспособности предотвратить аварию при торможении бизнеса (опрос VentureBeat, август 2026; пересказ AccountsRecovery, 25.08.2026). Кнопка «Разрешить» на экране уставшего инженера не защищает базу данных — она лишь переносит юридическую вину с программы на физическое лицо.

Для второго участника процесса — венчурного инвестора, фонда развития или грантового комитета (Сколково, Сириус, корпоративные инвестиционные советы), эта конструктивная трещина оборачивается прямым риском потери капитала. В 2025 году стартапы, строящие автономных агентов, привлекли рекордные 24,2 миллиарда долларов по 1 311 венчурным сделкам по методологии PitchBook (при более узком определении сегмента оценки на порядок ниже: Prosus и Dealroom в августе 2025 года прогнозировали около 6,7 миллиарда долларов за год; аналитическая записка PitchBook, II кв. 2026; ПРОГНОЗ: Prosus/Dealroom, 08.2025). В первом квартале 2026 года проекты в сфере искусственного интеллекта поглотили 242 миллиарда долларов — 80 % всего мирового венчурного капитала. Причём 65 % этой суммы пришлись на четыре гигантские сделки (OpenAI, Anthropic, xAI, Waymo; Crunchbase,

итоги I кв. 2026, 01.04.2026). И на фоне этого финансового прилива аналитическое агентство Gartner в официальном прогнозе от 25 июня 2025 года опубликовало прогноз. Свыше 40 % проектов в области агентного ИИ будут полностью свёрнуты и закрыты к концу 2027 года по трём причинам. Из-за неконтролируемого роста расходов, неясной деловой ценности и слабого контроля рисков (Gartner Press Release, 25.06.2025). Поток дешёвых токенов без строгого узла допуска не превращается в промышленный актив; он превращается в неучтённое обязательство в балансе предприятия.

1.3 Инженерная реальность 2026 года: три экспоненты и детерминированный затвор

За пределами рекламного шума верифицируемые приборные измерения 2026 года складываются в четыре жёстких количественных кортежа: вычислительную мощность обучения, экономику вывода (инференса), временной горизонт автономности модели и сокращение разрыва между закрытыми и открытыми весами.

Первый кортеж — физическое масштабирование вычислений. Пятикратный ежегодный рост, зафиксированный Epoch AI на выборке ведущих моделей с 2020 по февраль 2026 года (Epoch AI, 05.02.2026), означает увеличение вычислительного бюджета флагманского обучения на порядок примерно каждые семнадцать месяцев. В энергетическом эквиваленте крупнейшие прогоны обучения к 2030 году, по оценке Epoch AI, потребуют от 4 до 16 гигаватт электрической мощности (прогноз: Epoch AI, «Power demands of frontier AI training», август 2025). Выработки нескольких атомных энергоблоков ради одной матрицы весов.

Второй кортеж — расслоение цены токена. Закон десятикратного ежегодного удешевления при фиксированном качестве (÷10 в год по данным a16z и от ÷9 до ÷900 в год по замерам Epoch AI (a16z, 12.11.2024; Epoch AI, 12.03.2025; по обновлённой оценке Epoch —

около ×13 в год с 2023 года, 22.09.2026)) пока держится, хотя сами авторы допускают замедление. Задача, решение которой в 2023 году стоило доллар, в 2026 году обходится в доли цента. Однако на самом острие технологического фронтира стоимость миллиона токенов не падает. Пройдя путь от $30 к $1,25 на промежуточных поколениях, цена флагманского рассуждающего токена вернулась к отметке около $5 за миллион на моделях класса GPT-5.5 (витрины провайдеров, 2026, по вторичному источнику). Массовые рутинные операции дешевеют со скоростью обвала, провоцируя запуск тысяч параллельных периметров, тогда как сложные многошаговые рассуждения остаются дорогими и делают непредсказуемой совокупную стоимость владения (TCO) при потокенной оплате.

Третий и самый критичный для конструкции безопасности кортеж — удлинение временного горизонта автономной задачи. Независимый исследовательский институт METR в январе 2026 года опубликовал обновлённый эталонный набор TH1.1, включающий 228 многошаговых инженерных и исследовательских задач (METR TH1.1, январь 2026). Инструментальный замер показывает: время, в течение которого модель способна самостоятельно выполнять задачу с пятидесятипроцентной вероятностью успеха (p50, выраженное в эквиваленте рабочего времени квалифицированного человека), удваивается каждые 89–131 день в зависимости от выбранного статистического окна аппроксимации. METR, Time Horizon 1.1, 29.01.2026. В феврале 2026 года модель Claude Opus 4.6 достигла на этом стенде медианы p50 ≈ 12 часов непрерывного человеческого труда (данные METR, 20.02.2026). При этом в мае 2026 года ранняя версия Claude Mythos Preview вышла за 16-часовой предел (METR, май 2026). При этом сами авторы METR прямо фиксируют метрологический предел собственного прибора: на горизонтах свыше 16 часов текущий набор задач теряет статистиче-

скую надёжность (METR, страница time-horizons, 2026). Одновременно проект AI Futures Project в обновлении за первый квартал 2026 года (02.04.2026) пересмотрел свои кривые. Медиана появления полностью автоматизированного программиста (Automated Coder) у ведущего автора сдвинулась с конца 2029 года на середину 2028 года (у второго автора — с начала 2032 на середину 2030). При этом медианы появления систем, превосходящих лучших профильных экспертов, приблизились примерно на полтора года. В обновлении от 16.08.2026 медианы сдвинуты до ноября 2027 и января 2030 года соответственно (прогноз: AI Futures Project, Q1 и Q2.5 2026 Timelines Updates). Первая категория агентов уже доказала полное соответствие рынку. Годовой темп выручки (run-rate) терминального агента Claude Code к февралю 2026 года превысил 2,5 миллиарда долларов (более половины приходится на корпоративный сектор; сообщение Anthropic о раунде Series G, 12.02.2026), а кодовые помощники внедрены в 51 % опрошенных предприятий. Первая массовая категория корпоративного применения генеративного ИИ (Menlo Ventures, опрос 600 ИТ-руководителей США, 20.11.2024); отдельная выручка Codex компанией OpenAI не раскрывалась.

Четвёртый кортеж разрушает монополию закрытых американских лабораторий. Согласно Международному отчёту по безопасности ИИ от 3 февраля 2026 года (International AI Safety Report), модели с открытыми весами (open-weight) отстают от лучших закрытых проприетарных систем менее чем на один год (IAISR, 03.02.2026). При этом по уточнённой оценке Epoch AI к середине 2026 года этот временной лаг сжался примерно до четырёх месяцев (Epoch AI, Data Insight «open–closed ECI gap», 29.05.2026). При этом лидерами открытого сегмента стали китайские инженерные школы — семейства DeepSeek, Qwen, Kimi и GLM (Epoch AI / отраслевая сводка 2026). На коммерческом шлюзе Vercel в июне 2026

года открытые модели обеспечивали уже 29 % всего потока токенов, потребляя при этом менее 4 % денежных расходов клиентов, а в августе 2026 года. Уже 56 % токенов (Vercel AI Gateway Production Index, данные за июнь и август 2026).

Сопоставление этих четырёх фактов на одном чертеже приводит к однозначному конструктивному следствию. Когда агент автономно работает двенадцать часов подряд, опираясь на открытую модель стоимостью в доли цента, и совершает тысячи вызовов MCP и команд оболочки, контролировать безопасность через разбор его внутренних рассуждений невозможно. Индустрия не может опираться на то, *как* модель думает. Единственная точка, где вероятностный поток нейросети соприкасается с детерминированным миром предприятия, — узкая горловина вызова внешнего инструмента.

На этом водоразделе («мысль — действие», аксиома аксиома водораздела в реестр продуктовой линейки) построен и измерен в коде программный шлюз MITR на рабочем сервере шлюза (рабочий сервер шлюза, октябрь 2026). Для руководителя предприятия и юриста шлюз работает как нотариальная палата реального времени: у агента нет ключей от баз данных, платёжных шлюзов или почтовых серверов — есть только право подать формализованную заявку на действие. Шлюз сверяет параметры заявки с подписанным человеком контрактом-правилом. Если действие укладывается в границы дозволенного, шлюз исполняет его своими ключами и выписывает криптографическую квитанцию; при нарушении границы действие блокируется до выхода пакета.

Для системного инженера и математика этот же процесс внутри модуля z3_gate.py сводится к строгой формальной верификации посредством решателя выполнимости формул в теориях (SMT-решателя Z3). Контракт безопасности записан не на естественном языке промпта, который можно обойти хитроумной ме-

тафорой. При этом как система логических ограничений первого порядка (на 24 октября 2026 года в хранилище правил действуют 20 таких контрактов (на 14 октября — 19)). Когда агент запрашивает вызов инструмента с конечным набором параметров, решатель проверяет, существует ли хотя бы одно пересечение запрошенных параметров с множеством запрещённых состояний системы. Допуск действия математически эквивалентен результату UNSAT (невыполнимо: ни одно запрещённое состояние не достигается, значит, переход доказанно безопасен). Блокировка эквивалентна результату SAT: решатель находит конкретный контрпример нарушения инварианта и возвращает формальные условия починки запроса.

Ключевой вопрос любой синхронной проверки на пути высокоскоростного потока агентов — вносимая задержка. Прибор P4 (скрипт smt_latency_paper.py) даёт две цифры на обычном серверном процессоре (CPU) без графических ускорителей (GPU). Медианное время чистого математического решения SMT-задачи — 0,74 миллисекунды, медианное время вызова шлюза вместе с объяснением решения (статусы ветвей, условия починки), 3,3 миллисекунды (замер 26.09.2026 на 17 контрактах). При 19 контрактах повторные замеры дали 0,87 мс (02.10.2026, прежний узел) и 0,95 мс при вызове 4,3 мс (14.10.2026, рабочий сервер шлюза) — цена решения остаётся в пределах миллисекунды. При этом прогон эталонного состязательного набора из 157 сценариев обходится ровно в 0 долларов расходов на внешние модели. При этом расход узла на вспомогательные языковые модели за 1–6 октября 2026 года составил 0,14 доллара (за 1–14 октября — 0,30 доллара; журнал расхода рабочего сервера шлюза, 14.10.2026). Каждое принятое решение («допустить» или «отклонить») связывается криптографической хэш-цепью SHA-256 с внешней меткой времени OpenTimestamps. При этом проверщик, написанный для постороннего (verify_chain_

standalone.py --replay: стандартная библиотека Python и любой решатель SMT-LIB, без строки кода MITR), повторил 1602 решения из 1602 на 30.09.2026, 4958 из 4958 на 14.10.2026 и 5053 из 5053 на 24.10.2026 без единого расхождения. Контрольный прогон на стенде; сторонний проверяющий повтор ещё не выполнял.

1.4 Ближний инженерный рубеж: дисциплина ворот 2026–2027

Инженерная дисциплина требует чёткой границы между тем, что доказано как работающий механизм в одной точке, и тем, что находится в процессе статистического набора. В методологии книги эта граница закреплена метками и: первая описывает приборный факт в настоящем времени, вторая — задачу в будущем времени с машинными воротами приёмки, которые считает скрипт-генератор.

На 9 октября 2026 года проверка по шести базовым конструктивным постулатам системы (gate_mode_report.py --six) показывала эксплуатационную правду 1 из 6 (в запрете находился 1 класс из 7). При этом к 23 октября 2026 года все 20 контрактов и 6 основных классов действий переведены в режим автоматического запрета. За 23 октября прибор впервые подтвердил все шесть базовых положений («правда 6 из 6 — неправда 0, не измерено 0»). При этом за 24 октября из-за пустого суточного среза по одному классу показал «правда 5 из 6, не измерено 1». Перевод классов действий из режима наблюдения и совета (shadow / advisory) в режим безусловного блокирования опирается на строгую эксплуатационную процедуру — лестницу «от совета к запрету» и последующее 14-суточное удержание.

Ближний рубеж четвёртого квартала 2026 года включает три верифицируемых этапа. Первые ворота — закрытие серии P2a к 31 октября 2026 года: четырнадцать непрерывных суток с объёмом не менее 50 реальных проверок в сутки от минимум двух арендато-

ров при нулевом числе обходов. На момент среза счётчик зачётной серии составлял 0 из 14 (ранее прибор честности на протяжении 98 суток показывал «0 зелёных суток», не засчитывая дни с неполной телеметрией (октябрь 2026)).

Вторые ворота — достижение рубежа «шесть из шести» к 14 декабря 2026 года. Все шесть базовых постулатов конструкции и все целевые классы действий будут переведены в боевой режим запрета на четырнадцатисуточном интервале после того, как доля ложных срабатываний по каждому классу упадёт ниже установленного порога. Параллельно в модуле словаря mitr_vocab.py (направление C1, где первое правило уже было предложено и принято самим шлюзом под квитанцией № 305) доля нераспознанных действий будет снижаться до нуля. По первой выборке внутреннего периметра за 16–26 сентября она составляла 27,8 %, во второй выборке внутреннего периметра за 9 октября — 32 %.

Третьи ворота касаются второй петли конструкции — автономного испытателя INDRA (система разведки S4 в модели Бира). На 30 сентября 2026 года из пяти приёмочных ворот INDRA (A–E) пройдено 0 из 5. С 17 октября по 12 ноября 2026 года испытатель пройдёт ворота A–C (изолированный полигон clean-room, вход только через внешнюю дверь арендатора MCP/REST, подсчёт доли ложных срабатываний генератором — к 31.10.2026) и ворота D–E. Замыкание иммунной петли, где находка INDRA превращается в проект контракта MITR — к 12.11.2026, пилот к 15.11.2026. Ни одно число по INDRA не входит в книгу без артефакта приёмочного скрипта.

1.5 Проекция 2035: три ветви будущего и инвариантность конструкции

Стратегическое планирование на десятилетие вперёд сталкивается с глубоким расхождением оценок среди создателей технологии.

С одной стороны, Дарио Амодеи (Anthropic) в эссе «Machines of Loving Grace» и выступлениях 2026 года прогнозирует появление «мощного ИИ» уровня нобелевских лауреатов в ближайшие один–два года. Иными словами, к 2027–2028 году (эссе «The Adolescence of Technology», январь 2026) и сжатие «столетия прогресса в десятилетие». При этом Сэм Альтман (OpenAI) в работе «The Gentle Singularity» констатирует начало взлёта, ожидая в 2026 году самостоятельных научных результатов и в 2027 году. Выхода роботов в физический мир (заявления лидеров лабораторий, 2025–2026). Модель GATE от Epoch AI при параметрах по умолчанию показывает, что уже при автоматизации 30 % задач экономический рост может превысить 20 % в год. Условный результат модели, а не прогноз Epoch (прогноз: Epoch AI, «AI and explosive growth redux», июнь 2025).

С другой стороны, нобелевский лауреат по экономике Дарон Асемоглу оценивает прирост ВВП от внедрения ИИ менее чем в 2 % за целое десятилетие. Демис Хассабис (Google DeepMind) отводит до появления общего искусственного интеллекта (AGI) около десяти лет. При этом Андрей Карпатый указывает на гораздо более длинный путь до создания по-настоящему сверхчеловеческого инженера-программиста (обзор позиций лидеров индустрии, 2025–2026, по вторичному источнику). Принстонские исследователи Арвинд Нараянан и Саяш Капур в программной работе «AI as Normal Technology» (2025) утверждают, что искусственный интеллект будет распространяться подобно электричеству и интернету — через десятилетия постепенной институциональной диффузии. Там узким местом выступает не алгоритм, а перестройка организаций (Narayanan & Kapoor, 2025).

Синергетика Ильи Пригожина и Германа Хакена, а также теория нестационарных сложных систем показывают, что вблизи точек бифуркации точечный статистический прогноз принципи-

ально невозможен. Поэтому вместо гадания о единой дате в основу книги положен сценарный метод с тремя ветвями (по образцу сценарных матриц OECD AI Futures и британского UK Government Office for Science). Там каждая ветвь снабжена жёстким внешним *репером* — датой и независимым источником, по которым любой наблюдатель в 2028, 2030 и 2035 годах сможет однозначно определить, по какой траектории пошла реальность:

- Ветвь A («Нормальная технология»): темп масштабирования замедляется; диффузия растягивается на десятилетия; автоматизация преимущественно дополняет труд специалиста. Реперы к 2028 году: горизонт задач METR удваивается медленнее, чем раз в 7 месяцев. Сбывается прогноз Gartner об отмене свыше 40 % агентных проектов. Доля открытых моделей в потоке токенов по рынку в целом превышает 50 % (на отдельных шлюзах порог пройден уже в 2026 году: у Vercel в августе — 56 % токенов). К 2035 году вычисления фронтира вырастут примерно в 1 000 раз (≈ ×2 в год) к уровню 2026 года, горизонт задачи составит около месяца, Automated Coder появится после 2035 года. При этом отставание открытых весов сожмётся до нуля (прогноз: ветвь A; реперы: Epoch data/trends, METR, aifuturesmodel.com, IAISR).
- Ветвь B («Продолжение тренда» — базовая расчётная ветвь книги). Вычисления обучения растут в 4–5 раз в год до 2030 года (к 2030 году — в 250–600 раз; по оценке Epoch AI от 20.08.2024, до 2030 года такой темп выполним, а дальше упирается прежде всего в электроэнергию), затем около 2 раз в год. К 2035 году — примерно в 10 000 раз; при сохранении ×4–5 в год было бы ×2,6-10^5...2-10^6. Цена токена фиксированного качества падает в 10 раз ежегодно. К 2030 году — удешевление в 10 000 раз, к 2035 году, в миллион и более раз; качество 2026 года становится бесплатным и встраивается в ОС и устройства. Гори-

зонт задачи METR при удвоении раз в 3–7 месяцев достигает недели–месяца работы к 2027–2028 годам и квартала–года («проект полного цикла») к 2027–2030 годам. После этого нынешний набор задач METR его уже не измеряет. Медиана Automated Coder реализуется в 2028–2030 годах (медианы двух авторов AI Futures Project по обновлению 16.08.2026 — ноябрь 2027 и январь 2030 года: граница ветвей B и C проходит между ними; консенсус Metaculus по вторичным обзорам: 25 % вероятности AGI к 2029 году и 50 % к 2033 году). Открытые веса отстают примерно на полгода (прогноз: ветвь B; реперы: Epoch AI, METR, AI Futures Project, Gartner).

– Ветвь C («Ускорение»): модели полностью автоматизируют сам цикл исследований и разработки искусственного интеллекта (сценарий AI 2027 Дэниела Кокотайло и соавторов). Automated Coder возникает до 2028 года (медиана ведущего автора AI Futures Project по обновлению 16.08.2026 — ноябрь 2027 года). Вычисления упираются исключительно в физический предел выработки электроэнергии. Горизонт задач выходит за пределы измеримости существующими тестами METR, а закрытые модели энерго-вычислительных корпораций уходят в отрыв от открытых весов за счёт монополии на гигаваттные кластеры. Прогноз: ветвь C; реперы: aifuturesmodel.com, макростатистика роста ВВП США > 5 % и доли ИИ-капвложений > 5 % ВВП к 2028.

Из этого следует главный кибернетический вывод главы, построенный по принципу инвариантности решения в ТРИЗ. Независимо от того, какая из трёх ветвей. A, B или C — реализуется в действительности, потребность в детерминированном шлюзе контроля действий остаётся абсолютной. Даже в самой консервативной ветви A удешевление токена и распространение бесплатных открытых моделей приводят к тому, что число машинных обращений к корпора-

тивным базам данных и внешним сетям возрастает на порядки, многократно перекрывая предел человеческого внимания по Эшби. В базовой ветви B к 2028 году 15 % всех повседневных рабочих решений принимаются агентами полностью автономно (против 0 % в 2024 году). При этом 33 % корпоративных приложений получают агентный ИИ (менее 1 % в 2024 году; прогноз: ветвь B; репер: пресс-релиз Gartner, 25.06.2025, горизонт 2028). В ветви C без математического затвора на канале исполнения любая корпоративная или государственная инфраструктура разрушается в первые минуты после делегирования полномочий. Меняется лишь плотность потока через трубу, но сам затвор S3 *с проверкой контракта менее чем за миллисекунду и неизменяемой квитанцией остаётся обязательным элементом конструкции во всех трёх мирах. А к 2030 году, когда качество базовой модели 2026 года станет бесплатным, конкурентное преимущество окончательно уйдёт из самой нейросети в слой верификации действия и в юридически закреплённое право на данные.*

1.6 Экономика доказанного действия и аргумент для капитала

Перенос фокуса с обучения моделей на детерминированный контроль действий меняет инвестиционную оптику. В 2026 году основная масса венчурного капитала продолжает финансировать создание новых агентов (исполнительный уровень S1 в модели жизнеспособной системы Стаффорда Бира) или шин их координации (S2). Однако десятикратное удешевление токена и четырёхмесячный лаг открытых китайских моделей превращают уровень S1 в бесплатное сырьё — подобно тому как ранее обесценилось базовое создание веб-сайтов.

По мере того как рынок осознаёт причину предстоящего закрытия 40 % агентных проектов по прогнозу Gartner, платёжеспособный спрос смещается в периметр аудита и допуска (S3*) и в пери-*

метр независимых полигонных испытаний (S4). Сам Gartner выделяет системы присмотра и ограничения агентов в отдельный быстрорастущий сегмент. Guardian agents («агенты-стражи»), прогнозируя, что к 2030 году они займут от 10 до 15 % всего рынка агентного искусственного интеллекта (прогноз: ветвь B; репер: пресс-релиз Gartner от 11.06.2025, горизонт 2030). Принципиальное отличие конструкции MITR внутри этого сегмента состоит в том, что роль стража в ней выполняет не ещё одна вероятностная языковая модель (которая стоит денег за каждый токен, вносит задержку в сотни миллисекунд и подвержена сговору или обходу промптом, как показали работы по ложной покорности и сговору моделей 2024–2026 годов), а детерминированный логический решатель над контрактами-данными.

Экономика этого решения опирается на твёрдые приборные показатели, которые кладутся на стол перед инвестиционным комитетом, грантодателем (четвёртая ступень коммерческой лестницы: грант института развития на переход от лабораторного образца к опытному периметру и софинансирование) и корпоративным заказчиком (Ступень 3: первый внешний арендатор с кодовыми и операционными агентами через протоколы MCP и REST;, план-график проекта):

1. Формула цены от защищённой ценности: стоимость подписки за защищённого агента на Ступени 3 и корпоративной лицензии на Ступени 5 определяется разностью между средневзвешенной величиной предотвращённого ущерба (удаление промышленной базы как в PocketOS, простой региона как в Kiro, регуляторный штраф или отказ страховщика) плюс экономия на внешнем аудите без доверия минус цена самой проверки.
2. Пять машинных доказательств для инвестора (P1–P5 на Ступени 4): вместо декларативных презентаций инвестиционному фонду предъявляются пять инженерных артефактов: P3 — живая хэш-цепь квитанций с внешними якорями

OpenTimestamps. P4 — приборный профиль задержек решателя: 0,74 / 3,3 мс на 26.09.2026 и 0,95 / 4,3 мс на рабочем сервере шлюза на 14.10.2026. P1 — отражение набора атак с независимой чужой подписью; P5, непрерывный промышленный пилот длительностью не менее 30 суток. Вся потенциальная выручка по внешним ступеням лестницы до момента получения первого оплаченного счёта от стороннего заказчика маркируется в финансовой модели исключительно какгипотеза — без произвольных множителей и бумажных миллиардов.

1.7 Протокол независимой верификации

Дисциплина книги, которая сама работает как измерительный прибор, предполагает, что ни одно базовое утверждение о работоспособности кода или динамике внешних трендов не должно приниматься на веру. Служба технического аудита инвестора, проектировщик безопасности или независимый исследователь могут самостоятельно воспроизвести показания приборов в двух периметрах — внутреннем (по набору верификации рабочего сервера шлюза) и внешнем (по открытым реперам мировых исследовательских центров).

Периметр А. Проверка инженерных артефактов MITR (внутренний периметр и открытый пакет повтора):

При наличии штатного аудиторского доступа к рабочему серверу шлюза проверка живости сервисов, статуса шести постулатов, задержки решателя (P4) и целостности цепи якорей (P3) выполняется четырьмя стандартными командами оболочки:

Полный перечень терминальных команд для проверки служб шлюза, суточных срезов, журналов задержки и автономного повтора хэш-цепи приведён в приложении Б.

Для внешнего аудитора без доступа к рабочему серверу шлюза открыт набор проверки (asi5.ru/console/verify/, скрипт публикации

publish_verify_kit.py): якоря цепей, метки времени OpenTimestamps, проверяемые против сети Bitcoin чужим клиентом, и проверщик целости цепи. Повтор самих решений требует цепи квитанций и текстов правил (в стандарте SMT-LIB, по одному на редакцию правила, адресованных отпечатком SHA-256). Их аудитору передают вместе с проверщиком verify_chain_standalone.py --replay, и любой SMT-решатель (Z3, CVC5) перепроверяет решения без обращения к серверам института. Подменить цепь нельзя — её голова стоит под меткой Bitcoin в открытом наборе. На 24.10.2026 так повторено 5 053 решения из 5 053 на 24.10.2026 (на 14.10.2026 — 4 958 из 4 958), контрольный прогон; сторонний аудитор этого ещё не делал.

Периметр Б. Сверка внешних макроэкономических и технологических реперов:

1. *Темп роста вычислений обучения фронтира:* сверяется ежегодно 1 февраля по реестру Epoch AI (epoch.ai/data/trends). До 2030 года снижение множителя ниже ×3 означает переход на ветвь А, удержание в коридоре ×4–5 подтверждает базовую ветвь В. После 2030 года ветвь В сама закладывает замедление до ≈ ×2 в год, и ветви А и В различаются уже не темпом, а достигнутым уровнем (×1 000 против ≈ ×10 000 к 2035 году).
2. *Дефляция стоимости токена:* сверяется ежегодно по витрине epoch.ai/data-insights/ai-inference-price-trends.
3. *Горизонт автономной задачи (p50):* сверяется по публикациям эталона METR (TH1.1 и далее). Пересечение отметки p50 >= 40 часов (рабочая неделя специалиста) до конца 2028 года служит репером ветви В.
4. *Медиана появления Automated Coder:* сверяется ежеквартально по обновлениям AI Futures Project (blog.aifutures.org).

1.8 Паспорт главы: реестр утверждений, прогнозов и источников

Таблица 1.1 — Реестр утверждений главы 1

№	Утверждение	Метка	Прибор / Источник	Дата	Условие погашения / сверки
A1.1	Вычисления обучения ведущих моделей растут ≈ ×5 в год с 2020 г. (удвоение за 5,2 мес.); с 2010 г. — ×4–5 в год (точечно ×4,1–4,4)	ИЗМЕРЕНО	Epoch AI (data/trends; анализ 28.05.2024)	05.02.2026	Ежегодная сверка с Epoch AI
A1.2	Мощность крупнейших прогонов обучения оценивается в 4–16 ГВт к 2030 г.	ПРОГНОЗ	Epoch AI (power-demands)	08.2025	Отчёты МЭА и ЦОД к 2030 г.
A1.3	Цена токена фиксированного качества падает ≈ ÷10 в год (Epoch: ÷9… ÷900 в год по планкам за 2022 — начало 2025 г.; ≈ ×13 в год по обновлённой оценке 2026 г.)	ИЗМЕРЕНО	a16z LLMflation (12.11.2024); Epoch AI (12.03.2025; 22.09.2026)	12.11.2024 / 12.03.2025	Ежегодный пересчёт инференса
A1.4	Цена фронтирного токена не падает: от $30 к $1,25 и снова к ≈ $5 за 1 млн (GPT-5.5)	ИЗМЕРЕНО (вторичн.)	Сводка цен API (вторичн.)	14.10.2026	Сверка с прайсом до 01.04.2027
A1.5	Горизонт задач METR (TH1.1, 228 задач): удвоение за 89–131 день; Opus 4.6 в феврале 2026 г. — p50 ≈ 12 ч; Mythos Preview в мае 2026 г., >16 ч, где замер ненадёжен	ИЗМЕРЕНО	METR TH1.1 (29.01.2026); данные METR (20.02.2026, 08.05.2026)	01–05.2026	Выгрузка METR при плановом пересчёте
A1.6	Медиана Automated Coder у ведущего автора сдвинута с конца 2029 г. на середину 2028 г. (02.04.2026), затем на ноябрь 2027 г. (16.08.2026); экспертный уровень — на 1,5 года раньше	ПРОГНОЗ	AI Futures Project (Q1 и Q2.5 2026 Timelines Updates)	02.04.2026; 16.08.2026	Ежеквартально: blog.aifutures.org
A1.7	Открытые веса отстают от закрытых < 1 года (IAISR) и в среднем ≈ 4 мес. (Epoch, 29.05.2026); на Vercel — 29 % токенов и < 4 % затрат (06.2026), 56 % токенов (08.2026)	ИЗМЕРЕНО	IAISR (03.02.2026); Epoch AI (29.05.2026); Vercel (06–08.2026)	02–08.2026	Отчёт IAISR 2027
A1.8	Агент Cursor на Opus 4.6 удалил продбазу и копии PocketOS (на том же томе) одним вызовом API, найдя токен в чужом файле; данные восстановила Railway	ИЗМЕРЕНО	Рассказ основателя PocketOS, заявление Railway (25.04.2026); AI Incident Database, инцидент № 1469; Lumenova (25.08.2026)	04.2026	Первичный отчёт верифицирован

№	Утверждение	Метка	Прибор / Источник	Дата	Условие погашения / сверки
A1.9	У 49 % из 108 организаций от 100 сотрудников (самоотобранная выборка) агент, прошедший внутренние проверки, вызвал сбой для клиентов; тесты не поймали ни один из 4 разобранных инцидентов 2026 г. (причина — доступ)	ИЗМЕРЕНО	VentureBeat (волна 07.2026, публикация 08.2026); Lumenova (25.08.2026)	07–08.2026	Первичные публикации сверены
A1.10	88 % организаций за прошедший год подтвердили или подозревают инцидент с агентом (опрос вендора, 919 респондентов)	ИЗМЕРЕНО	Gravitee, «State of AI Agent Security 2026»	04.02.2026	Сверено с первоисточником 14.10.2026
A1.11	В 52,9 % публичных описаний внедрений агентов в России (37 из 70) главный способ сдерживания — подтверждение человеком; 38,7 % публичных серверов MCP с определимой аутентификацией работают без неё	ИЗМЕРЕНО	отраслевой обзор института, 40-я неделя 2026 (70 записей); Bloomberry (24.02.2026)	2026	Архив отраслевого обзора и сканирование Bloomberry
A1.12	Свыше 40 % проектов агентного ИИ будут свёрнуты к концу 2027 г. из-за расходов, неясной ценности и слабого контроля	ИЗМЕРЕНО (прогноз)	Gartner Press Release	25.06.2025	Репер разрешения — конец 2027 г.
A1.13	Решатель z3_gate.py: решение 0,74 мс, вызов 3,3 мс на CPU (26.09, 17 контрактов). На рабочем сервере шлюза — 0,95 и 4,3 мс (14.10, 19 контрактов) и 1,08 и 5,22 мс (24.10, 20 контрактов). 157 Сценариев состязательного набора — $0; повтор проверщиком для постороннего, 1602 из 1602 (30.09), 4958 из 4958 (14.10)	ИЗМЕРЕНО		26.09–14.10.2026	Команды повтора Приложения Б
A1.14	За 23.10.2026 прибор впервые подтвердил 6 из 6 базовых положений (за 24.10 — 5 из 6), все 20 контрактов (6 основных классов) в запрете. На 09.10 — 1 из 6 и 1 класс из 7. Серия P2a и приёмка INDRA на внутреннем стенде пройдена 25.10 (12 из 12; на 30.09 — 0 из 5)	ИЗМЕРЕНО / СТРОИМ	gate_mode_report.py;	09.10–24.10.2026	Ворота P2a (31.10), INDRA (12.11), 14 суток - -six (14.12)

Таблица 1.2 — Реестр прогнозов главы 1

№	Показатель	2026 (источник)	Ветвь A / B / C к 2035	Репер	Разрешение
Г1	Вычисления обучения фронтира, рост в год	×5 с 2020 г., удвоение за 5,2 мес. (Epoch AI, 05.02.2026)	A: ≈ ×2/год (×1 000 к 2035) / B: ×4–5/год до 2030 (×250–600), затем ≈ ×2/год (≈ ×10 000 к 2035) / C: упор в энергию	Epoch AI (data/trends)	Ежегодно, 1 февраля

№	Показатель	2026 (источник)	Ветвь A / B / C к 2035	Репер	Разрешение
Г2	Цена токена фиксированного качества	÷10 в год (a16z, 12.11.2024; Epoch, 12.03.2025: ÷9…÷900)	A: замедление / B: ÷10/год (÷10^4 к 2030, не менее ÷10^6 к 2035 — ÷10^9 при сохранении темпа; качество 2026 бесплатно) / C: ÷100/год	Epoch AI (inference prices)	Ежегодно
Г3	Горизонт задач лучшей модели (p50)	≈ 12 ч (Opus 4.6, 02.2026), >16 ч (Mythos Preview, 05.2026); удвоение за 89–131 дн. (METR TH1.1)	A: месяц к 2035 / B: неделя–месяц к 2027–2028, квартал–год к 2027–2030 (удвоение раз в 3–7 мес.), дальше — вне нынешней шкалы / C: вне шкалы METR до 2028	METR time-horizons	Ежегодно (репер 2028: >= 1 нед.; выше 16 ч METR считает свой замер ненадёжным — нужен новый эталон)
Г4	Automated Coder (автоматический кодер)	Медиана ведущего автора — середина 2028 г. (02.04.2026), ноябрь 2027 г. (16.08.2026); второго автора — 2030 г.	A: после 2035 г. / B: 2028–2030 гг. / C: до 2028 г.	blog.aifutures.org	Ежеквартально
Г7	Доля повседневных рабочих решений, принимаемых агентами автономно	0 % в 2024 г., прогноз 15 % к 2028 г. (Gartner)	A: 10 % / B: 30 % / C: > 50 % к 2035 г.	Аналитические отчёты Gartner	Ежегодно (2028, 2035)
Г8	Доля систем надзора и допуска (guardian agents) на рынке агентного ИИ	Прогноз 10–15 % к 2030 г. (Gartner, 11.06.2025)	A: 5 % / B: 15 % / C: 25 %	Пресс-релизы и отчёты Gartner	2030, 2035

Источники Главы 1

1. Внутренние артефакты: MITR/core/verification/z3_gate.py; MITR/policies/contracts/ (20 на 24.10.2026; 19 на 14.10). Scripts/gate_mode_report.py, smt_latency_paper.py, verify_chain_standalone.py, publish_verify_kit.py, mitr_vocab.py;(§2a, §4); канон проектирования шлюза; 0273, 0275, 0280–0282; реестр продуктовой линейки; отраслевой обзор института (40-я неделя 2026, 70 описаний).
2. Модели и вычисления (первичные): Epoch AI (epoch.ai/data/trends от 05.02.2026; research/ai-and-compute; blog/power-demands-of-frontier-ai-training; data-insights/ai-inference-price-trends; blog/announcing-gate; data-insights/open-models-threshold). A16z (a16z.com/llmflation-llm-inference-cost/); AI Futures Project, Q1 2026 Timelines Update (blog.aifutures.org/p/q1-

2026-timelines-update, 02.04.2026) и Q2.5 2026 Timelines Update (16.08.2026). International AI Safety Report (internationalaisafetyreport.org, 03.02.2026).

3. Модели и сценарии (вторичные и аналитические разборы): Разбор METR TH1.1 (buttondown.com/anatol/archive/are-ai-time-horizons-still-doubling/); Metaculus (the-decoder.com); Позиции лидеров (letsdatascience.com); D. Amodei «Machines of Loving Grace»; S. Altman «The Gentle Singularity»; A. Narayanan & S. Kapoor «AI as Normal Technology» (technologyreview.com, asteriskmag.substack.com).
4. Инциденты, протоколы, венчур: AI Incident Database № 7311 (incidentdatabase.ai/reports/7311). Lumenova (lumenova.ai/blog/agentic-ai-security-incidents/); Aviatrix; Gartner (gartner.com, пресс-релизы 25.06.2025 и 11.06.2025 — Guardian Agents). Bloomberry (bloomberry.com/blog/we-analyzed-1400-mcp-servers-heres-what-we-learned/, 24.02.2026). DreamFactory (mcp-server-statistics, пересказ); Stacklok; Sacra / MindStudio (Claude Code ARR). PitchBook / Dealroom; Tianpan (approval-fatigue, 25.06.2026); AccountsRecovery (25.08.2026); Alignment faking — Anthropic и Redwood Research (arxiv.org/abs/2412.14093, 18.12.2024); сговор при взаимном надзоре (arxiv.org/abs/2602.20628, 24.02.2026).

ISS

Приложения

Приложение Б. Команды независимого повтора

Команды ниже повторяют инструментальные замеры, на которые ссылаются главы книги. Они проверяют активность служб шлюза, суточный срез шести базовых положений за 23 октября 2026 года (первое достижение отметки «правда 6 из 6») и срез за 24 октября 2026 года. «Правда 5 из 6, не измерено 1». Отдельная команда выводит реестр активных арендаторов. Далее приведены команды для выгрузки журнала задержки решателя (вторая строка — замер 26 сентября 2026 года, последняя, актуальный срез) и проверки последнего якоря цепи квитанций. Завершает набор автономный повтор всех решений цепи проверщиком для постороннего (5 053 из 5 053 на 24 октября 2026 года; 4 958 из 4 958 на 14 октября). Последняя строка проверяет квитанцию № 305 первого правила словаря, принятого под доказательством.

```
systemctl is-active v-bot v-mcp

cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --six --
  date 2026-10-23

cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --six --
  date 2026-10-24

cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --tenants -
  -date 2026-10-24

sed -n 2p /opt/v/data/logs/smt_latency_paper.jsonl; tail -1
  /opt/v/data/logs/smt_latency_paper.jsonl

tail -1 /opt/v/data/gate_anchors.jsonl

cd /opt/v && sudo -u v venv/bin/python scripts/verify_chain_standalone.py
  data/logs/gate_actions.jsonl --anchors data/gate_anchors.jsonl --replay
```

```
grep -m1 'seq.: 305,' /opt/v/data/logs/gate_actions.jsonl
```

Читателю без доступа к серверу шлюза открыт набор проверки asi5.ru/console/verify/ (его формирует скрипт publish_verify_kit.py). В него входят якоря цепей, их метки времени OpenTimestamps, проверяемые против сети Bitcoin сторонним клиентом, и проверщик целости цепи. Повтор самих решений требует цепи квитанций и текстов правил в формате SMT-LIB. Их передают аудитору вместе с проверщиком verify_chain_standalone.py, который повторяет каждое решение любым решателем на собственной машине (стандартная библиотека Python, ни строки кода шлюза). Проверщик с повтором выходит в открытый набор после публикации статьи; до тех пор контрольный повтор выполнялся на стенде разработки.