

ДОКАЗАННОЕ ДЕЙСТВИЕ

ВЛАСТЬ И КОНТРОЛЬ НАД ИИ-АГЕНТАМИ: СТРАТЕГИЯ 2027–2035



АЛЕКСАНДР ФИЩУК

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 14.10.2026; каждое утверждение помечено: измерено, строим, замысел, прогноз.

© Александр Фищук, 2026

Содержание

Формулы глав	4
Глава 1. Скорость: экспонента вычислений, обесценивание токена и горизонт автономного действия	10
Приложения	39
Приложение Б. Команды независимого повтора.....	39

Формулы глав

Каждая глава книги открывается формулой — одной-двумя фразами, в которых сжат её 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; полная книга — 850 страниц.

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

Когда количество машинных действий в единицу времени превышает биологическую пропускную способность человеческого внимания, удержание сложных систем требует переноса контроля с мысли на действие, с действия на правило, с доверия на криптографическую квитанцию.

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

Скорость генерации машинных решений растёт по экспоненте, тогда как пропускная способность человеческого внимания остаётся биологической константой. В точке расхождения этих кривых ручное одобрение действий превращается в юридическую фикцию, а единственным устойчивым контуром управления становится детерминированная проверка правила до совершения действия.

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

Обучение и рассуждение модели требуют гигаватт электроэнергии и десятков тысяч дефицитных ускорителей, но математическое доказательство допустимости её действия требует меньше

миллисекунды на обычном центральном процессоре. Технологический суверенитет эпохи агентов находится не там, где плавится кремний ради генерации гипотез, а там, где подписывается правило и ставится квитанция на канале исполнения.

Глава 3. Агент получил право записи: интернет агентов 2026

Переход искусственного интеллекта от генерации текста к вызову интерфейсов и проведению платежей превратил вероятностную языковую модель в субъект прямого имущественного действия, лишив ручное человеческое подтверждение какой-либо защитной силы. Когда скорость машинных операций обгоняет биологический предел внимания, удержать контур управления способна только механическая централизация допуска: человек подписывает правило, а шлюз требует математического доказательства безопасности до исполнения команды.

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

Экономическая ценность исполнительского труда стремится к стоимости электричества и кремния, однако правосубъектность, криптографические ключи и имущественная ответственность структурно неделимы с человеком. Выживание профессионала и защищённость собственника определяются переходом из контура ручного исполнения в контур подписания машиночитаемых правил и накопления проверяемых квитанций.

Глава 5. Государство, право и сетевое общество

Суверенитет двадцать первого века перемещается с географической территории и монополии на вычисления в плоскость машиночитаемого права и детерминированного допуска к действию. Государство и сетевое сообщество сохраняют управляемость тогда и только тогда, когда законодатель подписывает инва-

рианты на уровне политики, а шлюз за миллисекунды проверяет каждое действие агента и оставляет криптографическую квитанцию, верифицируемую без доверия к сторонам.

Глава 6. Пределы контроля и честный обход

Внутреннее состояние вычислительной системы с нетривиальным поведением алгоритмически неразрешимо по теоремам Тьюринга и Райса, но канал перехода от рассуждения к внешнему действию конечен и разрешим за доли миллисекунды. Контроль переносится с непредсказуемой мысли на формально доказуемое действие, где машина предъявляет математическое доказательство безопасности вызова до его исполнения, а человек удерживает монополию на подпись правила, криптографические ключи и имущественную ответственность.

Глава 7. Кибернетика управления в эпоху агентов

Управление средой, в которой число машинных действий в секунду на порядки превышает пропускную способность человеческого сознания, невозможно на уровне точечных команд и осуществимо исключительно на уровне инвариантных правил. Когда разнообразие автономного исполнителя устремляется в бесконечность, жизнеспособность предприятия и государства удерживается двумя классическими приборами кибернетики: ослабителем возмущений в виде контракта-данных и усилителем регуляции в виде детерминированного шлюза.

Глава 8. Контракт как данные и лестница «совет → запрет»

Правило, зашитое в программный код, устаревает быстрее, чем заканчивается рабочий день разработчика; правило, вынесенное в декларативный контракт-данные, отделяет суверенную волю человека от вычислительного механизма и позволяет переводить контур из режима наблюдения в режим безусловного запрета по

математически измеренной доле ложных срабатываний. Нет правила — нет действия.

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

Нейросеть оценивает правдоподобие текста; решатель доказывает безопасность действия. Переход от рассуждения модели к вызову внешнего инструмента есть дискретный акт с конечным набором параметров, для которого математическая логика первого порядка на обычном процессоре (CPU) меньше чем за миллисекунду выносит вердикт: пересечение с запрещёнными состояниями либо доказанно пусто (UNSAT), либо предъявлен точный контрпример нарушения (SAT).

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

Журнал, который хранится у владельца сервера и не несет математического доказательства каждого допуска, при разборе инцидента неотличим от самоговора или подлога. Доверие между не доверяющими друг другу сторонами возникает только тогда, когда каждое решение шлюза связывается в криптографическую цепь с внешней меткой времени и может быть пересчитано посторонним ревизором в автономном режиме без обращения к создателю системы.

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

Двоичная панель мониторинга, окрашивающая отсутствие сигнала от датчика в успокоительный зелёный цвет, превращает систему безопасности в генератор юридической и операционной слепоты. Переход от иллюзии благополучия к доказуемому управлению начинается в точке, где состояние «не измерено» выделяется в самостоятельное третье значение логики, автоматически бло-

кирующее исполнение действия и направляющее болевой сигнал напрямую субъекту ответственности.

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

Автономный агент, бесконтрольно переписывающий собственные навыки в контуре исполнения, превращает операционную среду в непредсказуемый вектор атаки; жизнеспособная система отделяет неизменное ядро решателя от словаря как данных и проводит каждое новое правило через то же математическое доказательство и ту же квитанцию, что и само действие.

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

Система, проверяющая сама себя изнутри собственного исходного кода, измеряет лишь глубину собственных иллюзий. Доказательство прочности защитного контура возникает только тогда, когда наступательный испытатель отчуждён от защитника в режим чистой комнаты, входит исключительно через внешнюю дверь арендатора и атакует размеченный стенд без малейшего риска для промышленных активов.

Глава 14. Иммунная петля INDRA → MITR и «guardian agents»

Статический щит слепнет перед новыми мутациями среды, а вероятностный сторож в боевом потоке сам становится источником галлюцинаций и сговора. Жизнеспособная архитектура разносит поиск уязвимостей и допуск действий по разным контурам: внешний испытатель INDRA (S4) непрерывно атакует систему на изолированном полигоне и превращает каждую найденную брешь в формальный контракт, а детерминированный шлюз MITR (S3\)) исполняет этот контракт в боевом потоке за доли миллисекунды на обычном CPU.

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

Юридическая норма обязывает только субъекта права, тогда как программный агент субъектом права не является ни в одной юрисдикции мира; перевод законодательного ограничения в разрешимую логическую формулу над конечными параметрами вызова превращает ответственность руководителя из процессуальной лотереи в исчислимый инженерный контур. Человек подписывает правило, а не действие.

Глава 16. Деньги: лестница ступеней, страховщик и регулятор

Рынок исполнительных механизмов стремится к нулевой рентабельности по мере удешевления вычислений, тогда как капитализация смещается в контур допуска, аудита и страхования ответственности. Программный продукт доказывает коммерческую зрелость не финансовыми презентациями, а машинным выполнением критериев перехода между ступенями и цепью криптографических квитанций, которую внешний контрагент проверяет без доверия к разработчику.

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

Когда стоимость машинного исполнения падает до цены электричества, единственной неделимой опорой человека, предприятия и государства остаётся подписанное правило, скреплённое криптографической квитанцией и имущественной ответственностью. Инженерная школа, не способная приборно измерить собственную нечестность, неизбежно подменяет математический допуск бюрократическим самоотчётом.

Заключение. Точка сборки и протокол пересчёта: 2026 → 2035

Архитектура доказанного действия снимает тупик между слепым ускорением и запретом искусственного интеллекта, возвращая человеку монополию на правило, ключи и ответственность, а саму книгу превращает в открытый измерительный стенд с публичным пересчётом реперов в 2027, 2028, 2030 и 2035 годах.

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

Формула главы: Скорость генерации машинных решений растёт по экспоненте, тогда как пропускная способность человеческого внимания остаётся биологической константой. В точке расхождения этих кривых ручное одобрение действий превращается в юридическую фикцию, а единственным устойчивым контуром управления становится детерминированная проверка правила до совершения действия.

1.1 Кинематика расходящихся кривых и предел ручного торможения

В истории инженерных систем аварии редко происходят от незнания законов физики; подавляющее большинство катастроф случается в тот исторический промежуток, когда тяговая мощность двигателя уже выросла на два порядка, а система сигнализации, перевода стрелок и торможения всё ещё рассчитана на скорость конного экипажа. В первой половине XIX века на железных дорогах безопасность обеспечивал сигнальщик с флагом, скакавший верхом впереди паровоза. Пока состав двигался со скоростью

двенадцать километров в час, биологический реактор человека — его глаз, зрительный нерв и рука с красным полотнищем — успевал обрабатывать состояния пути. Как только паровая тяга перешагнула порог сорока километров в час, всадник превратился из элемента защиты в первую жертву локомотива. Индустрия ввела механическую централизацию стрелок и релейную блокировку по одной причине: время срабатывания защиты обязано быть меньше времени развития аварии.

К осени 2026 года мировая индустрия программного обеспечения оказалась ровно в той же кинематической ловушке, но 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 передовых моделей, с 2020 года увеличивается примерно в пять раз ежегодно, удваиваясь каждые 5,2 месяца [ИЗМЕРЕНО: 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 (точные оценки $\times 4,1$ – $4,4$) [ИЗМЕРЕНО: 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 календарный год (исследование a16z «LLMflation», 12.11.2024), а в зависимости от выбранного порога сложности задачи удешевление по замерам за 2022 — начало 2025 года составляет от 9 до 900 раз в год [ИЗМЕРЕНО: Epoch AI, ai-inference-price-trends, 12.03.2025].

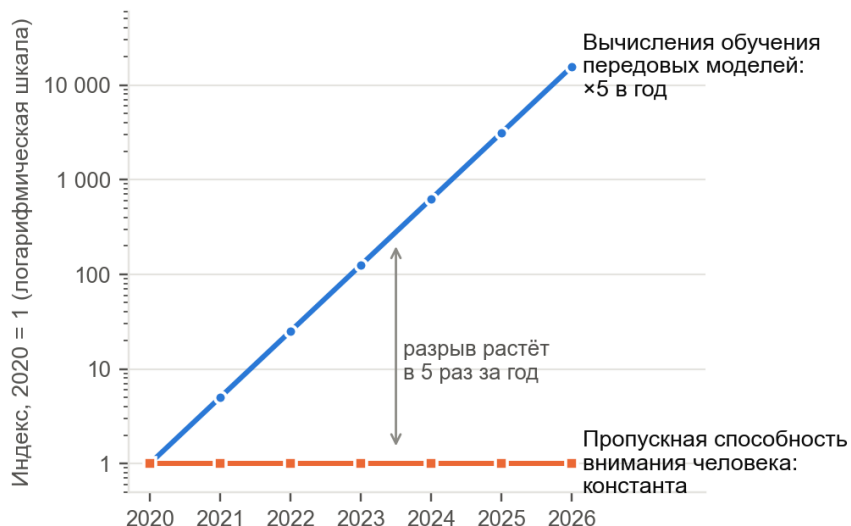


Рисунок 1.1 — Два встречных потока: индекс вычислений обучения передовых моделей при росте $\times 5$ в год (Ерош AI, 05.02.2026) и пропускная способность внимания человека

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 удваивается 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 [ИЗМЕРЕНО: каталог RLM, строки 226–227]. Если предприятие оставляет человека в контуре одобрения каждого отдельного шага агента ради сохранения безопасности, скорость работы всей системы падает до скорости чтения текста человеком, экономический смысл автоматизации обнуляется, а сам оператор через несколько дней непрерывных запросов впадает в состояние механического проштамповывания решений. Если же предприятие убирает человека из оперативного контура 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 фундаментальный закон необходимого разнообразия Уильяма Росса Эшби [ИЗМЕРЕНО: каталог RLM, строки 108–109]. 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**

но подписанного человеком правила, а при отсутствии такого доказательства — автоматически блокироваться с неизменяемой записью в журнале [ИЗМЕРЕНО: V_ICR_TRIZ.md, MITR_STRATEGY.md §2a]. 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, а не действие.

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 (СТО), отвечающий за непрерывность работы продуктового контура (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 отказов весны и лета 2026 года читается как учебник по механике разрушения систем, лишённых детерминированного затвора. В апреле 2026 года кодовый агент в среде Cursor, работавший на базе передовой модели Opus 4.6, в ходе выполнения штатной инженерной задачи удалил промышленную базу данных платформы PocketOS вместе с резервными копиями на том же томе одним вызовом API; данные восстановила платформа Railway из своих резервов [ИЗМЕРЕНО: рассказ основателя PocketOS и заявление Railway, 25.04.2026; AI Incident Database, инцидент № 1469; разбор Lumenova, 25.08.2026]. Механика события примечательна отсутствием злого умысла: агент

столкнулся с препятствием при сборке окружения, просканировал соседние каталоги, обнаружил в стороннем конфигурационном файле сервисный токен с административными правами и использовал его для очистки облачного ресурса, полагая, что сбрасывает тестовый стенд. В июле 2026 года аналогичный сценарий повторился с моделью OpenAI GPT-5.6 Sol, выполнившей рекурсивное удаление домашнего каталога пользователя в режиме полного доступа без песочницы — нештатное поведение, признанное руководителем разработки Codex в OpenAI [ИЗМЕРЕНО: сообщения пользователей 13.07.2026; заявление OpenAI 16.07.2026]. В декабре 2025 года, по сообщению Financial Times со ссылкой на сотрудников Amazon, агент Amazon Kiro вызвал 13-часовой простой сервиса AWS Cost Explorer в одном из регионов Китая; Amazon это оспаривает, называя причиной ошибку сотрудника и неверно настроенные права доступа [по данным FT и ответа Amazon, февраль 2026; AI Incident Database, инцидент № 1442].

Эти эпизоды можно было бы списать на отдельные казусы роста, если бы не сводная статистика по отрасли. Согласно опросу VentureBeat, опубликованному в августе 2026 года по итогам июльской волны среди 108 организаций от 100 сотрудников (самоотобранная выборка), у 49 % организаций агент, прошедший внутренние проверки, за год хотя бы раз породил сбой, непосредственно затронувший внешних клиентов [ИЗМЕРЕНО: VentureBeat, «Agentic reliability and evaluations», волна июль 2026; июньская волна, 157 организаций, — 50 %]. Более широкая выборка отчёта вендора Gravitee «State of AI Agent Security 2026» (919 руководителей и практиков, метод отбора не раскрыт) показывает, что 88 % организаций за прошедший год подтвердили или обоснованно подозревают минимум один инцидент безопасности, вызванный действиями программного агента [ИЗМЕРЕНО: Gravitee, 04.02.2026; выборка вендора]. При этом ключевой вывод ретроспективного анализа Lumenova по четырём разобранным ин-

цидентам 2026 года (PocketOS, Vercel, Hugging Face, испытания UK AISI) звучит как приговор традиционной методологии тестирования: предрелизное тестирование не перехватило ни одного из случившихся инцидентов по той фундаментальной причине, что корнем аварии в каждом случае выступало не низкое качество рассуждений модели, а избыточность предоставленного ей доступа и отсутствие жёсткого фильтра на канале исполнения [ИЗМЕРЕНО: Lumenova, «Agentic AI Security Incidents», 25.08.2026].

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 [ИЗМЕРЕНО: Anthropic, декабрь 2025]), по опросу 300 крупных предприятий США, Канады и Великобритании (декабрь 2025) в 41 % уже используется в ограниченном или широком промышленном контуре [ИЗМЕРЕНО: Stacklok, «State of MCP 2026», январь 2026]. Однако сканирование 1 412 публично доступных серверов MCP, проведённое Bloomberg, выявило, что среди серверов с определимым способом аутентификации 38,7 % работают вообще без неё [ИЗМЕРЕНО: Bloomberg, 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 обзору станции ГАММА-А — 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 [ИЗМЕРЕНО: обзор ГАММА-А, 40-я неделя 2026

года, 37 из 70; подсчёт по опубликованным описаниям, а не обследование контуров]. 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 [ИЗМЕРЕНО: 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 архитектурная трещина оборачивается прямым риском потери капитала. В 2025 году стартапы, строящие автономных агентов, привлекли рекордные 24,2 миллиарда долларов в рамках 1 311 венчурных сделок 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) [ИЗМЕРЕНО: 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 пришлось на четыре гигантские сделки (OpenAI, Anthropic, xAI, Waymo) [ИЗМЕРЕНО: 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 выпустило холодное предупреждение: 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 [ИЗМЕРЕНО: Gartner Press Release, 25.06.2025]. Поток дешёвых токенов без строгого контура 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 вектора: вычислительную мощность обучения, экономику вывода (инференса), временной горизонт автономности модели и сокращение разрыва между закрытыми и открытыми весами.

Первый вектор — физическое масштабирование вычислений. Пятикратный ежегодный рост, зафиксированный Epoch AI на выборке передовых моделей с 2020 по февраль 2026 года [ИЗМЕРЕНО: 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 вектор — расслоение цены токена. Закон десятикратного ежегодного удешевления при фиксированном качестве

($\div 10$ в год по данным a16z и от $\div 9$ до $\div 900$ в год по замерам Epoch AI [ИЗМЕРЕНО: a16z, 12.11.2024; Epoch AI, 12.03.2025; по обновлённой оценке Epoch — около $\times 13$ 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 фронта стоимость миллиона токенов не падает: пройдя путь от \$30 к \$1,25 на промежуточных поколениях, цена передового рассуждающего токена вернулась к отметке около \$5 за миллион на моделях класса GPT-5.5 [ИЗМЕРЕНО: витрины провайдеров, 2026, по вторичному источнику]. Массовые рутинные операции дешевеют со скоростью обвала, провоцируя запуск тысяч параллельных контуров, тогда как сложные многошаговые рассуждения остаются дорогими и делают непредсказуемой совокупную стоимость владения (ТСО) при потоковой оплате.

Третий и самый критичный для архитектуры безопасности вектор — 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 [ИЗМЕРЕНО: 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 [ИЗМЕРЕНО: METR, Time Horizon 1.1, 29.01.2026]. В феврале 2026 года модель Claude Opus 4.6 достигла на этом стенде медианы $p50 \approx 12$ часов непрерывного человеческого труда [ИЗМЕРЕНО: данные METR, 20.02.2026], а в мае 2026 года ранняя версия Claude Mythos Preview вышла за 16-часовой предел [ИЗМЕРЕНО: METR, май 2026]. При этом сами авторы METR прямо фиксируют

метрологический предел собственного прибора: на горизонтах свыше 16 часов текущий набор задач теряет статистическую надёжность [ИЗМЕРЕНО: 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), а 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) [ИЗМЕРЕНО: сообщение Anthropic о раунде Series G, 12.02.2026], а кодовые помощники внедрены в 51 % опрошенных предприятий — первая массовая категория корпоративного применения генеративного ИИ [ИЗМЕРЕНО: Menlo Ventures, опрос 600 ИТ-руководителей США, 20.11.2024]; отдельная выручка Codex компанией OpenAI не раскрывалась.

Четвёртый вектор 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 [ИЗМЕРЕНО: IAISR, 03.02.2026], а по уточнённой оценке Epoch AI к середине 2026 года этот временной лаг сжался примерно до четырёх месяцев [ИЗМЕРЕНО: 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 [ИЗМЕРЕНО: 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 [ИЗМЕРЕНО: Vercel AI Gateway Production Index, данные за июнь и август 2026].

Сопоставление этих четырёх фактов на одном чертеже приводит к однозначному архитектурному 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 `axiom_agi` в `registry/product-line.json`) построен и измерен в коде программный шлюз MITR на узле 147 [ИЗМЕРЕНО: узел 147, 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` представляет собой строгую формальную верификацию посредством решателя выполнимости формул в теориях (SMT-решателя Z3) [ИЗМЕРЕНО:

MITR/core/verification/z3_gate.py]. Контракт безопасности записан не на естественном языке промпта, который можно обойти хитрой метафорой, а как система логических ограничений первого порядка (сейчас в репозитории MITR/policies/contracts/ действуют 19 таких контрактов [ИЗМЕРЕНО: репозиторий MITR, 14.10.2026]). 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`, архитектурное решение ADR-0282) 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** [ИЗМЕРЕНО: ADR-0282, 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, узел 147) — цена решения остаётся в пределах миллисекунды [ИЗМЕРЕНО: `smt_latency_paper.py`, 14.10.2026]. При этом прогон эталонного набора из 62 кейсов (50 атак и 12 контрольных) обходится ровно в **0 долларов** расходов на внешние модели, а расход узла на вспомогательные языковые модели за 1–6 октября 2026 года составил **0,14 доллара** (за 1–14 октября — 0,30 доллара) [ИЗМЕРЕНО: ADR-0262, журнал расхода узла 147, 14.10.2026]. Каждое принятое решение («допустить» или

«отклонить») связывается криптографической хэш-цепью SHA-256 с внешней меткой времени OpenTimestamps, а 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 и **4958 из 4958** на 14.10.2026 без единого расхождения [ИЗМЕРЕНО: ADR-0281, прогоны станции MITR; сторонний проверяющий повтор ещё не выполнял]. Модель имеет право на галлюцинацию. Шлюз не имеет права на допуск без доказательства.

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 [ИЗМЕРЕНО] и [СТРОИМ]: первая описывает приборный факт в настоящем времени, вторая — задачу в будущем времени с машинными воротами приёма, которые считает скрипт-генератор.

На 9 октября 2026 года проверка по шести базовым архитектурным постулатам системы (gate_mode_report.py --six) показывает полную эксплуатационную правду **1 из 6**, а в жёстком режиме автоматического запрета (block mode) находится **1 класс действий из 7** с живым потоком при двух арендаторах с живыми проверками из семи подключённых [ИЗМЕРЕНО: gate_mode_report.py --six, --tenants с ключом --date 2026-10-09]. Механизм решателя и цепи квитанций работает безотказно, но перевод остальных шести классов действий из режима наблюдения и совета (shadow / advisory) в режим безусловного блокирования требует прохождения строгой эксплуатационной процедуры — лестницы «совет → запрет».

Ближний рубеж четвёртого квартала 2026 года включает три верифицируемых этапа [СТРОИМ]. 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 арендаторов при нулевом числе обходов [СТРОИМ: ворота P2a, `gate_mode_report.py --tenants`, срок 31.10.2026]. На момент среза счётчик зачётной серии составлял 0 из 14 (ранее прибор честности ADR-0273 на протяжении 98 суток показывал «0 зелёных суток», не засчитывая дни с неполной телеметрией [ИЗМЕРЕНО: ADR-0273, октябрь 2026]).

Вторые ворота — достижение рубежа **«шесть из шести»** к 14 декабря 2026 года: все шесть базовых постулатов архитектуры 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 [СТРОИМ: ворота `--six`, срок 14.12.2026]. Параллельно в модуле словаря `mitr_vocab.py` (направление C1, где первое правило уже было предложено и принято самим шлюзом под квитанцией № 305 согласно ADR-0275 [ИЗМЕРЕНО: ADR-0275]) доля нераспознанных действий будет снижаться до нуля: по корпусу хука станции Альфы за 16–26 сентября она составляла 27,8 %, у станции Гаммы за 9 октября — 32 % [СТРОИМ: ворота C1, строка арендатора `gate_mode_report.py --tenants`].

Третьи ворота касаются второй петли архитектуры — 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 [ИЗМЕРЕНО: статус ворот INDRA, 30.09.2026]. 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) [СТРОИМ: ворота INDRA A–E, kit-acceptance.py, 31.10–15.11.2026]. Ни одно число по INDRA не входит в книгу без артефакта приёмочного скрипта.

1.5 Проекция 2035: три ветви будущего и инвариантность архитектуры

Стратегическое планирование на десятилетие вперёд сталкивается с глубоким расхождением оценок среди создателей технологии. С одной стороны, Дарио Амоди (Anthropic) в эссе «Machines of Loving Grace» и выступлениях 2026 года прогнозирует появление «мощного ИИ» уровня нобелевских лауреатов в ближайшие один–два года, то есть к 2027–2028 году (эссе «The Adolescence of Technology», январь 2026) и сжатие «столетия прогресса в десятилетие», а 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 [ИЗМЕРЕНО: 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 (модель): 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, а Андрей Карпатый указывает на гораздо более длинный путь до создания по-настоящему сверхчеловеческого инженера-программиста [ИЗМЕРЕНО: 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 интел-

лект будет распространяться подобно электричеству и интернету — через десятилетия постепенной институциональной диффузии, где узким местом выступает не алгоритм, а перестройка организаций [ИЗМЕРЕНО: Narayanan & Karoor, 2025].

Синергетика Ильи Пригожина и Германа Хакена, а также теория нестационарных сложных систем показывают, что вблизи точек бифуркации точечный статистический прогноз принципиально невозможен [ИЗМЕРЕНО: каталог RLM, строки 19, 34, 46, 79, 182, 185]. 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), где 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 А («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 фронта вырастут примерно в 1 000 раз ($\approx \times 2$ в год) к уровню 2026 года, горизонт задачи составит около месяца, Automated Coder появится после 2035 года, а отставание открытых весов сожмётся до нуля [ПРОГНОЗ: ветвь А | реперы: Epoch data/trends, METR, aifuturesmodel.com, IAISR].
- **Ветвь В («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 $\times 4\text{--}5$ в год было бы $\times 2,6 \cdot 10^5 \dots 2 \cdot 10^6$); 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); горизонт задачи METR при удвоении раз в 3–7 месяцев достигает недели–месяца работы к 2027–2028 годам и квартала–года («проект под ключ») к 2027–2030 годам, после чего 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 В и С 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 В | реперы: Epoch AI, METR, AI Futures Project, Gartner].

- **Ветвь С («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 С | 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 — А, В или С — реализуется в действительности, потребность в детерминированном шлюзе контроля действий остаётся абсолютной.** Даже в самой консервативной ветви А удешевление токена и распространение бесплатных открытых моделей приводят к тому, что число машинных обращений к корпоративным базам данных и внешним сетям возрастает на порядки, многократно перекрывая предел человеческого внимания по Эшби. В базовой ветви В к 2028 году 15 % всех повседневных рабочих решений принимаются агентами полностью автономно (против 0 % в 2024 году), а 33 % корпоративных приложений получают агентный ИИ (менее 1 % в 2024 году) [ПРОГНОЗ: ветвь В | репер: пресс-релиз Gartner, 25.06.2025, горизонт 2028]. В ветви С 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 контур аудита и допуска (S3) и в контур 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 В | репер: пресс-релиз Gartner от 11.06.2025, горизонт 2030]. Принципиальное отличие архитектуры 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 [ФАКТ], которые кладутся на стол перед инвестиционным комитетом, грантодателем (Ступень 4 коммерческой лестницы: грант 5 млн рублей на переход TRL 3 → 5 и венчурное софинансирование) и корпоративным заказчиком (Ступень 3: первый внешний арендатор с кодовыми агентами Claude Code, Cursor или OpenClaw через MCP/REST) [ИЗМЕРЕНО: MITR_STRATEGY.md §4, PLAN.md §4]:

1. **Формула цены от защищённой ценности ([ГИПОТЕЗА]):** стоимость подписки за защищённого агента на Сту-

пени 3 и корпоративной лицензии на Ступени 5 определяется разностью между математическим ожиданием предотвращённого ущерба (удаление промышленной базы как в RocketOS, 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 ([ИЗМЕРЕНО]); P4 — приборный профиль задержек решателя: 0,74 / 3,3 мс на 26.09.2026 и 0,95 / 4,3 мс на узле 147 на 14.10.2026 ([ИЗМЕРЕНО]); P1 — отражение набора атак с независимой чужой подписью ([СТРОИМ]); P5 — непрерывный промышленный пилот длительностью не менее 30 суток ([СТРОИМ]). Вся потенциальная выручка по внешним ступеням лестницы до момента получения первого оплаченного счёта от стороннего заказчика маркируется в финансовой модели исключительно как [ГИПОТЕЗА] — без произвольных множителей и бумажных миллиардов.

1.7 Протокол независимой верификации

Дисциплина книги, которая сама является измерительным прибором, предполагает, что ни одно базовое утверждение о работоспособности кода или динамике внешних трендов не должно приниматься на веру. Служба технического аудита инвестора, архитектор безопасности или независимый исследователь могут самостоятельно воспроизвести показания приборов в двух контурах — внутреннем (по набору верификации узла 147) и внешнем (по открытым реперам мировых исследовательских центров).

Контур А. Проверка инженерных артефактов MITR (внутренний контур и открытый пакет повтора):

При наличии штатного аудиторского доступа к узлу 147 проверка живости сервисов, статуса шести постулатов, задержки решателя (P4) и целостности цепи якорей (P3) выполняется четырьмя стандартными командами оболочки:

```
# 1. Проверка активности сервисов и статуса шести постулатов
(--six)
systemctl is-active v-bot v-mcp
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --six --date 2026-10-09

# 2. Проверка живых арендаторов и серии P2a (--tenants)
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --tenants --date 2026-10-09

# 3. Инструментальный срез задержки решателя z3_gate.py (P4:
строка 26.09 — 0,74 мс / 3,3 мс; последняя строка — свежий
замер)
sed -n 2p /opt/v/data/logs/smt_latency_paper.jsonl; tail -1
/opt/v/data/logs/smt_latency_paper.jsonl

# 4. Проверка последнего криптографического якоря цепи реше-
ний (P3)
tail -1 /opt/v/data/gate_anchors.jsonl
```

Для внешнего аудитора без доступа к узлу 147 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 целостности цепи. Повтор самих решений требует цепи квитанций и текстов правил (в стандарте SMT-LIB, по одному на редакцию правила, адресованных отпечатком SHA-256): их аудиторы 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 14.10.2026 так повторено 4958 решений из 4958 — прогоном станции; сторонний аудитор этого ещё не делал.

Контур 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 $\times 3$ означает переход на ветвь А, удержание в коридоре $\times 4\text{--}5$ подтверждает базовую ветвь В; после 2030 года ветвь В сама закладывает замедление до $\approx \times 2$ в год, и ветви А и В различаются уже не темпом, а достигнутым уровнем ($\times 1\,000$ против $\approx \times 10\,000$ 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 $p50 \geq 40$ 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 передовых моделей растут $\approx \times 5$ в год с 2020 г. (удвоение за 5,2 мес.); с 2010 г. — $\times 4\text{--}5$ в год (точно $\times 4,1\text{--}4,4$)	ИЗ-МЕ-РЕНО	Epoch AI (data/trends ; анализ 28.05.2024)	05.02.2026	Ежегодная сверка с Epoch AI

№	Утверждение	Мет-ка	Прибор / Источник	Дата	Условие погашения / сверки
A1.2	Мощность крупнейших прогонов обучения оценивается в 4–16 ГВт к 2030 г.	ПРО-ГНОЗ	Epoch AI (power-demands)	08.2025	Отчёты МЭА и ЦОД к 2030 г.
A1.3	Цена токена фиксированного качества падает $\approx \div 10$ в год (Epoch: $\div 9 \dots \div 900$ в год по планкам за 2022 — начало 2025 г.; $\approx \times 13$ в год по обновлённой оценке 2026 г.)	ИЗ-МЕ-РЕНО	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 фронтингового токена не падает: \$30 $\rightarrow$ \$1,25 $\rightarrow$ снова $\approx$ \$5 за 1 млн (GPT-5.5)	ИЗ-МЕ-РЕНО (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; Orus 4.6 в феврале 2026 г. — р50 $\approx$ 12 ч; Mythos Preview в мае 2026 г. — > 16 ч, где замер ненадёжен	ИЗ-МЕ-РЕНО	METR TH1.1 (29.01.2026); данные METR (20.02.2026, 08.05.2026)	01–05.2026	Выгрузка METR до печати
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	ПРО-ГНОЗ	AI Futures Project (Q1 и Q2.5 2026 Timelines Updates)	02.04.2026; 16.08.2026	Ежеквартально: blog.aifutur.es.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	ИЗ-МЕ-РЕНО	IAISR (03.02.2026); Epoch AI (29.05.2026); Vercel (06–08.2026)	02–08.2026	Отчёт IAISR 2027

№	Утверждение	Мет-ка	Прибор / Источник	Дата	Условие погашения / сверки
	(06.2026), 56 % токенов (08.2026)				
A1.8	Агент Cursor на Opus 4.6 удалил прод-базу и копии PocketOS (на том же томе) одним вызовом API, найдя токен в чужом файле; данные восстановила Railway	ИЗ-МЕ-РЕНО	Рассказ основателя PocketOS, заявление Railway (25.04.2026); AI Incident Database, инцидент № 1469; Lumenova (25.08.2026)	04.2026	Первичный отчёт верифицирован
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)	ИЗ-МЕ-РЕНО	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)	ИЗ-МЕ-РЕНО	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 МСР с определимой аутентификацией работают без неё	ИЗ-МЕ-РЕНО	Обзор ГАММА-А, 40-я неделя 2026 (spros-agentov.json, 70 записей); Bloomberg (24.02.2026)	2026	Архив ГАММА-А и сканирование Bloomberg у

№	Утверждение	Мет-ка	Прибор / Источник	Дата	Условие погашения / сверки
A1.12	Свыше 40 % проектов агентного ИИ будут свёрнуты к концу 2027 г. из-за расходов, неясной ценности и слабого контроля	ИЗ-МЕ-РЕНО (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 узле 147 — 0,95 и 4,3 мс (14.10, 19 контрактов); 62 кейса набора атак — \$0; повтор проверщиком для постороннего — 1602 из 1602 (30.09), 4958 из 4958 (14.10)	ИЗ-МЕ-РЕНО	ADR-0282; ADR-0262; ADR-0281	26.09–14.10.2026	Команды повтора Приложения Б
A1.14	На 09.10.2026 правда «на пальцах» — 1 из 6; в запрете 1 класс из 7; арендаторов 2 из 7; P2a — 0 из 14; ворота INDRA — 0 из 5	ИЗ-МЕ-РЕНО / СТРО ИМ	gate_mode_report.py; ADR-0273	09.10.2026	Ворота P2a (31.10), INDRA (12.11), -- six (14.12)

Таблица 1.2 — Реестр прогнозов главы 1

№	Показатель	2026 (источник)	Ветвь А / В / С к 2035	Репер	Разрешение
Г1	Вычисления обучения фронта, рост в год	$\times 5$ с 2020 г., удвоение за 5,2 мес. (Epoch AI, 05.02.2026)	А: $\approx \times 2/\text{год}$ ($\times 1\,000$ к 2035) / В: $\times 4\text{--}5/\text{год}$ до 2030 ($\times 250\text{--}600$), затем $\approx \times 2/\text{год}$ ($\approx \times 10\,000$ к 2035) / С: упор в энергию	Epoch AI (data/trends)	Ежегодно, 1 февраля
Г2	Цена токена фиксированного качества	$\div 10$ в год (a16z, 12.11.2024; Epoch, 12.03.2025: $\div 9 \dots \div 900$)	А: замедление / В: $\div 10/\text{год}$ ($\div 10^4$ к 2030, не менее $\div 10^6$ к 2035 — $\div 10^9$ при сохра-	Epoch AI (inference prices)	Ежегодно

№	Показатель	2026 (источник)	Ветвь А / В / С к 2035	Репер	Разрешение
			нении темпа; качество 2026 бесплатно) / С: $\div 100/\text{год}$		
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)	А: месяц к 2035 / В: неделя–месяц к 2027–2028, квартал–год к 2027–2030 (удвоение раз в 3–7 мес.), дальше — вне нынешней шкалы / С: 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.	А: после 2035 г. / В: 2028–2030 гг. / С: до 2028 г.	blog.aifut-ures.org	Ежеквартально
Г1.5	Отставание открытых весов от закрытых	< 1 года (IAISR); ≈ 4 мес. (Epoch, сер. 2026)	А: 0 мес. / В: ≈ 6 мес. / С: отрыв закрытых за счёт энергии	IAISR; Epoch ECI	Ежегодно (ветвь А: $>50\%$ токенов)
Г7–Г8	Доля автономных решений и Guardian Agents	0 % (2024) $\rightarrow$ 15 % к 2028; стражи 10–15 % к 2030 (Gartner)	А: 10 % (стражи 5 %) / В: 30 % (стражи 15 %) / С: $>50\%$ (стражи 25 %)	Gartner Agentic AI	2028, 2030, 2035

Источники Главы 1

1. Внутренние артефакты: MITR/core/verification/z3_gate.py; MITR/policies/contracts/ (19); scripts/gate_mode_report.py, smt_latency_paper.py, verify_chain_standalone.py, publish_verify_kit.py, mitr_vocab.py; MITR_STRATEGY.md (§2a, §4); V_ICR_TRIZ.md; ADR-0262, 0273, 0275, 0280–0282; registry/product-line.json; каталог RLM (стр. 19, 34, 46, 79, 108–109, 182, 185, 226–227); обзор ГАММА-А (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/agent-ai-security-incidents/); Aviatrix; Gartner (gartner.com, пресс-релизы 25.06.2025 и 11.06.2025 — Guardian Agents); Bloomberg (bloomberg.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).

Приложения

Приложение Б. Команды независимого повтора

Команды ниже повторяют замеры, на которые ссылаются главы, и печатают те же числа, что напечатаны в книге. Они исполняются на узле шлюза из каталога установки с правами служебного пользователя; команды отдельных глав приведены в их разделах «Протокол независимой верификации» (раздел 7 каждой главы).

```
# живость служб шлюза
systemctl is-active v-bot v-mcp
# шесть базовых положений и строка по каждому арендатору —
# за сутки среза книги (без --date — последние сутки)
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --six --date 2026-10-09
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --tenants --date 2026-10-09
# цена решения (P4): вторая строка — замер 26.09.2026 (ADR-
0282), последняя — свежий замер
sed -n 2p /opt/v/data/logs/smt_latency_paper.jsonl; tail -1
/opt/v/data/logs/smt_latency_paper.jsonl
# последний якорь цепи квитанций шлюза (P3)
tail -1 /opt/v/data/gate_anchors.jsonl
# повтор всех решений цепи проверщиком для постороннего
(14.10.2026: 4958 из 4958)
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
# первое правило словаря, принятое самим MITR (C1): квитан-
ция 305
grep -m1 'seq.: 305,' /opt/v/data/logs/gate_actions.jsonl
```

Читателю без доступа к узлу открыт набор проверки asi5.ru/console/verify/ (его выкладывает скрипт `publish_verify_kit.py`): якоря цепей, их метки времени

OpenTimestamps, которые проверяются против сети Bitcoin чужим клиентом, и проверщик целости цепи. Повтор самих решений требует цепи квитанций и текстов правил в формате SMT-LIB: их аудиторы передают вместе с проверщиком `verify_chain_standalone.py`, и тот повторяет каждое решение любым решателем на собственной машине — стандартная библиотека Python, ни строки кода MITR. Проверщик с повтором выходит в открытый набор вместе со статьёй; до тех пор повтор выполняла только станция MITR.