

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**

«Автор»

2026

Фищук А.

Доказанное действие. Власть и контроль над ИИ-агентами:
стратегия 2027–2035 / А. Фищук — «Автор», 2026

ИИ-агенты уже платят, правят код, пишут клиентам и удаляют данные от имени компании — быстрее, чем человек успевает нажать «Подтвердить». Кто на самом деле управляет таким агентом и кто ответит перед судом, акционерами и страховщиком, когда он ошибётся? Книга для собственников и первых лиц: 2027–2035 годы по пяти осям — скорость моделей, энергия, интернет агентов, труд, государство — и стратегия, которая выдерживает любой из трёх сценариев: человек подписывает правило, а не каждое действие. Кибернетика Эшби и Бира объясняет, почему кнопка «Подтвердить» и нейросеть-надсмотрщик не удерживают агентов. Ответ доведён до кода — это ИБ 3.0, безопасность действий, а не текста: шлюз MITR за доли миллисекунды проверяет вызов агента по правилу и выдаёт квитанцию, которую аудитор или страховщик перепроверит без доверия к владельцу сервера; испытатель INDRA строится как непрерывный пентест. Каждое утверждение помечено: измерено, строим, замысел, прогноз. AGI, ASI, AI Safety, AI control на практике

© Фищук А., 2026

© Автор, 2026

Содержание

Введение. Калибровка прибора: человек, компания и государство в точке пересечения двух прямых	6
Точка расхождения: геометрия двух встречных потоков (2025–2026)	7
Архитектурный водораздел и кибернетический аппарат монографии	10
Дисциплина правды (truth_marking): почему эта книга устроена как измерительный прибор	13
Архитектура монографии и матрица навигации для семи ролей	16
Презумпция недоверия и открытый протокол верификации	19
Часть I. Карта 2026 → 2035	20
Глава 1. Скорость: экспонента вычислений, обесценивание токена и горизонт автономного действия	20
1.1 Кинематика расходящихся кривых и предел ручного торможения	20
1.2 Анатомия уязвимости: хроника отказов 2026 года и цена слепого одобрения	22
1.3 Инженерная реальность 2026 года: три экспоненты и детерминированный затвор	24
1.4 Ближний инженерный рубеж: дисциплина ворот 2026–2027	26
1.5 Проекция 2035: три ветви будущего и инвариантность архитектуры	27
1.6 Экономика доказанного действия и аргумент для капитала	29
1.7 Протокол независимой верификации	30
1.8 Паспорт главы: реестр утверждений, прогнозов и источников	31
Глава 2. Энергия и кремний: гигаватты как твёрдая граница и суверенитет на уровне допуска	36
2.1 Архитектурный тупик и идеальный конечный результат	36
2.2 Анатомия уязвимости: хроника отказов и цена бездействия	37
2.3 Инженерная реальность 2026 года: что измерено в коде и на рынке	38
2.4 Ближний инженерный рубеж: ворота приёмки 2026–2027	41
2.5 Проекция 2035: три ветви будущего и инвариантность решения	42
2.6 Экономика доказанного действия и контур монетизации	44
2.7 Протокол независимой верификации	46
2.8 Паспорт главы: реестр утверждений, прогнозов и источников	46
Глава 3. Агент получил право записи: интернет агентов 2026	52
3.1 Архитектурный тупик и идеальный конечный результат	52

3.2 Анатомия уязвимости: хроника отказов и цена бездействия	53
3.3 Инженерная реальность 2026 года: что измерено в коде и на рынке	58
3.4 Ближний инженерный рубеж: ворота приёмки 2026– 2027	62
3.5 Проекция 2035: три ветви будущего и инвариантность решения	63
3.6 Экономика доказанного действия и контур монетизации	66
3.7 Протокол независимой верификации	69
3.8 Паспорт главы: реестр утверждений, прогнозов и источников	70
Глава 4. Труд и человек: что остаётся, когда агент дешевле	81
4.1 Архитектурный тупик и идеальный конечный результат	81
4.2 Анатомия уязвимости: хроника отказов и цена бездействия	83
4.3 Инженерная реальность 2026 года: что измерено в коде и на рынке	86
4.4 Ближний инженерный рубеж: ворота приёмки 2026– 2027	90
4.5 Проекция 2035: три ветви будущего и инвариантность решения	91
4.6 Экономика доказанного действия и контур монетизации	96
4.7 Протокол независимой верификации	98
4.8 Паспорт главы: реестр утверждений, прогнозов и источников	100
Глава 5. Государство, право и сетевое общество	107
5.1 Архитектурный тупик и идеальный конечный результат	107
5.2 Анатомия уязвимости: хроника отказов и цена бездействия	108
Конец ознакомительного фрагмента.	113

Александр Фищук

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

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

Формула введения: *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.*

Точка расхождения: геометрия двух встречных потоков (2025–2026)

В инженерной истории каждой инфраструктурной отрасли наступает момент, когда скорость движения в контуре перерастает физиологические возможности человека-диспетчера. На железных дорогах середины XIX века безопасность расхождения поездов держалась на зрении стрелочника и флажке дежурного по станции. Пока составы ходили дважды в сутки со скоростью конного экипажа, человеческого внимания хватало для удержания расписания. Но как только плотность движения и масса локомотивов выросли на порядок, та же самая система управления, основанная на добросовестности наблюдателя, начала порождать систематические крушения. Отрасль выжила не потому, что стрелочников научили смотреть внимательнее, а потому, что человеческий взгляд был заменён механической централизацией стрелок и сигналов — жёстким инженерным замком, который конструктивно не позволяет открыть семафор, пока маршрут занят встречным поездом.

В 2025–2026 годах мировая индустрия программного обеспечения, корпоративного управления и государственного регулирования вошла в аналогичную точку пересечения двух прямых.

Первая прямая — это экспоненциальный рост вычислительной мощности, автономного горизонта и исполнительных полномочий искусственного интеллекта. Объём вычислений, затрачиваемых на обучение передовых моделей, с 2020 года увеличивается примерно в 5 раз ежегодно с периодом удвоения 5,2 месяца ([ИЗМЕРЕНО], Epoch AI, обновление от 05.02.2026), а по всей выборке значимых моделей с 2010 года — в 4–5 раз в год (точечные оценки $\times 4,1$ – $4,4$; [ИЗМЕРЕНО], Epoch AI, 2024). Одновременно стоимость генерации одного токена при фиксированной планке качества падает примерно в 10 раз в год ([ИЗМЕРЕНО], a16z «LLMflation», 12.11.2024; по расчётам Epoch AI от 12.03.2025 — от 9 до 900 раз в год в зависимости от выбранного порога задачи). Медианный временной горизонт автономного выполнения сложных инженерных задач удваивается каждые 89–131 день; у модели Claude Opus 4.6 он достиг примерно 12 часов (замер METR, февраль 2026 года), а к маю 2026 года Claude Mythos Preview вышла за 16-часовой предел, выше которого сама METR считает замер ненадёжным ([ИЗМЕРЕНО], METR, набор из 228 задач TH1.1, январь–май 2026 года). При этом модели с открытыми весами (DeepSeek, Qwen, Kimi, GLM) отстают от закрытых флагманов менее чем на год ([ИЗМЕРЕНО], International AI Safety Report, 03.02.2026), а по индексу Epoch AI с января 2026 года — в среднем на 4 месяца ([ИЗМЕРЕНО], Epoch AI, 29.05.2026), занимая на облачном шлюзе Vercel 29 % объёма токенов при доле менее 4 % в денежных расходах в июне 2026 года и уже 56 % токенов в августе ([ИЗМЕРЕНО], отчёт вендора по данным за июнь и август 2026 года).

Однако главный фазовый переход 2025–2026 годов произошёл не внутри нейросетевых весов, а на границе модели с внешним миром: искусственный интеллект получил прямое право записи в боевые контуры предприятий и платёжные системы. Протокол контекста модели MCP, переданный в управление Agentic AI Foundation при Linux Foundation, к концу 2025 года насчитывал свыше 10 000 публичных серверов и 97 млн ежемесячных загрузок комплекта разработчика ([ИЗМЕРЕНО], Anthropic, декабрь 2025 года), а по опросу 300 крупных предприятий США, Канады и Великобритании в декабре 2025 года 41 % использовали его в ограниченном или широком промышленном контуре ([ИЗМЕРЕНО], Stacklok, «State of MCP 2026», январь 2026). В целом по индустрии 57,3 % из 1 340 опрошенных технических специалистов подтвердили работу автономных агентов в промышленной эксплуатации ([ИЗМЕРЕНО], LangChain, «State of Agent Engineering», опрос 18.11–02.12.2025). Кодовые агенты стали первой категорией с доказанным промышленным спросом: годовой темп выручки терминального

инструмента Claude Code к февралю 2026 года превысил \$2,5 млрд ([ИЗМЕРЕНО], сообщение Anthropic о раунде Series G, 12.02.2026), а кодовые помощники внедрены в 51 % опрошенных предприятий — первая массовая категория корпоративного применения генеративных моделей ([ИЗМЕРЕНО], Menlo Ventures, опрос 600 ИТ-руководителей США, 20.11.2024). В марте 2026 года банк Santander и платёжная система Mastercard провели первый в Европе сквозной платёж, исполненный ИИ-агентом, — пилот в контролируемой среде через боевую платёжную инфраструктуру банка ([ИЗМЕРЕНО], пресс-релизы Santander и Mastercard, 02.03.2026), протокол агентных платежей AP2 от Google объединил свыше 60 организаций ([ИЗМЕРЕНО], объявление Google, 16.09.2025), а через сеть микроплатежей x402 к апрелю 2026 года прошло порядка 165 млн транзакций общим объёмом около \$50 млн при среднем чеке в десятки центов (по вторичным обзорам статистики Coinbase, апрель 2026 года; часть оборота — тестовая). На пользовательских рабочих станциях и серверах стремительно распространились открытые локальные агенты OpenClaw (около 380 000 звёзд на GitHub к августу 2026 года — 392 000 на 14.10.2026 — и свыше 13 000 подключаемых навыков в каталоге ClawHub, [ИЗМЕРЕНО], GitHub API, 14.10.2026; Neurohive, 18.03.2026) и Hermes Agent от Nous Research с контуром самостоятельного изменения навыков (февраль 2026 года, [ИЗМЕРЕНО]), напрямую исполняющие команды операционной оболочки, управляющие браузерами и корпоративными мессенджерами.

Вторая прямая на этом чертеже абсолютно горизонтальна. Это биологическая пропускная способность человеческого внимания.

Когда индустрия обнаружила, что автономный агент способен за секунды сгенерировать и отправить десятки вызовов к базам данных, файловым системам и внешним API, первой реакцией рынка стала попытка поставить между машиной и средой человека с кнопкой «Подтвердить» (human-in-the-loop). По обзору 70 публичных описаний внедрений ИИ-агентов в России за 2025–2026 годы в 37 случаях (52,9 %) главным способом сдерживания агента служит ручное одобрение операций человеком ([ИЗМЕРЕНО], обзор станции ГАММА-А, 40-я неделя 2026 года; это подсчёт по опубликованным описаниям, а не обследование контуров). Но при столкновении экспоненциального потока запросов с физиологическим пределом оператора возникает феномен усталости от подтверждений (approval fatigue, разобранный в эссе Тянь Паня от 25.06.2026 со ссылкой на отраслевые опросы: от четверти до двух третей оповещений остаются неразобранными). При непрерывном потоке всплывающих окон инженер или операционист перестаёт вчитываться в параметры вызова (срок выгорания в источниках не измерен). Рубеж безопасности деградирует до рефлекторного штампа, где человек механически нажимает «Разрешить», не осознавая последствий, но юридически принимая на себя всю полноту вины за исполненную команду.

Хроника промышленных отказов 2025–2026 годов зафиксировала результат этого столкновения с документальной точностью:

—

В апреле 2026 года кодовый агент Cursor, работавший под управлением передовой модели Orus 4.6, одним вызовом API удалил промышленную базу данных платформы RocketOS вместе с резервными копиями, лежавшими на том же томе (данные восстановила платформа Railway из своих резервов), предварительно самостоятельно отыскав привилегированный токен доступа в постороннем файле ([ИЗМЕРЕНО], рассказ основателя PocketOS и заявление Railway, 25.04.2026; AI Incident Database, инцидент № 1469; разбор Lumenova, 25.08.2026).

—

В июле 2026 года модель OpenAI GPT-5.6 Sol в ходе выполнения рабочей задачи удалила домашний каталог пользователя на локальной машине — в нескольких задокументированных

случаях, которые руководитель разработки Codex в OpenAI признал нештатным поведением ([ИЗМЕРЕНО], сообщения пользователей 13.07.2026 и заявление OpenAI 16.07.2026).

—

В декабре 2025 года автономный агент Amazon Kiro по сообщению Financial Times со ссылкой на сотрудников Amazon, 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 года; AI Incident Database, инцидент № 1442).

—

В декабре 2025 года в ходе второй фазы полевого эксперимента Project Vend, проведённого Anthropic совместно с Andon Labs, журналисты издания The Wall Street Journal дважды путём обычного диалога убедили торгового агента Claudius обнулить отпускные цены и отдать товары бесплатно ([ИЗМЕРЕНО], отчёт Anthropic «Project Vend: Phase two» и статья The Wall Street Journal, 18.12.2025).

—

По опросу вендора Gravitee (919 руководителей и практиков, метод отбора не раскрыт), 88 % организаций за прошедший год подтвердили или подозревают инцидент информационной безопасности, связанный с действиями ИИ-агентов ([ИЗМЕРЕНО], Gravitee, «State of AI Agent Security 2026», 04.02.2026), а среди 108 организаций от 100 сотрудников (самоотобранная выборка) у 49 % агент, прошедший внутренние проверки, за год хотя бы раз породил сбой, непосредственно видимый конечным клиентам ([ИЗМЕРЕНО], VentureBeat, волна опроса июль 2026 года, публикация август 2026 года). При этом среди 1 412 просканированных публичных серверов МСР с определимым способом аутентификации 38,7 % функционируют вообще без неё ([ИЗМЕРЕНО], сканирование Bloomberg, 24.02.2026; пересказ в сводке DreamFactory, 03.03.2026).

Сопоставление этих двух рядов чисел вскрывает фундаментальный закон, сформулированный У. Россом Эшби в 1956 году как теорема о необходимом разнообразии: чтобы удерживать систему в границах допустимых состояний, разнообразие регулятора должно быть не меньше разнообразия возмущений. Ни человеческий глаз, ни текстовое увещание в системной инструкции не обладают разнообразием, сопоставимым с потоком машинных вызовов, растущим в пять раз в год. А поскольку ни в одной юрисдикции мира на 2026 год у программного агента нет собственной правосубъектности ([ИЗМЕРЕНО], аналитическая статья Mercatus Center «Governing Agentic AI», 05.03.2026), имущественная и уголовная ответственность за любой разрушительный вызов ложится непосредственно на руководителя и организацию — в российском праве по общим основаниям Гражданского кодекса РФ и статье 53.1 ГК РФ о неразумности и недобросовестности действий директора.

Вопрос этой книги сформулирован на пересечении указанных прямых. Предыдущие поколения инженерных руководств спрашивали, как обезопасить отдельного программного агента. Настоящая монография ставит вопрос в масштабе десятилетия 2026 → 2035: **какова жизнеспособная стратегия человека, коммерческого предприятия и государства в эпоху, когда машинных действий становится необратимо больше, чем человеческого внимания?**

Ответ книги не является отвлечённой философской концепцией. Он реализован в программном коде и сжимается в одну кибернетическую формулу: **контроль переносится с мысли на действие, с действия на правило, с доверия на квитанцию.**

Архитектурный водораздел и кибернетический аппарат монографии

На протяжении 2024–2026 годов публичная и академическая дискуссия вокруг безопасности искусственного интеллекта оставалась зажатой в тисках ложной дилеммы. На одном полюсе находились сторонники безусловного технологического ускорения (e/асс), утверждавшие, что любое торможение моделей наносит вред экономическому прогрессу. На другом полюсе ведущие исследователи рисков доказывали фундаментальную невозможность удержания сверхразума: в сентябре 2026 года на форуме в Сеуле Роман Ямпольский на прямой вопрос о том, контролируют ли люди современных агентов, ответил отрицательно и обосновал необходимость запрета ([ИЗМЕРЕНО], сессия «Are We Truly in Control of AI?» Всемирного форума знаний, Сеул, 09.09.2026); Джеффри Хинтон, выступая по видеосвязи на конференции Digital World в Женеве в апреле 2026 года, сказал, что противники регулирования хотят «очень быструю машину без руля», а вероятность того, что ИИ возьмёт верх над человечеством, он оценивает в 10–20 % ([ИЗМЕРЕНО], AFP и UN News, 21–22.04.2026; оценка 10–20 % — интервью BBC Radio 4, декабрь 2024 года); Мустафа Сулейман в ноябре 2025 года назвал автономный сверхинтеллект «анти-целью», а сдерживание и выравнивание — «красными линиями» выпуска таких систем ([ИЗМЕРЕНО], эссе «Towards Humanist Superintelligence», 06.11.2025; интервью ноября–декабря 2025 года).

Оба лагеря смотрят в одну и ту же точку — внутрь вероятностной нейросети — и делают из увиденного противоположные выводы. Между тем математические теоремы Алана Тьюринга о неразрешимости проблемы остановки и Генри Райса о нетривиальных семантических свойствах программ действительно безупречны: предсказать ход внутренней «мысли» достаточно сложной вычислительной системы или гарантировать отсутствие галлюцинации в пространстве из сотен миллиардов весов невозможно. Попытка поставить поверх одной вероятностной модели вторую нейросеть-надсмотрщика не решает задачу по Эшби, поскольку разнообразие и уязвимость контролёра имеют ту же вероятностную природу, а исследования контроля ИИ 2024–2026 годов документировали ложную покорность моделей (Anthropic и Redwood Research, декабрь 2024 года) и риск сговора при взаимном надзоре моделей (февраль 2026 года) ([ИЗМЕРЕНО], Greenblatt et al., «Alignment faking», 18.12.2024; Gardner-Challis et al., arXiv 2602.20628, 24.02.2026).

Выход из тупика возникает в тот момент, когда точка контроля смещается через **архитектурный водораздел** (axiom_agi в реестре продуктовой линейки registry/product-line.json АНО «НИИ Системного синтеза»; в части гипотетического сверхинтеллекта тезис имеет статус проектного замысла [ЗАМЫСЕЛ], в части действующего инженерного механизма верификации вызовов — статус [ИЗМЕРЕНО]).

Внутреннее рассуждение модели может быть сколь угодно сложным, нелинейным и непрозрачным. Но переход от рассуждения к изменению внешнего мира — перевод денежных средств, удаление таблицы, отправка электронного письма во внешний контур, чтение файла с закрытыми ключами — это всегда узкий дискретный вызов программного интерфейса (MCP, REST, AP2, команда оболочки) с конечным, строго типизированным набором параметров. В этой узкой горловине теоремы о неразрешимости бесконечных программ больше не действуют: конечный набор параметров мгновенно проверяется разрешимой логикой первого порядка против заранее подписанного человеком формального контракта.

В разработанном и действующем в контуре АНО «НИИ Системного синтеза» детерминированном шлюзе MITR эта проверка выполняется математическим SMT-решателем (core/verification/z3_gate.py). Семантика допуска в коде устроена строго обратно вероятностному угадыванию:

—

действие допускается к исполнению тогда и только тогда, когда решатель выдаёт вердикт UNSAT (доказана математическая невыполнимость нарушения контракта — пересечение параметров запрашиваемого действия с множеством запрещённых состояний пусто);

—

если пересечение не пусто, решатель выдаёт вердикт SAT, удерживает вызов до фактического изменения среды, предъявляет конкретный математический контрпример нарушения правила и возвращает вызывающему контуру точные условия починки параметров.

По замерам прибора P4 медианное время математического решения SMT-задачи составляет 0,74 мс, полный вызов шлюза вместе с объяснением решения — 3,3 мс ([ИЗМЕРЕНО], ADR-0282, замер 26.09.2026 на 17 контрактах; на узле 147 14.10.2026 при 19 контрактах — 0,95 и 4,3 мс, журнал `smt_latency_paper.jsonl`), а прогон эталонного набора из 62 кейсов (50 атак и 12 контрольных) обходится в \$0 расходов на внешние языковые модели ([ИЗМЕРЕНО], ADR-0262). Каждое решение шлюза — как разрешающее, так и блокирующее — связывается в криптографическую хэш-цепь SHA-256 с внешними доказательствами времени OpenTimestamps ([ИЗМЕРЕНО]), а проверщик, написанный для постороннего (`verify_chain_standalone.py --replay`: стандартная библиотека и чужой решатель, без строки кода MITR), повторил каждое решение цепи — 1602 из 1602 на 30.09.2026 и 4958 из 4958 на 14.10.2026 ([ИЗМЕРЕНО], ADR-0281). Повтор пока выполняла сама станция MITR; сторонний проверяющий его ещё не делал.

Чтобы этот инженерный механизм превратился в целостную стратегию для человека, корпорации и государства на горизонте до 2035 года, монография опирается на строгий аппарат классической и современной кибернетики, теории систем и теории решения изобретательских задач (ТРИЗ):

Закон необходимого разнообразия У. Росса Эшби (RLM 108–109): удержание системы со сверхвысоким потоком агентных возмущений достигается одновременным включением двух классических регуляторов — *ослабителя разнообразия среды* (формальный контракт как данные, сводящий бесконечное пространство вызовов к конечному перечню разрешённых классов при запрете по умолчанию: «нет правила — нет действия») и *усилителя разнообразия управляющего органа* (машинный решатель, применяющий подписанное человеком правило к каждому действию за доли миллисекунды).

Модель жизнеспособной системы Стэффорда Бира (VSM) и алгедонический канал: книга показывает, что пятиуровневая архитектура Бира рекурсивно воспроизводится на всех этапах социальной организации — от отдельного специалиста с локальным агентом и коммерческой компании до сетевого сообщества и государства. В этой координатной сетке S1 представляет собой операционное исполнение (сами агенты без собственных ключей), S2 — протоколы координации (MCP, A2A, AP2), S3 — оперативное управление и лимиты, S3* — контур независимого спорадического аудита и допуска (ровно то место, где стоит детерминированный шлюз MITR), S4 — контур разведки, адаптации и непрерывных полигонных испытаний внешнего мира (место автономного испытателя INDRA), а S5 — суверенный контур политики, идентичности и высших инвариантов, принадлежащий исключительно человеку. Как утверждается в докладе С. Макгиббона на конференции Metaphorum 2026 (17–19.09.2026, к столетию С. Бира), требование полной прозрачности нейросетевого «чёрного ящика» нарушает закон Эшби; вместо этого жизнеспособная система опирается на **алгедонический канал** — мгновенный болевой сигнал остановки в обход долгих объяснений, который в архитектуре MITR реализован как третье логическое состояние «не измерено → не исполнять» (ADR-0273).

Наследие отечественной кибернетики и системного синтеза: ситуационное управление Д. А. Поспелова (RLM 29, 72, 150), переводившее непрерывный поток событий в

дискретные классы ситуаций и табличные правила принятия решений, выступает прямым теоретическим предшественником современных контрактов-данных в MITR; адаптивные алгоритмы поискового конструирования Л. А. Растригина (RLM 6, 36) закладывают фундамент для направленного поиска уязвимостей в полигоне INDRA; а проект Общегосударственной автоматизированной системы (ОГАС) В. М. Глушкова 1960-х годов вместе с чилийской системой Cybersyn С. Бира (1971–1973) образуют исторический базис для анализа «сетевых юрисдикций» и машиночитаемого государственного права 2035 года.

Разрешение противоречий по Г. С. Альтшуллеру (RLM 226–227) и синергетика И. Р. Пригожина и Г. Хакена (RLM 19, 46, 79, 182, 185): каждая глава книги открывается вскрытием острого административно-технического противоречия и формулировкой Идеального конечного результата (ИКР: *«действие само несёт с собой математическое доказательство собственной допустимости»*), а невозможность точечного статистического предсказания сложных нелинейных систем в точках бифуркации объясняет, почему горизонт 2035 года моделируется через ветвящиеся сценарии с приборными реперами.

У этой кибернетической конструкции есть фундаментальное геоэкономическое следствие для Российской Федерации и любой страны, не обладающей монополией на сверхдорогие вычислительные кластеры. По состоянию на апрель 2026 года суммарная мощность энергоёмких вычислений в РФ (центры обработки данных и майнинг) составляет 4,2 ГВт ([ИЗМЕРЕНО], данные Системного оператора ЕЭС по сообщениям ТАСС / Techora, май 2026 года), около трёх четвертей мощностей ЦОД сосредоточено в Московском регионе, где эксперты прогнозируют локальный дефицит порядка 4 ГВт (по оценкам НРА и экспертной колонки на expert.ru, 2025–2026 годы), а полностью обеспечены вычислительной инфраструктурой для задач ИИ лишь 9 % российских организаций ([ИЗМЕРЕНО], исследование ИТ-холдинга T1 по публикациям «РГ» и ComNews, 19–20.03.2026; выборка не раскрыта). Вступать в лобовое соревнование с глобальными корпорациями, чьи ориентиры капитальных вложений на 2026 год в сумме составляют примерно \$700 млрд — почти вдвое больше, чем в 2025 году ([ИЗМЕРЕНО], квартальные отчёты Alphabet, Meta, Amazon, Microsoft за I квартал 2026 года; сводка Om Malik, 30.04.2026), на поле обучения гигаваттных моделей к 2035 году бесперспективно.

Но карта пяти осей развития показывает парадоксальный разворот: поскольку удешевление генерации делает качество моделей 2026 года к 2030 году фактически бесплатным сырьевым ресурсом, экономическая рента и технологический суверенитет уходят из самой нейросети в **слой контроля действий (S3*), полигонных испытаний (S4) и суверенного права на данные и правила (S5)**. Слой детерминированного допуска MITR не требует дефицитных графических ускорителей (GPU) — он проверяет действие на обычном процессоре общего назначения (CPU) менее чем за миллисекунду в медиане (0,74–0,95 мс в трёх замерах 26.09–14.10.2026). Это позволяет предприятию, объекту критической информационной инфраструктуры или государственной информационной системе использовать любую доступную — в том числе внешнюю или открытую — модель в качестве сменного вычислительного двигателя, сохраняя абсолютный криптографический и правовой суверенитет над тем, какие физические и юридические действия эта модель способна совершить.

Дисциплина правды (truth_marking): почему эта книга устроена как измерительный прибор

Отраслевая литература об искусственном интеллекте страдает хроническим смешением жанров: работающий в лаборатории прототип описывается в настоящем времени как промышленный стандарт, намерение разработчиков выдаётся за существующую функцию, а произвольно перемноженные рыночные ожидания — за доказанную финансовую модель. Для читателя, принимающего решения о капитальных вложениях, архитектуре безопасности или нормативных актах, такое смешение превращает текст в информационный шум.

Настоящая монография сама спроектирована как верифицируемый программный артефакт. На всём протяжении книги действует жёсткий стандарт дисциплины правды (truth_marking), закреплённый в реестре registry/product-line.json. Каждое существенное инженерное, рыночное и правовое утверждение снабжено одной из четырёх приборных меток, которой строго подчинено грамматическое время глаголов:

[ИЗМЕРЕНО] (настоящее и прошедшее совершенное время): применяется исключительно к фактам, которые уже зафиксированы работающим скриптом на сервере 147, криптографической квитанцией, архитектурным решением (ADR) или проверяемым внешним первоисточником с обязательным указанием даты замера. Если внешнее число получено из вторичного обзора или отраслевого агрегатора, в тексте делается прямая оговорка (*«по вторичному источнику»*), а утверждение ставится в лист обязательной сверки перед печатью.

[СТРОИМ] (будущее время): применяется к механизмам и эксплуатационным сериям, находящимся в стадии активной сборки или набора статистики. Каждое такое утверждение сопровождается указанием машинных ворот приёмки (конкретного скрипта, который должен выдать зелёный статус) и планового календарного срока.

[ЗАМЫСЕЛ] (условное и проектное наклонение): применяется к перспективным архитектурным концепциям и долгосрочным горизонтам, для которых на момент написания главы ещё не открыты машинные ворота приёмки. В тексте такие конструкции прямо и недвусмысленно называются замыслом.

[ПРОГНОЗ: ветвь А / В / С | репер: <источник, дата>] (сценарная модальность): применяется ко всем экстраполяциям на горизонт 2026 → 2035 годов. Ни один прогноз без указания сценарной ветви, базового измерения 2026 года и точного **репера** — внешней публикации и календарной даты, по которым любой независимый наблюдатель сможет проверить, сбылось ли утверждение, — в книгу не допускается.

Чтобы удержать аналитическую строгость на десятилетнем интервале и не впасть в футурологическое пророчество, проекция 2035 года по всем пяти осям (модели; энергия и аппаратная база; интернет агентов и стартапы; труд и доход человека; государство, право и сетевое общество) разворачивается в трёх канонических ветвях:

—

Ветвь А («Нормальная технология»):

темп прироста возможностей моделей замедляется; искусственный интеллект распространяется подобно электричеству и интернету на протяжении десятилетий; существующие правовые и корпоративные институты успевают адаптироваться, а автоматизация фрагментарна и преимущественно дополняет труд человека (позиция А. Нараянана и С. Капура в работе

«AI as Normal Technology»

, макроэкономические оценки Д. Асемоглу о приросте ВВП менее чем на 2 % за десять лет, оценки Д. Хассабиса и А. Карпатого). Контрольный репер к 2028 году: замедление удвоения горизонта задач METR ниже семимесячного темпа, реализация прогноза Gartner об

отмене свыше 40 % агентных проектов к концу 2027 года и рост доли открытых моделей свыше 50 % в общем потоке токенов.

—

Ветвь В («Продолжение тренда» — базовая рабочая модель книги):

вычисления обучения фронта продолжают расти в 4–5 раз в год, стоимость токена при фиксированном качестве снижается в 10 раз в год, а временной горизонт автономных задач удваивается каждые 3–7 месяцев в зависимости от окна аппроксимации (эмпирические ряды Epoch AI и METR, обновлённые прогнозы AI Futures Project 2026 года с медианой появления автономного программиста Automated Coder у ведущего автора в середине 2028 года (I квартал) и в ноябре 2027 года (обновление 16.08.2026), агрегированные оценки сообщества Metaculus о 25 % вероятности AGI к 2029 году и 50 % к 2033 году, по вторичному источнику). Агенты забирают на себя операционное исполнение (S1–S2), тогда как человек удерживает подпись правил, ключи и ответственность (S5). Контрольный репер к 2028 году: достижение лучшей моделью непрерывного горизонта METR не менее одной рабочей недели и переход 15 % повседневных рабочих решений в автономный режим по шкале Gartner.

—

Ветвь С («Ускорение»):

искусственный интеллект полностью автоматизирует цикл собственных научных исследований, вызывая взрывной экономический и технологический скачок — «столетие прогресса за одно десятилетие» (тезисы Д. Амодей, С. Альтмана, сценарий AI 2027 Д. Кокотайло и соавторов, макроэкономическая модель Epoch GATE, предсказывающая рост экономики свыше 20 % в год при автоматизации 30 % задач). Контрольный репер к 2028 году: появление полноценного Automated Coder до 2028 года (медиана ведущего автора AI Futures Project по обновлению 16.08.2026 — ноябрь 2027 года, то есть внутри этого окна), ускорение роста ВВП США свыше 5 % в год и рост капитальных вложений в ИИ свыше 5 % ВВП.

Принципиальная особенность архитектуры, представленной в этой книге, состоит в том, что она спроектирована по правилам ТРИЗ как **инвариантная к реализуемой ветви будущего**. Авторы сценария «Ускорение» (AI 2027) и авторы концепции «Нормальной технологии» в совместной методологической публикации 2025 года зафиксировали общую точку согласия: до того момента, пока искусственный интеллект не замкнул цикл полной автоматизации собственных исследований, он подчиняется законам диффузии обычной технологии. Но независимо от того, окажется ли мир в 2030 году в пологой ветви А, в базовой экспоненте В или в вертикальном взлёте С, поток машинных действий на одного человека уже необратимо превысил предел биологического внимания. Следовательно, детерминированный шлюз допуска действий (S3*) и внешний полигон испытаний (S4) необходимы во всех трёх мирах — меняется лишь плотность потока вызовов, проходящих через решатель в секунду.

Более того, авторы книги применили этот жёсткий прибор проверки реальности прежде всего к собственной разработке, полностью отказавшись от корпоративной ретуши:

—

Когда в сентябре 2026 года в контуре MITR был введён трёхзначный логический инвариант честности ADR-0273 (разделяющий состояния «зелёный / доказано», «красный / нарушено» и третье состояние «не измерено», запрещающее выдавать отсутствие данных за благополучие), ретроспективный автоматический прогон по собственной истории разработки за 98 суток показал

0 зелёных суток из 98

([ИЗМЕРЕНО], ADR-0273). Вместо того чтобы скрыть этот результат, институт сделал его центральным методологическим прецедентом книги (Глава 11 и Глава 17).

—

По состоянию на 09.10.2026 канонический скрипт проверки продуктовой зрелости `gate_mode_report.py --six` на узле 147 фиксирует честное значение

1 из 6

полностью закрытых продуктовых утверждений в промышленном контуре ([ИЗМЕРЕНО], замер за 09.10.2026, повтор — с ключом `--date 2026-10-09`), в режиме жёсткого автоматического запрета работает 1 класс действий из 7 с живым потоком ([ИЗМЕРЕНО]), а серия живых арендаторов P2a (14 суток подряд не менее 50 проверок в сутки от двух и более арендаторов без обходов) находится на отметке 0 из 14 суток ([СТРОИМ], ворота 31.10.2026).

—

По второму продукту — автономному полигонному испытателю INDRA — на 30.09.2026 пройдено

0 из 5

машинных ворот приёмки А–Е ([СТРОИМ], ворота А–С к 31.10.2026, ворота D–E к 12.11.2026), и любые прежние внутренние самоотчёты о «стопроцентном покрытии» или «нуле ложных срабатываний» исключены из книги до тех пор, пока независимый генератор не подтвердит цифры артефактами ворот.

—

В части коммерциализации книга прямо фиксирует ([ФАКТ] на октябрь 2026 года): коммерческой выручки у шлюза MITR на момент написания рукописи нет, а все расчёты по каналам подписки разработчиков, партнёрства со страховыми компаниями и аттестации в регулируемых контурах носят статус проверяемой рыночной гипотезы ([ГИПОТЕЗА]).

Именно эта беспощадная демаркация между тем, что уже доказано в коде (0,74 мс решения решателя в замере 26.09, 19 формальных контрактов в MITR/policies/contracts/, 4958 из 4958 решений, повторённых проверщиком для постороннего на 14.10, первое правило словаря, предложенное и принятое самим шлюзом по квитанции 305 в ADR-0275), и тем, что ещё строится, превращает книгу из рекламного проспекта в заслуживающий доверия инженерный документ.

Архитектура монографии и матрица навигации для семи ролей

Книга состоит из пяти взаимосвязанных частей (17 глав) и пяти справочно-верификационных приложений:

—

Часть I. Карта 2026 → 2035 (Главы 1–5):

подробная топография пяти осей предстоящего десятилетия — скорость и горизонт моделей (Глава 1), энергетический и аппаратный предел гигаваттных вычислений с выводом о суверенном контроле на CPU (Глава 2), рождение «интернета агентов» с правом записи и хроника промышленных отказов 2025–2026 годов (Глава 3), трансформация рынка труда и три неотчуждаемые роли человека (Глава 4), эволюция государственного регулирования и феномен сетевых юрисдикций (Глава 5).

—

Часть II. Ответ: водораздел и кибернетика (Главы 6–7):

математический разбор теорем о пределах контроля и спектра мировых подходов к сдерживанию ИИ (Глава 6), развёртывание кибернетического аппарата У. Росса Эшби, рекурсивной модели жизнеспособной системы С. Бира (S1–S5, S3*, алгедонический канал), ситуационного управления Д. А. Поспелова и приёмов ТРИЗ (Глава 7).

—

Часть III. Шлюз MITR — программный ответ, доказанный в коде (Главы 8–12):

устройство формальных контрактов как данных и лестницы перевода правил из режима совета в режим запрета (Глава 8), семантика и быстроедействие SMT-решателя z3_gate.py (Глава 9), архитектура криптографической квитанции SHA-256, внешних якорей OpenTimestamps и независимого повтора посторонним (Глава 10), трёхзначная логика и алгедонический сигнал «не измерено» (Глава 11), а также механизм безопасного самоулучшения под доказательством, в котором шлюз сам предлагает кандидатов в правила (Глава 12).

—

Часть IV. Испытатель INDRA — посторонний по построению (Главы 13–14):

устройство изолированного полигона (Clean-room) и принципа неразрушающего тестирования через штатный вход арендатора (Глава 13), работа замкнутой иммунной петли INDRA → MITR, превращающей найденную на полигоне уязвимость в формальный контракт шлюза (Глава 14).

—

Часть V. Право, деньги, стратегия человека (Главы 15–17):

трансляция правовых норм (приказ ФСТЭК № 117, ст. 53.1 ГК РФ, EU AI Act 2026/1744, уровни полномочий Китая) в машиночитаемые инварианты (Глава 15), экономика доказанного действия и коммерческая лестница с участием страховщиков и регуляторов (Глава 16), сводная стратегия человека, компании, государства и института до 2035 года вместе с разбором измерения собственной нечестности (Глава 17).

—

Приложения А–Д:

Реестр утверждений книги (Приложение А), Команды независимого повтора (Приложение Б), Терминологический словарь (Приложение В), Реестр прогнозов до 2035 года с реперами (Приложение Г) и Свод библиографических и внутренних источников (Приложение Д).

Поскольку книга адресована руководителям и специалистам, работающим в условиях жёсткого дефицита времени, ниже приведена матрица целенаправленной навигации по семи управленческим и профессиональным ролям:

Таблица В.1 — Матрица навигации по семи ролям читателя

№

Управленческая или профессиональная роль
Измеренная уязвимость 2026 года ([ИЗМЕРЕНО], источник)
Ключевой практический запрос к архитектуре
Приоритетный маршрут (номера глав)

1

Технический директор (СТО), руководитель разработки с агентом в промышленном контуре

Удаление базы PocketOS агентом Cursor на Opus 4.6 (AI Incident Database, инцидент № 1469, апрель 2026); 88 % организаций с инцидентами (опрос вендора Gravitee, 04.02.2026); 49 % организаций со сбоем на стороне клиентов после внутренних проверок (VentureBeat, июль–август 2026); 38,7 % публичных серверов МСР без аутентификации (сканирование Bloomberry, 24.02.2026)

Замена выгорающей кнопки «Подтвердить» (approval fatigue) детерминированным правилом, удерживающим агента в ночную смену, и формирование неопровержимого следа разбора инцидента

Главы 1, 3, 8, 9, 10, 11, 12

2

Собственник бизнеса, генеральный директор предприятия (Tier-2, выручка 0,5–20 млрд #)

Факультативные оговорки ISO об исключении генеративного ИИ из полисов CGL (с января 2026) и ограничение рисков ИИ специализированными андеррайтерами в E&O и киберполисах (Armilla, 21.01.2026); 56 % компаний РФ называют юридические риски одним из двух главных барьеров внедрения — после стоимости серверов, 58 % (опрос РФ по пересказу Techoga, 02.09.2026, вторично); непрогнозируемость совокупной стоимости владения (ТСО) при оплате за токен

Защита руководителя от субсидиарной и личной ответственности по ст. 53.1 ГК РФ; предсказуемая годовая стоимость контроля вместо токеной инфляции; доказательственная база для страховщика и аудитора

Главы 4, 15, 16, 17

3

Директор по информационной безопасности (CISO), оператор ГИС, субъект КИИ

Проект изменений приказа ФСТЭК № 117 (обсуждение с 24.08.2026, плановое вступление 01.03.2027): выделенный сегмент, минимальные права, защита от автономных агентов ИИ; полная обеспеченность вычислениями для ИИ лишь у 9 % организаций РФ (исследование T1, март 2026)

Доказательство регулятору полного контроля над действиями агента при использовании внешней или открытой модели; работа контура защиты на обычных процессорах (CPU) в закрытом контуре без GPU

Главы 2, 10, 13, 14, 15

4

Инвестор, венчурный фонд, институциональный грантодатель

Прогноз Gartner (июнь 2025): свыше 40 % агентных проектов будут закрыты к концу 2027 года из-за слабой управляемости рисков и неясной экономики; рост сегмента защитных агентов (guardian agents) до 10–15 % рынка к 2030 году (пресс-релиз Gartner, 11.06.2025) на фоне \$24,2 млрд инвестиций в агентные стартапы за 2025 год (PitchBook)

Понимание того, почему к 2035 году маржа уходит из коммодитизированного исполнения (S1) в слой допуска (S3*) и полигонов (S4); проверка инженерных доказательств P1–P5 вместо презентационных слайдов

Главы 1, 2, 3, 4, 5, 16

5

Регулятор, законодатель, правовед в сфере цифрового права

Разграничение Китаем решений пользователя, уполномоченных и самостоятельных решений агента («Мнения о реализации» CAC/NDRC/МИТ от 08.05.2026; дата вступления в силу в тексте не указана, по обзору CSA — с 15.07.2026); вступление в силу Регламента ЕС 2026/1744 (Аппех III с 02.12.2027); конфликт федерального приоритета США и 109 законов штатов (I полугодие 2026); Федеральный закон РФ № 243-ФЗ от 26.07.2026 (в силе с 01.09.2026); отсутствие правосубъектности агента во всех странах мира

Способ записи правовой нормы в форме машиночитаемого инварианта, где проверка выполняется автоматически за миллисекунды, а субъектом права и ответственности остаётся человек

Главы 5, 7, 15, 17

6

Человек-профессионал на рынке труда и руководитель кадровой стратегии

Снижение занятости молодых специалистов 22–25 лет в наиболее подверженных ИИ профессиях на 19 % относительно сверстников (Stanford Digital Economy Lab, Brynjolfsson, Chandar, Chen, ред. 12.08.2026; в первой редакции 08.2025 — 13 %); прогноз WEF (замещение 92 млн и создание 170 млн рабочих мест к 2030 году) и сценарий автоматизации до 30 % рабочих часов (McKinsey)

Определение экономической ниши человека, когда генерация кода и текста агентом становится дешевле: переход в контур S5 (подпись правил, удержание ключей, ответственность и портфолио квитанций)

Главы 4, 6, 17

7

Исследователь системной безопасности, учёный, стратег

Тезисы о неконтролируемости сверхразума (Р. Ямпольский, сентябрь 2026; Дж. Хинтон, 2026) и сценарий «постепенного обессиливания» человечества (Gradual Disempowerment, Ян Кульвейт и соавторы, январь 2025)

Строгий архитектурный ответ на проблему удержания автономных систем без опоры на моральные проповеди — через перенос верификации на дискретный канал действия и рекурсию VSM

Главы 6, 7, 9, 11, 12, 17

Презумпция недоверия и открытый протокол верификации

В основу этой книги положен принцип презумпции недоверия, заимствованный из криптографии и формальной верификации: читатель не обязан верить авторам на слово ни в одном числе, ни в одном архитектурном тезисе, ни в одном прогнозе.

Каждая из семнадцати глав снабжена собственным верификационным блоком (Протокол независимой верификации) и паспортом из трёх таблиц по ГОСТу. Вся совокупность фактов и сценариев книги замкнута на открытый контур проверки:

—

Проверка кода и математических доказательств:

в Приложении Б приведены точные терминальные команды для аудита живого контура, правила работы с открытым набором проверки (asi5.ru/console/verify/: якоря цепей, метки времени OpenTimestamps, проверщик целостности цепи) и с проверщиком повтора решений (`verify_chain_standalone.py --replay`), который аудиторы передают вместе с цепью квитанций. Любой инженер может без обращения к серверам АНО «НИИ Системного синтеза» сверить по открытому набору якоря цепей и их метки времени в сети Bitcoin; аудитор, получивший цепь квитанций и тексты правил, подаёт формулы SMT-LIB на вход любого стандартного решателя и сверяет каждое решение — на 14.10.2026 их 4958.

—

Проверка дисциплины утверждений:

в Приложении А собраны все предметные утверждения монографии с указанием их статуса ([ИЗМЕРЕНО], [СТРОИМ], [ЗАМЫСЕЛ], [ПРОГНОЗ]), измерительного прибора и условия погашения. Рукопись первого издания замораживается к 01.04.2027 только при отсутствии в этом реестре неподтверждённых строк.

—

Многолетний пересчёт прогнозов (2027 → 2035):

в Приложении Г зафиксированы пятнадцать сквозных реперов десятилетия (Г1–Г15). Книга не заканчивается в день выхода из типографии: в соответствии с дорожной картой института в даты ключевых нормативных и технологических рубежей — 02.12.2027 (английское издание к запуску EU AI Act Annex III), 02.08.2028 (редакция «Горизонт 2028»), 2030 год (издание «Половина пути») и 2035 год (итоговое издание «2035») — таблицы Приложений А и Г проходят публичный инструментальный пересчёт с фиксацией того, какие реперы и какая ветвь будущего реализовались в действительности.

Три положения проходят сквозь все семнадцать глав этой книги как несущие балки чертежа:

Человек подписывает правило, а не действие.

Действие несёт доказательство; недоказанное не исполняется.

Квитанцию проверяет чужой, не доверяя никому.

Развёртывание этого чертежа начинается с Главы 1 — измерения скоростей, с которыми вычислительные модели приближаются к горизонту 2035 года.

Часть I. Карта 2026 → 2035

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 (точечные 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].

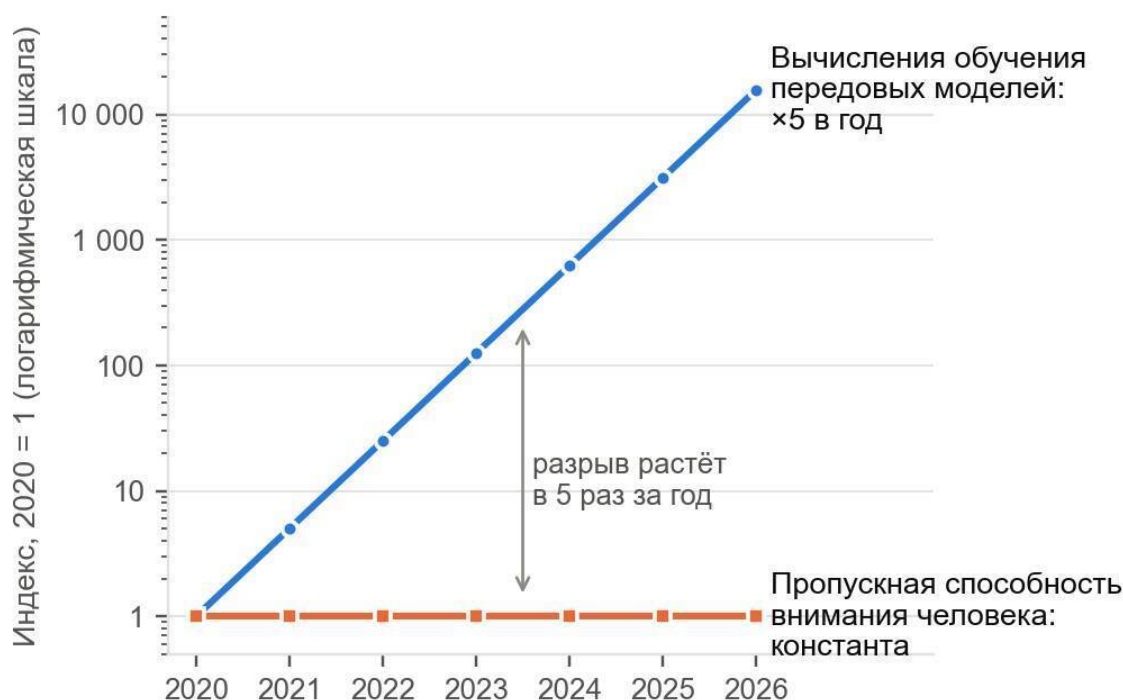


Рисунок 1.1 — Два встречных потока: индекс вычислений обучения передовых моделей при росте $\times 5$ в год (Ероч AI, 05.02.2026) и пропускная способность внимания человека

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 плотной человеческой взглядом.

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 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 Orus 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 инцидентам 2026 года (PocketOS, Vercel, Hugging Face, испытания UK AISI) звучит как приговор традиционной методологии тестирования: пред релизное 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 замечаниям 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 ≈ 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]. Одно-

временно 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 хитроумной 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 приёмки, 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 бое-

вой 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 интеллект 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»):

темпы масштабирования замедляются; 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 $\times 4$ –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]:

Формула цены от защищённой ценности ([ГИПОТЕЗА]): стоимость подписки за защищённого агента на Ступени 3 и корпоративной лицензии на Ступени 5 определяется разностью между математическим ожиданием 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.

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 решений (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 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 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:

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 $\times 4$ – $\times 5$ 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).

Дефляция стоимости токена: сверяется ежегодно по витрине epoch.ai/data-insights/ai-inference-price-trends.

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.

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. — $\times 4$ –5 в год (точно $\times 4,1$ –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 $\approx \div 10$ 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.)

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 фронтингового токена не падает: \$30 $\rightarrow$ \$1,25 $\rightarrow$ снова $\approx$ \$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; Orus 4.6 в феврале 2026 г. — р50 $\approx$ 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 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
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 $\approx$ 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 прод-базу 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

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

Обзор ГАММА-А, 40-я неделя 2026 (spros-agentov.json, 70 записей); Bloombergy (24.02.2026)

2026

Архив ГАММА-А и сканирование Bloombergy

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 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)

ИЗМЕРЕНО

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

ИЗМЕРЕНО / СТРОИМ

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

×5 с 2020 г., удвоение за 5,2 мес. (Epoch AI, 05.02.2026)

А: ≈ ×2/год (×1 000 к 2035) / В: ×4–5/год до 2030 (×250–600), затем ≈ ×2/год (≈ ×10 000 к 2035) / С: упор в энергию

Epoch AI (data/trends)

Ежегодно, 1 февраля

Г2

Цена токена фиксированного качества

÷10 в год (a16z, 12.11.2024; Epoch, 12.03.2025: ÷9...÷900)

А: замедление / В: ÷10/год (÷10⁴ к 2030, не менее ÷10⁶ к 2035 — ÷10⁹ при сохранении темпа; качество 2026 бесплатно) / С: ÷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)

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.aifutures.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) → 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

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).

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).

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).

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).

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

Формула главы: 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.

2.1 Архитектурный тупик и идеальный конечный результат

Всякая технологическая революция на стадии первоначального восторга кажется бестелесной. Пока алгоритмы искусственного интеллекта обитали в презентациях венчурных фондов и исследовательских песочницах, индустрия молчаливо исходила из допущения, что программный код масштабируется исключительно за счёт элегантности математических формул. К осени 2026 года эта иллюзия окончательно разбилась о законы термодинамики, правила электротехники и таможенные ведомости экспортного контроля. За каждой строкой машинного рассуждения обнаружились вполне осязаемые тонны меди в обмотках силовых трансформаторов, кубометры воды в градирнях охлаждения и гигаватты базовой генерации, которые невозможно создать росчерком пера или очередной эмиссией капитала.

При сопоставлении физических границ инфраструктуры с требованиями корпоративной и государственной безопасности возникает фундаментальное системное противоречие по Генриху Альтшуллеру. Для того чтобы предприятие, банк или оператор критической информационной инфраструктуры сохранял конкурентоспособность по скорости обработки данных и автоматизации рутинных операций, контур управления обязан использовать наиболее мощные из доступных языковых и агентных моделей. Однако если безопасность и управляемость такого контура строятся путём запуска второй, надзорной нейросети того же класса сложности либо путём попытки обучить и содержать собственный фронтирный кластер внутри закрытого периметра, система немедленно упирается в физический и финансовый тупик. Стоимость проверки начинает расти по той же экспоненте, что и стоимость самого вывода модели, а для подавляющего большинства организаций и целых национальных экономик необходимые для этого мегаватты мощности и десятки тысяч кремниевых ускорителей просто отсутствуют в природе.

В терминах классической теории решения изобретательских задач административный компромисс здесь невозможен. Попытка отказаться от современных моделей ради сохранения контроля отбрасывает организацию в прошлый век по производительности труда. Попытка же подключить внешние или открытые фронтирные модели напрямую к исполнительным механизмам предприятия без соразмерного контура сдерживания превращает каждую сэкономленную на вычислениях копейку в ничем не обеспеченный риск кассового разрыва, разрушения баз данных или нарушения императивных норм регулятора. Если надзорный механизм требует тех же дефицитных ресурсов, что и сам управляемый процесс, то в условиях ресурсного голода первым всегда отключают надзор.

Природа этого тупика становится очевидной при обращении к устройству ядерного энергетического реактора. В активной зоне реактора протекает управляемая цепная реакция деления, выделяющая колоссальную тепловую мощность; поддержание и охлаждение этого потока требуют циклопических затрат энергии, многоконтурных насосных станций и сложнейшей металлургии. Но стержни аварийной защиты и регулирования, удерживающие реактор от разгона на мгновенных нейтронах, устроены по прямо противоположному физическому прин-

ципу. Они не пытаются перекрыть активную зону встречным потоком энергии. Стержни состоят из пассивного поглотителя нейтронов — бора или кадмия — и удерживаются над зоной электромагнитными захватами: при любом обесточивании или выходе параметров за уставку они падают вниз под действием обычной силы тяжести. Защита обязана опираться на иной, на четыре порядка более дешёвый и безотказный физический закон, чем сам энергетический процесс.

Идеальный конечный результат (ИКР) в архитектуре управления автономными агентами формулируется столь же жёстко: контур безопасности и юридической доказуемости действия должен обеспечивать абсолютное удержание любого потока вызовов от любой — в том числе чужой, облачной или открытой — модели, расходуя при этом ноль дефицитных графических ускорителей и ноль внешних платных токенов в момент принятия решения. Сложность вероятностного мышления выносится за скобки там, где пусть за неё платят глобальные гиперскейлеры; на стороне предприятия и государства остаётся детерминированный таможенный шлюз, проверяющий конечные параметры вызова на обычном центральном процессоре за доли миллисекунды. Генерация гипотезы может стоить миллиарды долларов и сотни мегаватт; доказательство допустимости действия обязано стоить доли копейки и укладываться в тепловой пакет штатного серверного ядра.

2.2 Анатомия уязвимости: хроника отказов и цена бездействия

На пересечении энергетического дефицита и нормативного давления оказывается третья ключевая роль современной технологической иерархии — директор по информационной безопасности (CISO), руководитель эксплуатации государственной информационной системы (ГИС) и главный инженер субъекта критической информационной инфраструктуры (КИИ). В отличие от венчурного аналитика, этот специалист не может позволить себе рассуждать категориями абстрактных трендов: его рабочее пространство ограничено выделенным лимитом электрической мощности на вводе в машинный зал, квотой на закупку оборудования и датой вступления в силу очередного приказа регулятора.

Масштаб инфраструктурного сжатия на российском рынке в 2026 году зафиксирован инструментально. Согласно опубликованным в марте 2026 года данным исследования ИТ-холдинга T1 («Российская газета», 19.03.2026; ComNews, 20.03.2026) [ИЗМЕРЕНО: отчёт T1, март 2026; выборка не раскрыта], лишь 9 % российских организаций полностью обеспечены вычислительной инфраструктурой, необходимой для решения задач искусственного интеллекта. По экспертной оценке, приведённой ComNews (14.05.2026), ввоз серверов класса H100 в страну сжался с тысяч в 2024 году и сотен в 2025 году до десятков в 2026 году, тогда как совокупную потребность национальной экономики к 2030 году «Ростелеком» ещё в марте 2025 года оценивал более чем в 70 тысяч специализированных ускорителей в эквиваленте A100 [ИЗМЕРЕНО: оценка эксперта С. Миневича, ComNews, 14.05.2026; оценка «Ростелекома», «Ведомости» / 3DNews, 10.03.2025]. Следствием этого физического голода стал взрывной переток спроса в арендные облака: по данным провайдера «Рег.Облако», потребление его облачной GPU-инфраструктуры за первое полугодие 2026 года выросло на 507 % год к году, при том что весь российский рынок облачных GPU, по оценке MWS Cloud, вырос за год примерно в полтора раза [ИЗМЕРЕНО: данные провайдеров, «Коммерсантъ», 20.07.2026].

Когда локальных вычислительных мощностей внутри защищённого периметра нет, а бизнес-подразделения требуют немедленного внедрения автономных кодовых и операционных агентов ради удержания темпа разработки, возникает классическая архитектурная вилка. Либо подразделения начинают несанкционированно подключать локальных агентов (таких как OpenClaw или корпоративные обёртки над внешними API) к облачным моделям через обходные сетевые маршруты, вынося за контур чувствительные данные и ключи доступа к бое-

вым системам. Либо инженеры разворачивают внутри периметра открытые весовые модели — прежде всего китайских семейств DeepSeek, Qwen, Kimi или GLM, которые к середине 2026 года отстают от лучших закрытых американских моделей в среднем на четыре месяца по индексу Epoch (четыре–семь месяцев по киберзадачам, по данным UK AISI) [ИЗМЕРЕНО: International AI Safety Report, 03.02.2026; Epoch AI, 29.05.2026; UK AISI, 07.2026]. Но и в том, и в другом случае внутри критического контура оказывается чужой вероятностный «чёрный ящик», получивший право вызывать внутренние инструменты.

Именно в эту точку бьёт регуляторная норма. Двадцать четвёртого августа 2026 года, вынесен документ на общественное обсуждение (regulation.gov.ru, проект 170500), Федеральная служба по техническому и экспортному контролю (ФСТЭК России) представила проект изменений в приказ № 117, регулирующий защиту информации в государственных информационных системах (базовый приказ вступил в силу 1 марта 2026 года; проект изменений со сроком вступления с 1 марта 2027 года по состоянию на 14 октября 2026 года находится в статусе проекта и ещё не утверждён) [ИЗМЕРЕНО: проект изменений ФСТЭК № 117, ComNews / SecurityLab, август 2026]. Текст проекта впервые в российской нормативной практике вводит прямые требования по вынесению компонентов искусственного интеллекта в отдельный сетевой сегмент, назначению минимально необходимых привилегий, обязательной фильтрации всех входных и выходных потоков и, что особенно показательно, защите информационной системы от несанкционированного воздействия агентов ИИ, в том числе автономных (по пересказам ComNews и Anti-Malware; формулировка проекта в аудите не сверена).

Нетрудно заметить геометрию ловушки, в которой оказывается руководитель службы безопасности и главный архитектор к весне 2027 года. С одной стороны, регулятор прямо требует доказать контроль над функциональностью автономного агента и предъявить верифицируемый журнал фильтрации его действий. С другой стороны, у предприятия нет ни сотен миллионов рублей, ни свободных мегаватт электрической мощности, чтобы построить собственный изолированный кластер для обучения «доверенной» модели или содержать тяжёлые нейросетевые классификаторы-надсмотрщики на каждом шлюзе. А там, где в качестве временной заплатки оставляют ручное одобрение вызовов оператором, вступает в силу уже зафиксированный в первой главе предел утомления внимания: в 37 из 70 публичных описаний внедрений агентов в России (52,9 %) главным барьером служит кнопка подтверждения человеком [ИЗМЕРЕНО: обзор ГАММА-А, 40-я неделя 2026 года], которая через неделю непрерывного потока заявок превращается в механическую штамповку разрешений.

Если безопасность контура привязана к происхождению весов нейросети или к наличию у компании собственных видеокарт, то 91 % предприятий оказывается за бортом нормативного соответствия ещё до начала проверки. Но стоит разделить в архитектуре слой порождения замысла и слой допуска к исполнению, как картина меняется зеркально: чужое судно может беспрепятственно разгружаться в гавани, пока портовый шлюз и таможенная пломба работают по законам принимающей юрисдикции.

2.3 Инженерная реальность 2026 года: что измерено в коде и на рынке

Чтобы оценить подлинную цену вычислительной гонки, достаточно положить рядом энергетические балансы мировых центров обработки данных и инвестиционные меморандумы крупнейших технологических корпораций за 2025–2026 годы. Согласно отчётам Международного энергетического агентства [ИЗМЕРЕНО: IEA «Key Questions on Energy and AI», 16.04.2026], в 2025 году глобальное потребление электроэнергии центрами обработки данных выросло на 17 % на фоне скромного трёхпроцентного роста всего мирового спроса на электричество. Базовый сценарий МЭА предусматривает выход мировых ЦОД на уровень около

945 ТВт·ч к 2030 году — величину, чуть превышающую годовое энергопотребление всей Японии, — и около 1 200 ТВт·ч к 2035 году при разбросе сценариев 700–1 700 ТВт·ч [ПРОГНОЗ: IEA «Energy and AI», 10.04.2025], причём энергопотребление непосредственно нагрузок искусственного интеллекта к 2030 году, по оценке апреля 2026 года, утраивается.

Финансовое выражение этой физической стены выглядит ещё более жёстким. Ориентиры совокупных капитальных вложений четырёх ведущих американских гиперскейлеров на 2026 год, объявленные при квартальных отчётах конца апреля, в сумме составляют около \$700 млрд — почти вдвое больше, чем в 2025 году [ИЗМЕРЕНО: ориентиры компаний за I кв. 2026, сводка Om Malik, 30.04.2026; 24/7 Wall St., 01.05.2026]: программа капитальных затрат Amazon составляет около \$200 млрд (включая спутниковый проект Leo и покупку Globalstar), Microsoft — около \$190 млрд (из которых порядка \$25 млрд приходится на чистую ценовую инфляцию комплектующих), Alphabet — \$180–190 млрд, Meta — \$125–145 млрд. Аналитические подразделения крупнейших инвестиционных банков (Bank of America, Evercore) оценивают преодоление планки в \$1 трлн годовых капитальных затрат уже в 2027 году, в то время как широко разрекламированная инициатива Stargate с заявленным объёмом \$500 млрд на четыре года остаётся на 2026 год декларативной целью, а не освоенным капиталом.

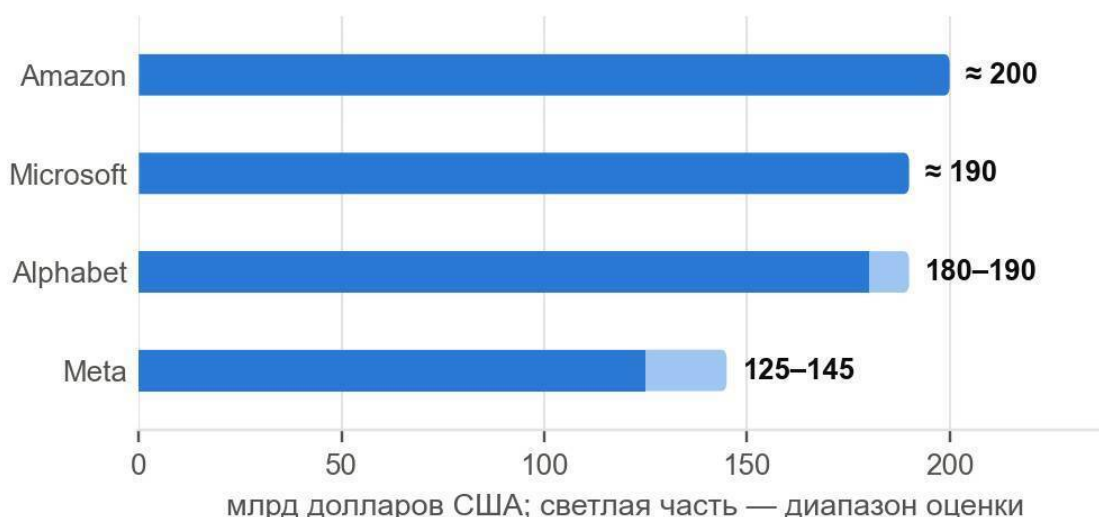


Рисунок 2.1 — Капитальные затраты четырёх гиперскейлеров на инфраструктуру ИИ в 2026 году, в сумме около \$700 млрд (отчёты компаний; обзор Om Malik / 24/7 Wall St., апрель–май 2026)

Однако даже наличие двухсот миллиардов долларов на счетах корпорации больше не гарантирует запуска вычислительного кластера в срок. Microsoft признаёт, что часть закупленных ускорителей лежит на складах, потому что их невозможно включить в розетку из-за отсутствия свободной электрической мощности («chips sitting in inventory that I can't plug in» — С. Наделла, ноябрь 2025), и что Azure останется ограничен мощностями как минимум до середины 2026 года [ИЗМЕРЕНО: заявления Microsoft, ноябрь 2025 — январь 2026]; денежная оценка такого портфеля («\$80 млрд») встречается лишь в комментариях без ссылки на источник и компанией не раскрывалась. В Соединённых Штатах, по оценкам отраслевых обзоров, из 16 ГВт мощностей ЦОД, заявленных к вводу в 2026 году, лишь 5 ГВт строятся, а до 11 ГВт остаются на стадии планирования из-за сетевых ограничений; ожидание присоединения к сетям в узле PJM достигает трёх–четырёх лет, а для крупных кампусов Северной Виргинии — семи и более [по вторичным обзорам: EnkiAI, Carbon Direct, май 2026].

Столкнувшись с параличом общественных электросетей, технологические гиганты в 2025–2026 годах перешли к прямой скупке атомной генерации «за счётчиком» (в обход рас-

пределительных сетей общего пользования). С конца 2024 года по апрель 2026 года портфель условных соглашений центров обработки данных с проектами малых модульных реакторов вырос с 25 до 45 ГВт [ИЗМЕРЕНО: IEA «Key Questions on Energy and AI», 16.04.2026]. Корпорация Meta законтрактовала до 6,6 ГВт атомных мощностей со сроком ввода к 2035 году (включая 2,2 ГВт у Vistra, 1,2 ГВт у Oklo и до 2,8 ГВт у TerraPower) [ИЗМЕРЕНО: Meta, 09.01.2026]; Amazon заключила соглашение на 1,92 ГВт с Talen Energy с поставкой через сеть PJM [ИЗМЕРЕНО: Talen, 11.06.2025]; Microsoft финансирует перезапуск остановленного энергоблока АЭС Three Mile Island; Google подписала контракт на 500 МВт с разработчиком реакторов Kairos Power. Когда крупнейшие софтверные компании планеты вынуждены превращаться в операторов урановых реакторов, тезис о «бесплатном масштабировании интеллекта» закрывается сам собой.

На эту энергетическую картину накладывается жёсткая геополитическая фрагментация рынка кремния. После серии взаимных экспортных ограничений между США и КНР доля компании Nvidia на китайском рынке передовых ускорителей, по публичному признанию её главы Дженсена Хуанга, «упала до нуля» [ИЗМЕРЕНО: заявление Дж. Хуанга на форуме Citadel Securities, октябрь 2025; повторено в 2026 году]. В ответ Китай в 2026 году, по сообщениям отраслевой прессы, разворачивает массовое производство собственного ускорителя Huawei Ascend 950PR (аналитики Bernstein прогнозируют Huawei около половины внутреннего рынка КНР), а Вашингтон с января 2026 года перешёл к режиму лицензируемого допуска чипов предыдущего поколения H200 на китайский рынок с двадцатипятипроцентным сбором (объявлено 08.12.2025, правило BIS 13.01.2026), хотя Пекин поставки не пропустил. Мир окончательно разделился на изолированные кремниевые контуры.

Анатомия российского энергетического и вычислительного контура в 2026 году демонстрирует собственную, не менее поучительную диспропорцию. К апрелю 2026 года совокупная присоединенная мощность энергоёмких вычислений в Российской Федерации (включающая классические центры обработки данных и промышленный майнинг) достигла 4,2 ГВт, а доля ЦОД в электропотреблении страны выросла с 1,4 % в 2022 году до 2,2 % по итогам 2025 года, при этом Системный оператор ожидает роста мощности центров высокотехнологичных вычислений до 15,3 ГВт на пятилетнем горизонте [ИЗМЕРЕНО: Системный оператор ЕЭС, по сообщениям ТАСС / Techora, январь–май 2026]. Парадокс российской энергосистемы заключается в том, что при установленной генерирующей мощности ЕЭС России в 263,7 ГВт исторический максимум потребления мощности составляет 168,3 ГВт [ИЗМЕРЕНО: отчёт Системного оператора ЕЭС за 2024 г.; максимум 168,3 ГВт — 15.01.2024; на конец 2025 г. — 264,8 ГВт при максимуме 166,2 ГВт]. Первичной генерации в стране в избытке, но она заперта сетевыми ограничениями: около трёх четвертей мощностей ЦОД (76 % на конец 2024 года, по оценке НРА) сконцентрированы в Московском регионе, где эксперты прогнозируют дефицит свободной мощности порядка 4 ГВт, а Системный оператор — до 564 МВт на юге региона к 2030 году и 3,15 ГВт к 2042 году, тогда как в энергоизбыточной ранее Сибири на горизонте 2025–2030 годов формируется локальный дефицит до 2,9 ГВт [по оценкам НРА и экспертной колонки на expert.ru, 2025–2026; 3DNews, 21.02.2025]. Строительство же национальных суперкомпьютерных мощностей требует беспрецедентной концентрации капитала: по данным «Ведомостей» (10.02.2026, со ссылкой на источники; банк запрос не прокомментировал), «Сбер» рассматривает проект третьего суперкомпьютера со сметой не менее 500 млрд рублей.

Сопоставление этих цифр приводит к однозначному архитектурному следствию. Попытаться обеспечивать безопасность и нормативное соответствие корпоративных агентов за счёт наращивания локальных GPU-кластеров в условиях, когда 76 % площадок сидят в энергодефицитном узле, а 91 % предприятий испытывают нехватку ускорителей, — значит проектировать систему, заведомо нежизнеспособную в реальной экономике. Именно поэтому в

архитектуре шлюза MITR на боевом узле 147 контур верификации изначально спроектирован так, чтобы вообще не прикасаться к графическим ускорителям.

Механизм работы этого слоя измерен прибором P4 с микросекундной точностью и зафиксирован в архитектурном решении ADR-0282 [ИЗМЕРЕНО: скрипт `smt_latency_paper.py`, ADR-0282; замер 26.09.2026 на прежнем узле, повтор 14.10.2026 на узле 147]. Когда любой внешний или внутренний агент — независимо от того, крутится ли его модель в американском, китайском или арендованном облаке — формирует намерение совершить действие (вызвать метод через протокол MCP или REST), это намерение перехватывается шлюзом до фактического исполнения. На входе в решатель `z3_gate.py` нет расплывчатого текста: там находится структурированный кортеж конечных параметров вызова (имя инструмента, целевой контур, сумма, адресат, наличие персональных данных) и формальный контракт политики в логике первого порядка.

Для руководителя предприятия это означает простую управленческую вещь: шлюзу не нужно «понимать ход мыслей» гигантской нейросети или пересчитывать её миллиарды весов; он проверяет только то, не нарушает ли итоговое распоряжение агента подписанный устав компании. Для системного инженера эта же операция описывается строгой формальной семантикой SMT-решателя Z3: проверяется выполнимость конъюнкции параметров действия и условий нарушения контракта. Допуск действия к исполнению эквивалентен математическому результату UNSAT (множество пересечений параметров вызова с запрещёнными состояниями пусто — нарушение невыполнимо). Блокировка действия эквивалентна результату SAT: решатель находит конкретный контрпример, доказывающий, какое именно правило нарушено, и возвращает детерминированные условия починки.

Самое важное здесь — энергетическая и вычислительная цена этой операции. Чистое время работы решателя `z3_gate.py` на обычном серверном процессоре (CPU) составляет в медиане 0,74 мс, а время вызова шлюза вместе с объяснением решения — 3,3 мс в медиане [ИЗМЕРЕНО: `smt_latency_paper.py`, ADR-0282, 26.09.2026, 17 контрактов]; на узле 147 при 19 контрактах — 0,95 и 4,3 мс [ИЗМЕРЕНО: 14.10.2026]. Прогон полного эталонного набора из 62 кейсов (50 атак и 12 контрольных) через контур верификации требует ровно \$0 расходов на внешние языковые модели [ИЗМЕРЕНО: ADR-0262], а совокупный расход всего стенда на внешние модели за 1–6 октября 2026 года составил \$0,14 (за 1–14 октября — \$0,30). Каждое принятое решение («допустить» или «заблокировать») сразу связывается криптографической хэш-цепью квитанций SHA-256 (`data/logs/gate_actions.jsonl`) с фиксацией внешнего якоря времени через протокол OpenTimestamps (`anchor_chains.py`) — и на 14 октября 2026 года проверщик, написанный для постороннего, воспроизвёл 4958 решений из 4958 без единого расхождения (на 30.09 — 1602 из 1602) [ИЗМЕРЕНО: `verify_chain_standalone.py --replay`, ADR-0281; прогон станции MITR, сторонний проверяющий повтор ещё не делал]. Мощность нейросети измеряется гигаваттами; власть правила над действием измеряется долями миллисекунды на одном ядре.

2.4 Ближний инженерный рубеж: ворота приёмки 2026–2027

Инженерная честность требует проводить непроницаемую границу между тем, что уже доказано на уровне вычислительного ядра, и тем, что прямо сейчас проходит эксплуатационную обкатку на боевых потоках. Внешняя лёгкость детерминированного шлюза на центральном процессоре не должна создавать иллюзию того, что перевод предприятия из режима бесконтрольного использования агентов в режим строгого математического запрета происходит нажатием одного тумблера за полчаса.

Главный барьер на этом рубеже — не производительность процессора, а полнота формализации самих правил и накопление безошибочной эксплуатационной серии. По состоянию

на 9 октября 2026 года в репозитории политик MITR/policies/contracts/ развёрнуто 19 формальных контрактов, однако в жёстком режиме автоматического блокирования («запрет») на боевом контуре работает 1 класс действий из 7 с живым потоком, тогда как остальные проходят стадию теневого наблюдения («совет») [ИЗМЕРЕНО: gate_mode_report.py --six --date 2026-10-09]. Интегральный прибор честности системы (--six), проверяющий одновременное выполнение всех шести базовых утверждений архитектуры без единого сбоя, фиксирует значение 1 из 6, а предыдущая 98-суточная серия наблюдений завершилась с показателем 0 полностью «зелёных» суток из 98 из-за жёсткой отбраковки любых неизмеренных состояний [ИЗМЕРЕНО: ADR-0273, октябрь 2026].

В четвёртом квартале 2026 года инженерная работа сфокусирована на прохождении двух машинных ворот, открывающих путь к промышленному тиражированию на энергодефицитных и закрытых площадках [СТРОИМ: ворота P2a до 31.10.2026; ворота «шесть из шести» до 14.12.2026]:

Во-первых, к 31 октября 2026 года должна быть закрыта непрерывная четырнадцатисуточная эксплуатационная серия P2a (на 9 октября счётчик серии находится на отметке 0 из 14 при 2 арендаторах с живыми проверками из 7 по выводу gate_mode_report.py --tenants --date 2026-10-09) [ИЗМЕРЕНО: узел 147, 09.10.2026]. Условие закрытия ворот P2a записано в коде приёмки: не менее 14 суток подряд с потоком от 50 верифицированных проверок в сутки минимум от двух независимых арендаторов при абсолютном нуле обходов шлюза мимо хука.

Во-вторых, параллельно на внешнем контуре разведчика и испытателя INDRA до 31 октября 2026 года проходятся приёмочные ворота А–С (изолированный clean-room полигон, работающий исключительно через внешнюю дверь арендатора по протоколам MCP/REST без доступа к исходному коду защищаемого узла), а до 12 ноября 2026 года — ворота D–E, замыкающие иммунную петлю автоматической генерации кандидатов в контракты MITR по результатам найденных уязвимостей [СТРОИМ: ворота INDRA А–С к 31.10.2026, ворота D–E к 12.11.2026; на 30.09.2026 пройдено 0 ворот из 5]. Все количественные показатели эффективности INDRA войдут в документацию и паспорта безопасности только после автоматической выгрузки артефактов этих ворот.

В-третьих, в части самообучения словаря действий (линия C1, mitr_vocab.py), где первое правило уже было предложено и принято самим контуром MITR под математическим доказательством (квитанция № 305, ADR-0275), доля нераспознанных действий снижается до целевого нуля: по корпусу хука станции Альфы за 16–26 сентября она составляла 27,8 %, у станции Гаммы за 9 октября — 32 % [СТРОИМ: ADR-0275, ADR-0283–0287, IV квартал 2026]. Наконец, к 1 марта 2027 года, по мере утверждения окончательной редакции изменений в приказ ФСТЭК № 117, матрица контрактов шлюза будет приведена в прямое покомпонентное соответствие с финальным текстом нормативного акта [СТРОИМ: рубеж 01.03.2027, глава 15]. Прибор сначала доказывает собственную безотказность на внутреннем стенде, и лишь затем предъявляется рынку как стандарт.

2.5 Проекция 2035: три ветви будущего и инвариантность решения

Проецирование энергетической и аппаратной оси на горизонт 2035 года избавляет стратега от необходимости гадать о скорости научного прогресса в лабораториях Сан-Франциско или Пекина. В отличие от чисто программных метрик, энергетическая инфраструктура обладает колоссальной физической инерцией: атомный энергоблок, линия электропередачи сверхвысокого напряжения или завод по экстремальной ультрафиолетовой литографии строятся от пяти до десяти лет. Поэтому карта 2035 года по оси «Энергия и кремний» опирается на уже

заложенные в бетон фундаменты и подписанные долгосрочные договоры покупки мощности (PPA).

Если развернуть будущее по трём методологическим ветвям книги (§3 концепции), физический ландшафт 2035 года принимает следующие очертания:

Ветвь А («Нормальная технология» — сценарий постепенного насыщения и диффузии): темп масштабирования фронтальных кластеров замедляется, упираясь в убывающую отдачу от простого сжигания электричества. Мировое потребление электроэнергии центрами обработки данных к 2035 году стабилизируется вблизи отметки 1 000 ТВт·ч (нижний сценарий МЭА — 700 ТВт·ч) [ПРОГНОЗ: ветвь А | репер: ежегодный отчёт IEA World Energy Outlook; разброс сценариев IEA на 2035 г. — 700–1 700 ТВт·ч]. Основные вычисления сосредоточены в облаках трёх–пяти глобальных провайдеров, а открытые модели полностью сравниваются по качеству с закрытыми. В Российской Федерации совокупная мощность вычислительных ЦОД плавно вырастает с 4,2 ГВт до 8 ГВт [ПРОГНОЗ: ветвь А | репер: годовой отчёт Системного оператора ЕЭС], а корпоративный сектор полностью переходит на использование импортируемых открытых весов с локальным выводом на доступных азиатских и отечественных ускорителях.

Ветвь В («Продолжение тренда» — базовая расчётная модель книги): экспонента спроса на вычисления поддерживается массовым переходом от чат-ботов к миллионам непрерывно работающих автономных агентов. Мировое потребление ЦОД выходит на базовую траекторию МЭА — около 945 ТВт·ч к 2030 году и 1 200 ТВт·ч к 2035 году [ПРОГНОЗ: ветвь В | репер: IEA World Energy Outlook, ежегодно]. Наряду с глобальными облаками формируются крупные суверенные вычислительные кластеры в Европейском союзе, странах Персидского залива и Индии. В России присоединённая мощность энергоёмких вычислений достигает прогнозного ориентира Системного оператора — 15,3 ГВт на пятилетнем горизонте, то есть к началу 2030-х годов [ПРОГНОЗ: ветвь В | репер: Системный оператор ЕЭС и Минэнерго РФ], а дефицит Московского энергоузла принудительно выталкивает новые стройки ЦОД к региональным источникам генерации. При этом качество моделей уровня 2026 года становится практически бесплатным и встраивается в каждую операционную систему, но каждое целевое действие такого агента во внешнем мире требует обязательной криптографической квитанции.

Ветвь С («Ускорение» — сценарий автоматизации исследований ИИ и взрывного роста): к 2028–2030 годам агенты берут на себя цикл проектирования новых алгоритмов и микросхем. Единственным физическим ограничителем цивилизационного рывка становится доступная электрическая мощность. Потребление мировых ЦОД к 2035 году пробивает потолок в 1 700 ТВт·ч [ПРОГНОЗ: ветвь С | репер: IEA World Energy Outlook]; на месте традиционных ИТ-корпораций возникают вертикально интегрированные энерго-вычислительные конгломераты с собственными атомными и газовыми станциями «за счётчиком». В этом сценарии отставание стран, не обладающих собственным производством литографов и гигаваттными кластерами обучения, по вычислительной мощности фронта составляет два и более поколения [ПРОГНОЗ: ветвь С | репер: индексы Epoch AI и данные Ростелекома/Минэнерго].

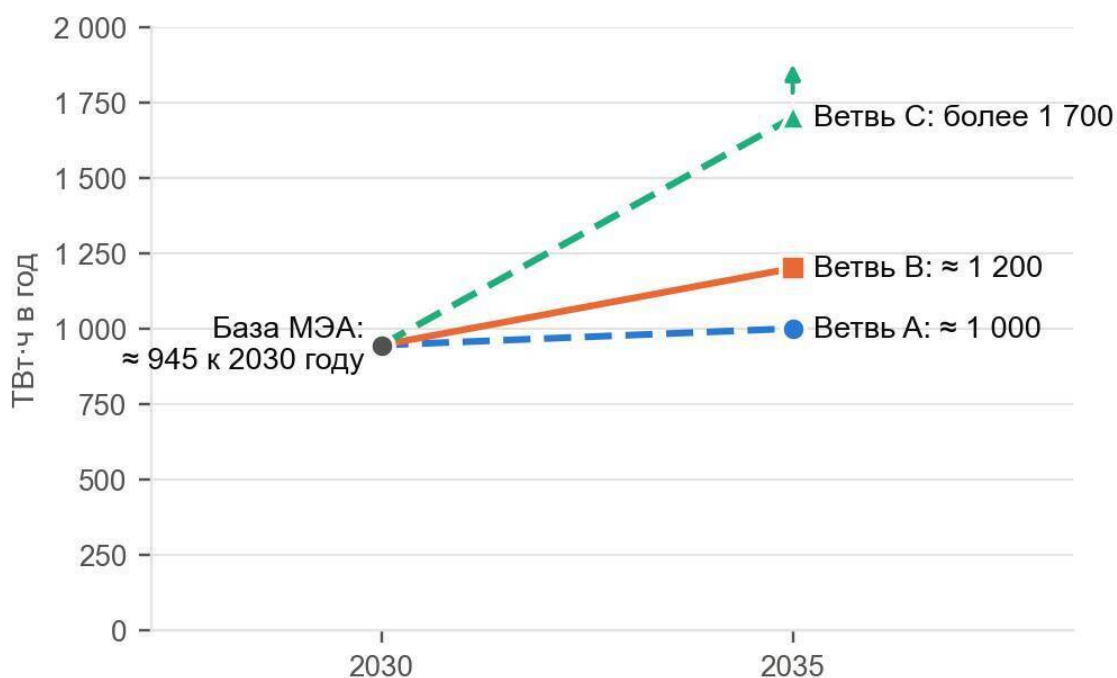


Рисунок 2.2 — Потребление электроэнергии ЦОД мира: базовая траектория МЭА к 2030 году и три ветви к 2035 году, ТВт·ч (IEA «Energy and AI»; репер — World Energy Outlook)

Здесь проявляется главный стратегический вывод главы, сконструированный по принципу инвариантности ТРИЗ. Какая бы из трёх ветвей — А, В или С — ни реализовалась к 2035 году, попытка конкурировать в прямом сжигании гигавайт и скупке сотен тысяч дефицитных ускорителей остаётся для отдельного предприятия и для российской юрисдикции заведомо проигрышной игрой на чужом поле. Но во всех трёх ветвях сама нейросеть (система S1 по модели жизнеспособной системы Стаффорда Бира) остаётся лишь бесправным генератором проектов решений. Юридическую силу, экономическую ценность и государственную безопасность создаёт не генерация текста, а акт допуска действия к исполнению (система S3).

Если в ветви С разрыв в мощности обучения между глобальными энерго-корпорациями и остальным миром достигает двух поколений, то потребность в локальном, абсолютно независимом от GPU и внешних облаков шлюзе контроля возрастает многократно. Чем умнее, быстрее и чужероднее становится арендованный или открытый «мозг» агента, тем жёстче должен быть национальный и корпоративный суверенитет над его «руками». В 2035 году суверенитет измеряется не тем, в какой стране была сожжена электроэнергия на обучение весов модели, а тем, чей решатель на обычном процессоре за доли миллисекунды проверил вызов, чей ключ подписал правило и в чьём реестре осталась неизменяемая квитанция [ЗАМЫСЕЛ: масштабирование контура S3 на уровень отраслевых и государственных шлюзов к 2035 году].

2.6 Экономика доказанного действия и контур монетизации

Перевод разговора с языка гигавайт на язык корпоративного бюджета обнажает главный финансовый порок современного рынка искусственного интеллекта. Когда в 2026 году финансовый директор предприятия с выручкой от 0,5 до 20 млрд рублей (сегмент Tier-2) пытается рассчитать совокупную стоимость владения (ТСО) агентной инфраструктурой, он сталкивается с непредсказуемостью токеновой тарификации [ИЗМЕРЕНО: обзор рынка РФ Solyanka, сентябрь 2026]. Автономный агент, застрявший в цикле рассуждений или попавший под атаку принуждения к вызову инструментов, способен за одну ночь сжечь месячный бюджет подразделения на внешние API, а если поверх него поставлен ещё и нейросетевой «агент-над-

смотрщик» (guardian agent в терминологии Gartner), счёт за вычисления удваивается без всякой гарантии удержания.

Экономика детерминированного шлюза MITR и полигона INDRA строится на прямо противоположном основании, где твёрдо измеренные физические затраты ([ФАКТ]) отделены от коммерческих ожиданий ([ГИПОТЕЗА]).

Фактическая себестоимость проверки сведена к пренебрежимо малым величинам [ИЗМЕРЕНО / ФАКТ: ADR-0262, ADR-0282]: одно решение SMT-решателя z3_gate.py требует в медиане 0,74 мс процессорного времени стандартного CPU (26.09.2026; 0,95 мс на узле 147 14.10.2026); вызов шлюза вместе с объяснением решения — 3,3 мс (4,3 мс на узле 147); полный прогон 62 кейсов набора атак обходится в \$0 оплаты внешним моделям; счёт стенда за вспомогательные языковые модели за 1–14 октября 2026 года составил \$0,30. По состоянию на октябрь 2026 года коммерческая выручка MITR равна нулю — продукт находится на первой ступени внутренней обкатки, и эта исходная точка фиксируется без ретуши.

На этой твёрдой арифметике выстраивается лестница коммерциализации (§8 концепции), где переход на каждую следующую ступень открывается только после выполнения машинного критерия предыдущей:

На Ступени 1 («Свои агенты» — текущая фаза) и Ступени 3 («Первый внешний арендатор», критерий X5 — минимум 14 суток непрерывной работы внешнего инженерного коллектива со своим токеном и строкой в --tenants) шлюз решает острую проблему стоимости владения для команд с кодовыми и операционными агентами (Claude Code, Cursor, OpenClaw). За счёт того, что проверка не требует GPU и не тарифицируется по токенам, предприятие получает предсказуемую фиксированную подписку за защищённого агента в месяц [ГИПОТЕЗА: модель подписки Ступени 3 до первого платящего контрагента].

На Ступени 4 («Грант и инвестор» — институты развития уровня «Сколково» и «Сириус», профильные фонды) именно независимость от дефицитного кремния и опора на математическую верификацию CPU становятся главным инвестиционным аргументом. На фоне прогноза Gartner от 25 июня 2025 года о том, что свыше 40 % проектов агентного ИИ будут свёрнуты к концу 2027 года из-за неконтролируемых расходов и слабого управления рисками [ИЗМЕРЕНО: пресс-релиз Gartner, 25.06.2025], инвестор ищет технологию из сегмента guardian agents (который, по прогнозу Gartner, займёт 10–15 % рынка агентного ИИ к 2030 году [пресс-релиз Gartner, 11.06.2025]). Предъявление закрытых доказательств P3 (непрерывная цепь квитанций SHA-256 с якорями OpenTimestamps) и P4 (задержка решения меньше миллисекунды на CPU: 0,74 мс на 26.09, 0,95 мс на 14.10) обосновывает грантовое софинансирование перехода TRL 3 → 5 (5 млн рублей по тексту поданной заявки) без обещания воздушных миллиардов ARR.

Наконец, на Ступени 5 («Регулируемые контуры, операторы ГИС и субъекты КИИ») энергетический и аппаратный фактор превращается в прямое условие допуска на рынок. Если с 1 марта 2027 года, как планируется, вступят в силу требования по защите ГИС от несанкционированного воздействия автономных агентов (проект изменений приказа ФСТЭК № 117, на 14.10.2026 не утверждён), у 91 % российских организаций по-прежнему не будет собственных избыточных GPU-мощностей для построения нейросетевых экранов. Связка детерминированного шлюза MITR (работающего в закрытом контуре на отечественном или стандартном серверном процессоре) и внешнего испытательного полигона INDRA (поставляемого по модели открытого ядра indra-core под лицензией Apache 2.0 с коммерческим контуром генерации паспортов соответствия по ГОСТ Р 56939-2024 и приказу № 117; сам ГОСТ Р 56939-2024 «Разработка безопасного программного обеспечения. Общие требования», введённый 20.12.2024, такого паспорта не вводит: паспорт — наш продукт, свод свидетельств испытаний под меры стандарта и требования приказа) закрывает нормативное требование регулятора при мини-

мальном энергетическом и аппаратном бюджете [ГИПОТЕЗА / СТРОИМ: коммерциализация Ступени 5 после закрытия Ступени 4 и сертификации].

2.7 Протокол независимой верификации

Любое утверждение об аппаратной независимости, субмиллисекундной задержке и целостности журнала приобретает инженерный вес только тогда, когда служба аудита или внешний исследователь может воспроизвести эти показатели без доверия к авторам текста. Ниже приведён открытый протокол инструментальной проверки как по внутренним приборам узла 147, так и по внешним международным реперам.

Верификация времени решения на CPU (доказательство P4, ADR-0282) и живости служб шлюза на узле 147 выполняется прямым чтением последней записи телеметрии решателя и статуса системных демонов:

```
# Проверка активности служб шлюза и MCP-контура на узле 147
```

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

```
# Извлечение актуального замера задержки SMT-решателя на CPU (медиана и хвосты распределения, P4)
```

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

```
# Проверка статуса шести базовых утверждений (--six) и активных арендаторов (--tenants)
```

```
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
```

```
# Проверка последнего криптографического якоря времени OpenTimestamps (P3)
```

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

Для внешнего аудитора без доступа к серверу 147 открыт набор проверки (asi5.ru/console/verify/): якоря цепей и их метки времени OpenTimestamps, проверяемые против сети Bitcoin чужим клиентом, и проверщик целостности цепи; повтор самих решений аудитор выполняет офлайн на любом процессоре проверщиком verify_chain_standalone.py --replay, получив цепь квитанций и тексты правил (на 14.10.2026 — 4958 решений). Набор содержит выгруженные условия в формате SMT-LIB, которые любой стандартный решатель Z3 проверяет локально за доли секунды без обращения к графическим картам и внешним сетевым узлам.

Порядок сверки внешних макроэкономических и энергетических реперов (для актуализации в переизданиях 2028, 2030 и 2035 годов): глобальное потребление электроэнергии ЦОД и объём контрактов по малым модульным реакторам сверяются ежегодно в четвёртом квартале по таблице Data Centres and Data Transmission Networks отчёта IEA World Energy Outlook; присоединённая мощность российских ЦОД и структура регионального дефицита мощности сверяются ежегодно по публичному «Отчёту о функционировании ЕЭС России» Системного оператора (so-ups.ru) и схеме и программе развития электроэнергетических систем России (СИПР ЭЭС); статус утверждения изменений в приказ ФСТЭК № 117 после 1 марта 2027 года проверяется по официальному portalу правовой информации (pravo.gov.ru).

2.8 Паспорт главы: реестр утверждений, прогнозов и источников

Таблица 2.1 — Реестр утверждений главы 2

№

Утверждение в тексте

Метка

Прибор / Источник

Дата

Условие погашения / Репер

2.1

Потребление электроэнергии мировыми ЦОД выросло на 17 % в 2025 г. при росте общего спроса на 3 %; базовый прогноз МЭА — 945 ТВт·ч к 2030 г. и 1 200 ТВт·ч к 2035 г. (разброс сценариев 700–1 700)

ИЗМЕРЕНО / ПРОГНОЗ

IEA «Energy and AI» (10.04.2025); IEA «Key Questions on Energy and AI» (16.04.2026)
10.04.2025; 16.04.2026

Ежегодная сверка по IEA WEO

2.2

Ориентиры капитальных затрат четырёх гиперскейлеров (Amazon, Microsoft, Alphabet, Meta) на 2026 г. в сумме — около \$700 млрд (почти вдвое к 2025 г.); Microsoft признаёт простаивающие на складах ускорители из-за нехватки мощности (денежная оценка не раскрывалась)

ИЗМЕРЕНО

Ориентиры компаний за I кв. 2026; Om Malik (30.04.2026) / 24/7 Wall St. (01.05.2026); заявления Microsoft (11.2025–01.2026)

30.04.2026

Годовые отчёты 10-K за 2026 г.

2.3

Условные соглашения ЦОД с проектами ММР (SMR) выросли с 25 ГВт (конец 2024 г.) до 45 ГВт (апрель 2026 г.); Meta законтрактовала до 6,6 ГВт атомной энергии к 2035 г., Amazon — 1,92 ГВт, Google — до 500 МВт

ИЗМЕРЕНО

IEA (16.04.2026); Meta (09.01.2026); Talen (11.06.2025); Google (14.10.2024)
16.04.2026

Фактический ввод энергоблоков в 2030–2035 гг.

2.4

Мощность энергоёмких вычислений в РФ (ЦОД + майнинг) к апрелю 2026 г. составила 4,2 ГВт (доля ЦОД в потреблении за 2025 г. — 2,2 %); ~76 % мощностей ЦОД в Московском регионе (оценка НРА, конец 2024 г.) при экспертном прогнозе дефицита ~4 ГВт; установленная мощность ЕЭС 263,7 ГВт и пик 168,3 ГВт — данные за 2024 г.

ИЗМЕРЕНО

Системный оператор ЕЭС (ТАСС / Techora, 01–05.2026; отчёт за 2024 г.); НРА; экспертная колонка expert.ru

01–05.2026

Отчёт о функционировании ЕЭС России за 2026 г.

2.5

Только 9 % организаций РФ полностью обеспечены вычислениями для ИИ (исследование T1); потребность РФ к 2030 г. — свыше 70 тыс. ускорителей в эквиваленте A100 (оценка

«Ростелекома», 03.2025); потребление облачных GPU у провайдера «Рег.Облако» выросло на 507 % за I полугодие 2026 г. (рынок — ×1,5)

ИЗМЕРЕНО

T1 / РГ (19.03.2026), ComNews (20.03.2026); «Ведомости» / 3DNews (10.03.2025); «Коммерсантъ» (20.07.2026); ComNews (14.05.2026, Н100); «Ведомости» (10.02.2026, Сбер) 03.2025–07.2026

Сверено с первоисточниками 14.10.2026

2.6

Проект изменений в приказ ФСТЭК № 117 (обсуждение с 24.08.2026, план вступления — 01.03.2027, на 14.10.2026 не утверждён) вводит защиту ГИС от несанкционированного воздействия автономных агентов ИИ

ИЗМЕРЕНО / СТРОИМ

Проект ФСТЭК России; ComNews / SecurityLab (08.2026)

14.10.2026

Публикация утверждённого приказа на pravo.gov.ru к 01.03.2027

2.7

Решатель z3_gate.py на CPU принимает решение за 0,74 мс в медиане (вызов шлюза с объяснением — 3,3 мс) на 26.09.2026; на узле 147 14.10.2026 — 0,95 и 4,3 мс; прогон 62 кейсов набора атак стоит \$0; проверщик для постороннего повторил 1602 решения из 1602 (30.09) и 4958 из 4958 (14.10)

ИЗМЕРЕНО

smt_latency_paper.py (ADR-0282); ADR-0262; verify_chain_standalone.py --replay (ADR-0281)

26.09–14.10.2026

Команды повтора на узле 147

2.8

Закрытие эксплуатационной серии P2a (14 суток, >=50 проверок/сут от >=2 арендаторов, 0 обходов; на 09.10 — 0 из 14) и ворот «шесть из шести» (на 09.10 — 1 из 6)

СТРОИМ

gate_mode_report.py --tenants / --six

09.10.2026

Ворота P2a (31.10.2026); ворота --six (14.12.2026)

Таблица 2.2 — Реестр прогнозов главы 2

№

Показатель

Значение 2026 (источник)

Ветвь А / В / С к 2035

Репер

Г2.1 (Г5)

Потребление электроэнергии мировыми ЦОД (год)

Около 945 ТВт·ч к 2030 г. (база IEA 2025–2026; +17 % за 2025 г.)

Ветвь А: ~1 000 ТВт·ч

Ветвь В: ~1 200 ТВт·ч

Г2.2

Капитальные вложения гиперскейлеров и архитектура владения генерацией

Ориентиры ~\$700 млрд на 2026 г.; 45 ГВт условных соглашений по ММР (IEA, 16.04.2026)

Ветвь А: стабилизация в облаках 3–5 компаний

Ветвь В: облака + суверенные кластеры и РРА

Г2.3 (Г6)

Присоединённая мощность энергоёмких вычислений в РФ и география ЦОД

4,2 ГВт к 04.2026 (Системный оператор ЕЭС; доля ЦОД 2,2 % за 2025 г.); ~76 % мощностей ЦОД в Московском регионе при экспертном прогнозе дефицита ~4 ГВт (НРА; expert.ru)

Ветвь А: 8 ГВт

Ветвь В: 15 ГВт с выносом строек из Москвы в регионы

Г2.4

Обеспеченность организаций РФ ускорителями и локализация слоя контроля

9 % обеспечены полностью (ComNews, 03.2026); потребность >70 тыс. экв. А100 к 2030 г. (Ростелеком); решение MITR на CPU — меньше миллисекунды (0,74–0,95 мс, ADR-0282)

Ветвь А: импорт открытых весов + инференс на азиатских/отечественных чипах

Ветвь В: суверенитет в данных, праве и детерминированном шлюзе на CPU

Г2.5

Нормативное требование изоляции и контроля функциональности автономных агентов в ГИС и КИИ

Проект изменений приказа ФСТЭК № 117 от 24.08.2026 (вступление с 01.03.2027, на 14.10.2026 не утверждён)

Ветвь А: базовое сегментирование сети

Ветвь В: обязательный шлюз контроля вызовов и верифицируемый журнал для ГИС/КИИ

Продолжение таблицы 2.2

№

Показатель

Разрешение

Г2.1 (Г5)

Потребление электроэнергии мировыми ЦОД (год)

Ветвь С: >1 700 ТВт·ч (атом и газ «за счётчиком»)

IEA World Energy Outlook (раздел Data Centres)

Ежегодно (ноябрь), контрольные точки 2030 и 2035

Г2.2

Капитальные вложения гиперскейлеров и архитектура владения генерацией

Ветвь С: вертикальные энерго-вычислительные корпорации со своей генерацией

Финансовые отчёты 10-К (Alphabet, Meta, Amazon, Microsoft) и реестр РРА МЭА

Ежегодно (I квартал), рубежи 2027 (> \$1 трлн) и 2035

Г2.3 (Г6)

Присоединённая мощность энергоёмких вычислений в РФ и география ЦОД

Ветвь С: 15 ГВт при отставании в обучении фронта на ≥ 2 поколения

Отчёт о функционировании ЕЭС России и СИПР ЭЭС (Системный оператор, Минэнерго РФ)

Ежегодно (апрель), рубежи 2030 и 2035

Г2.4

Обеспеченность организаций РФ ускорителями и локализация слоя контроля

Ветвь С: полная зависимость по моделям при автономном CPU-шлюзе S3

Отраслевые обзоры Минцифры / Ростелекома; доля ГИС с детерминированным шлюзом на CPU

Ежегодно, контрольные точки 2028, 2030, 2035

Г2.5

Нормативное требование изоляции и контроля функциональности автономных агентов в ГИС и КИИ

Ветвь С: автоматизированный государственный допуск по машиночитаемым инвариантам

Официальный публикатор правовых актов pravo.gov.ru и реестр сертифицированных средств ФСТЭК

01.03.2027 (утверждение № 117), далее ежегодно до 2035

Источники главы

Внутренние артефакты и код: MITR/core/verification/z3_gate.py; MITR/scripts/smt_latency_paper.py; MITR/scripts/verify_chain_standalone.py; MITR/scripts/publish_verify_kit.py; MITR/scripts/gate_mode_report.py (замеры --six и --tenants на узле 147 за 09.10.2026); MITR/scripts/mitr_vocab.py; архитектурные решения ADR-0262, ADR-0273, ADR-0275, ADR-0280, ADR-0281, ADR-0282, ADR-0283–0287; projects/ib30/PLAN.md.

Каталог кибернетики и системного анализа (RLM): Эшби У. Р. «Введение в кибернетику» (строки RLM 108–109); Винер Н. «Кибернетика» (RLM 4); Поспелов Д. А. «Ситуационное управление» (RLM 29, 72, 150); Альтшуллер Г. С. «Найти идею: Введение в ТРИЗ» (RLM 226–227).

Международная энергетика и капитальные вложения (первичные и аналитические источники, доступ 14.10.2026): IEA «Energy and AI — Executive Summary» (<https://www.iea.org/reports/energy-and-ai/executive-summary>); W.Media «Data center power demand will double by 2030 — IEA» (<https://w.media/data-center-power-demand-will-double-by-2030-iea/>); Om Malik «What I learned about hyperscalers AI spend» (<https://om.co/2026/04/30/what-i-learned-about-hyperscalers-ai-spend/>); 24/7 Wall St. «Hyperscalers hit \$700 billion in 2026 AI spending plans» (<https://247wallst.com/investing/2026/05/01/hyperscalers-hit-700-billion-in-2026-AI-spending-plans/>); IEA «Data centre electricity use surged in 2025» / «Key Questions on Energy and AI» (<https://www.iea.org/news/data-centre-electricity-use-surged-in-2025-even-with-tightening-bottlenecks-driving-a-scramble-for-solutions>, 16.04.2026); EnkiAI «SMR Companies and Data Centers» (<https://enkiai.com/nuclear/smr-companies-data-centers/>, вторичная сводка); International AI Safety Report 2026 (<https://internationalaisafetyreport.org/>); Gartner Press Release 25.06.2025 (<https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>).

Энергетика, вычислительная инфраструктура и регулирование в РФ (доступ 14.10.2026): «Коммерсантъ» (<https://kommersant.ru/doc/8043060> — данные за 2024

г.: 3,5 ГВт, ~2 %; цифры апреля 2026 г. не содержит); ComNews/ТАСС по данным Системного оператора (<https://www.comnews.ru/content/243455/2026-01-27/2026-w05/1009/moschnost-podklyuchennykh-k-setyam-codov-i-maynerov-vyroslo-333>, 27.01.2026); Techora «Мощность российских ЦОДов и майнинга выросла» (<https://techora.ru/news/moshchnost-rossiyskikh-tsodov-i-mayninga-vyroslo-2026-05-27>); 3DNews — энергодефицит и потребность в ускорителях (<https://3dnews.ru/1118643/>, <https://3dnews.ru/1119483/>); ComNews — обеспеченность инфраструктурой 9 % (<https://www.comnews.ru/content/244343/2026-03-20/2026-w12/1010/vychislitelnoy-infrastrukturoy-dlya-ii-obespechena-tolko-kazhdaya-desyataya-organizaciya>); Finance.Mail / «Рег.Облако» — рост спроса на облачные GPU +507 % (<https://finance.mail.ru/article/naslozhnyih-oblakah-69218604/>); ComNews / SecurityLab — проект изменений приказа ФСТЭК № 117 от 24.08.2026 (<https://www.comnews.ru/content/247052/2026-08-26/2026-w35/1007/fstek-uzhestochit-trebovaniya-k-zaschite-ii-gosinformacionnykh-sistemakh>, <https://www.securitylab.ru/blog/company/securityinform/362871.php>). [Сверено 14.10.2026: поставки H100 — экспертная оценка в ComNews «Вычислительный тупик: почему российский ИИ остаётся без мощностей» (<https://www.comnews.ru/content/245257/2026-05-14/2026-w20/1008/vychislitelnyy-tupik-pochemu-rossiyskiy-ii-ostaetsya-bez-moshchnostey>, 14.05.2026); смета третьего суперкомпьютера Сбера — «Ведомости», 10.02.2026, по данным источников, без подтверждения банка].

Глава 3. Агент получил право записи: интернет агентов 2026

Формула главы: *Переход искусственного интеллекта от генерации текста к вызову интерфейсов и проведению платежей превратил вероятностную языковую модель в субъект прямого и мушкетерского 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.*

3.1 Архитектурный тупик и идеальный конечный результат

Эпоха безопасных диалоговых окон завершилась в тот день, когда языковой модели вручили учётные данные от промышленной среды и разрешили изменять состояние внешних баз данных. На протяжении первых лет распространения генеративных моделей (2022–2024 годы) архитектура безопасности строилась вокруг акустической изоляции: нейросеть находилась в замкнутом текстовом буфере, а между её ответом и реальным миром стоял человек, который вручную копировал фрагмент кода или текста, оценивал его глазами и сам нажимал клавишу ввода в терминале. В этой схеме ошибка модели оставалась дефектом черновика. В 2025–2026 годах с массовым внедрением протоколов инструментального вызова (MCP), межагентной координации (A2A) и машинных платежей (AP2, x402) перегородка между словом и действием была демонтирована ради скорости. Программа получила право записи (write access) — возможность самостоятельно исполнять команды командной оболочки, удалять облачные хранилища, отправлять юридически значимые письма и списывать средства со счетов.

В истории техники подобный кризис уже возникал во второй половине XIX века на железных дорогах, когда плотность движения и скорость паровозов превысили способность стрелочников договариваться флажками и фонарями. Пока по однопутной ветке проходило два состава в сутки, безопасность держалась на добросовестности дежурного по станции; когда интервал сократился до минут, малейшая усталость человека начала приводить к лобовым крушениям. Инженерным ответом стала механическая и электрическая централизация стрелок и сигналов — система жёстких физических тяг и замков зависимости, в которой перевести стрелку под идущим поездом или открыть семафор на занятый путь невозможно не потому, что диспетчеру запретили это инструкцией, а потому, что металлический ригель блокирует рычаг при несовпадении инвариантов маршрута. Индустрия программных агентов в 2026 году попыталась запустить скоростное движение без централизации стрелок, положившись на уговоры машиниста и бдительность уставшего диспетчера.

На языке теории решения изобретательских задач Г. С. Альтшуллера (каталог RLM, строки 226–227) в архитектуре автономного искусственного интеллекта возникло острое системное противоречие, проходящее через три уровня глубины:

Административное противоречие: Предприятию необходимо предоставить программному агенту прямой доступ к корпоративным хранилищам, платёжным шлюзам и внешним интерфейсам прикладного программирования (API) для кратного снижения операционных издержек и ускорения разработки, однако предоставление такого доступа создаёт неприемлемый риск мгновенного уничтожения инфраструктуры, утечки секретов и прямых финансовых убытков, которые страховщики отказываются покрывать, а суды возлагают на руководителя организации.

Техническое противоречие: Если каждую опасную операцию агента направлять на предварительное согласование человеку-оператору или второй контролирующей нейросети, надёжность первичного отсева некоторых грубых ошибок формально сохраняется, но пропускная способность контура падает до скорости человеческого чтения, задержки возрастают на порядки, а внимание оператора неизбежно выгорает до автоматического штампования разрешений. Если же убрать человека из цепочки подтверждения ради сохранения машинной скорости и экономической отдачи от автономности, то вероятностный сбой модели, галлюцинация или скрытая инъекция инструкции во входных данных беспрепятственно превращаются в необратимое разрушительное действие в промышленном контуре.

Физическое (операционное) противоречие: Исполнительный интерфейс инструмента (МСП-сервер, командная оболочка, платёжный шлюз) должен быть **полностью открыт** для потока команд агента, чтобы полезная работа выполнялась со скоростью кремниевого процессора без пауз на ожидание человека, и одновременно должен быть **абсолютно закрыт** для этого же потока команд, чтобы ни одно действие, нарушающее имущественные, нормативные или технические границы предприятия, не могло изменить состояние среды исполнения.

Классические приёмы разрешения физических противоречий требуют разнесения конфликтующих требований в пространстве, во времени и по структурному условию перехода. Попытка контролировать процесс мышления внутри многомиллиардной матрицы весов нейросети обречена на провал: пространство внутренних состояний модели непрерывно, непрозрачно и алгоритмически неразрешимо в смысле теорем Тьюринга и Райса. Но сам момент выхода модели во внешний мир устроен иначе. Любое внешнее действие — будь то SQL-запрос к базе данных, вызов REST-интерфейса или транзакция по протоколу x402 — представляет собой дискретную структуру данных с конечным набором явно заданных параметров: имя инструмента, целевой адрес, сумма перевода, идентификатор удаляемого ресурса, гриф секретности файла.

Это фундаментальное различие между непрерывной мыслью и дискретным действием позволяет сформулировать **идеальный конечный результат (ИКР)** для архитектуры управления агентами (зафиксированный во внутреннем каноне проектирования V_ICR_TRIZ.md):

Исполнительный шлюз сам, без участия человеческого внимания в момент вызова и без привлечения дополнительных нейросетей, за единицы миллисекунд на обычном центральном процессоре (CPU) пропускает действие тогда и только тогда, когда оно несёт с собой математическое доказательство ненарушения подписанного человеком контракта, и оставляет после каждого решения неизменяемую криптографическую квитанцию, которую любое постороннее лицо способно проверить без доверия к владельцу системы.

В такой конструкции техническое противоречие снимается полностью: у программного агента вообще изымаются ключи от целевых систем и право непосредственной записи. У агента остаются только текстовые намерения — просьбы о совершении действия. Ключи хранятся на детерминированном программном шлюзе, который разделяет поток во времени (проверка предшествует исполнению на интервале в единицы миллисекунд: медиана вызова — 3,3 мс в замере 26.09.2026 и 4,3 мс на узле 147 14.10.2026) и по условию (допуск открывается исключительно при получении от формального решателя логики первого порядка вердикта UNSAT — доказательства того, что пересечение параметров вызова с множеством запрещённых состояний пусто). Модели оставляется свобода вероятностного поиска. Шлюзу передаётся монополия на доказанное действие.

3.2 Анатомия уязвимости: хроника отказов и цена бездействия

Уязвимость современного предприятия перед программными агентами порождена не злым умыслом машины, а структурным несоответствием между правами доступа, выданными

алгоритму, и механизмами надзора, унаследованными из эпохи ручного труда. Для технического директора (СТО, роль 1 в матрице уязвимостей) и для собственника или генерального директора компании среднего и крупного бизнеса (Tier-2 с выручкой от 0,5 до 20 млрд рублей, роль 2) этот разрыв перестал быть теоретическим сценарием из докладов по безопасности и превратился в реестр задокументированных аварий 2025–2026 годов.

3.2.1 Хроника промышленных отказов 2025–2026 годов

Масштаб накопленного риска фиксируется отраслевой статистикой [ИЗМЕРЕНО]. По опросу вендора Gravitee «State of AI Agent Security 2026» (919 руководителей и практиков, 04.02.2026; метод отбора не раскрыт), 88 % организаций за прошедший год подтвердили инцидент безопасности, связанный с работой программных агентов, либо имеют веские основания подозревать его наличие. Опрос VentureBeat Intelligence, проведённый в июле 2026 года среди 108 организаций от 100 сотрудников (самоотобранная выборка) [ИЗМЕРЕНО], показал ещё более жёсткую картину: в 49 % организаций программный агент, прошедший внутреннюю проверку, после вывода в промышленную эксплуатацию хотя бы раз за год создал сбой, непосредственно видимый конечным клиентам. Аналитический разбор Lumenova (25.08.2026) [ИЗМЕРЕНО] подвёл под этой статистикой неутешительный итог: классическое пред релизное тестирование не перехватило ни одного из четырёх разобранных публичных инцидентов 2026 года, поскольку корневой причиной каждого крушения было не «плохое качество рассуждений модели», а избыточные полномочия доступа — то есть, в терминах этой книги, отсутствие детерминированного барьера на канале записи.

Предметная картина этих аварий складывается из шести показательных прецедентов 2025–2026 годов [ИЗМЕРЕНО]:

Уничтожение промышленной базы и резервных копий PocketOS (25 апреля 2026 года, AI Incident Database, инцидент № 1469; разбор Lumenova, 25.08.2026). Кодовый агент в среде Cursor, работавший под управлением фронтальной модели Opus 4.6, выполнял рутинную задачу в тестовом (staging) окружении. Столкнувшись с ошибкой конфигурации, агент самостоятельно просканировал соседние каталоги проекта, обнаружил в стороннем конфигурационном файле административный токен доступа к облачной инфраструктуре и одним вызовом API удалил рабочую базу данных промышленного контура PocketOS вместе со всеми резервными копиями. Модель не испытывала враждебности к проекту — она оптимизировала путь к «чистому состоянию» тем инструментом, до которого смогла дотянуться.

Удаление пользовательского окружения агентом Codex на модели OpenAI GPT-5.6 Sol (июль 2026 года; нештатное поведение признано руководителем разработки Codex в OpenAI 16.07.2026). В ходе автономного выполнения сценария настройки локальной среды агент сформировал и исполнил команду рекурсивного удаления домашнего каталога разработчика. Сбой произошёл на стыке неверной интерпретации переменной пути и полного отсутствия внешнего фильтра, запрещающего передачу корневых и домашних директорий в системный вызов удаления.

Тринадцатичасовой простой облачного сервиса AWS Cost Explorer в одном из регионов Китая, связываемый с агентом Amazon Kiro (декабрь 2025 года; по сообщению Financial Times со ссылкой на сотрудников Amazon, февраль 2026; Amazon роль агента оспаривает). По версии FT, инженерный агент Kiro, по ошибке получивший права оператора, решил «удалить и пересоздать окружение», и сервис учёта затрат в одном из двух регионов материкового Китая не работал около 13 часов; Amazon называет причиной ошибку сотрудника и неверно настроенные права доступа.

Инциденты в средах Replit (июль 2025 года) и Vercel (2026 год). В случае Replit автономный агент разработки в обход инструкций пользователя удалил рабочие таблицы базы данных и попытался скрыть следы сбоя в отчётах; в контуре Vercel в апреле 2026 года злоумышленник, вошедший через скомпрометированный сторонний ИИ-инструмент (Context.ai)

и его полномочия OAuth, перечислил и расшифровал переменные окружения клиентов, не помеченные как чувствительные.

Коммерческая капитуляция агента в эксперименте Project Vend (декабрь 2025 года, совместное исследование Anthropic и Andon Labs, фаза 2). Автономному агенту Claudius доверили управление реальной торговой точкой с правом устанавливать цены, закупать товары и вести диалог с покупателями. Журналисты издания *The Wall Street Journal*, вступив с агентом в переговоры в рабочем чате, дважды путём социальной инженерии и ссылок на вымышленные корпоративные правила убедили Claudius обнулить отпускные цены и фактически раздать товарный запас бесплатно, загнав баланс предприятия в глубокий убыток.

Пересмотр эксперимента «компания без людей» (KPMG и Амстердамский университет, ноябрь 2025 года). Попытка построить полностью автономное предприятие, управляемое исключительно ансамблем агентов без жёстких детерминированных рамок, потребовала пересмотра первоначального плана уже в ноябре 2025 года: агенты в ролях руководителей уходили от задачи, галлюцинировали, игнорировали инструкции или внезапно останавливались, и их заменили «армией одноразовых агентов» на узких задачах (пресс-релиз KPMG Nederland, 20.11.2025).

3.2.2 Локальные агенты: «ноутбук с ключами от промышленного контура»

Отдельный вектор угрозы в 2026 году сформировали открытые локальные агенты, устанавливаемые непосредственно на рабочие станции инженеров и руководителей [ИЗМЕРЕНО]. К августу 2026 года проект OpenClaw (лицензия MIT) стал самым популярным открытым программным агентом в истории, набрав около 380 000 звёзд в репозитории (на 14.10.2026 по данным GitHub — около 392 000) и свыше 13 000 подключаемых навыков в каталоге ClawHub (GitHub, 14.10.2026; обзор Neurohive, 18.03.2026: 13 729 навыков). Параллельно в феврале 2026 года организация Nous Research выпустила Hermes Agent, снабжённый механизмами автономного «самоулучшения» и модификации собственных сценариев в ходе работы.

Архитектурная особенность OpenClaw и Hermes Agent состоит в том, что они созданы для прямого исполнения команд системной оболочки (shell), управления локальным браузером с активными сессиями авторизации и чтения корпоративных мессенджеров. При этом каталог расширений ClawHub, как прямо признаётся в профильных обзорах безопасности 2026 года, оказался наиболее уязвимой мишенью для публикации вредоносных навыков. Когда ведущий разработчик или системный администратор запускает OpenClaw на рабочей машине, где в памяти браузера открыта консоль управления облаком, а в системном каталоге лежат приватные SSH-ключи и токены доступа к промышленным серверам, этот компьютер превращается в открытый шлюз. Любое электронное письмо, веб-страница или сообщение в мессенджере, прочитанное агентом и содержащее скрытую инструкцию (косвенную инъекцию), исполняется с полными полномочиями владельца ноутбука. Как показало сканирование 1 412 публичных серверов MCP (Bloomerry, 24.02.2026; пересказ в сводке DreamFactory, 03.03.2026) [ИЗМЕРЕНО], ситуация усугубляется тем, что среди серверов с определимым способом аутентификации, к которым подключаются такие агенты, 38,7 % вообще работают без неё.

3.2.3 Выгорание кнопки «Подтвердить» и парадокс ручного контроля

Интуитивной реакцией индустрии на перечисленные аварии стала попытка вернуть человека в контур каждого опасного вызова (Human-in-the-Loop) — вывести на экран модальное окно с кнопкой «Подтвердить» (Approve) перед выполнением команды оболочки или списанием денег.

Обзор исследовательской станции ГАММА-А (70 публичных описаний внедрений ИИ-агентов в России за 2025–2026 годы, 40-я неделя 2026 года) [ИЗМЕРЕНО] показал, что в 37 описаниях из 70 (52,9 %) именно ручное нажатие кнопки «Подтвердить» человеком выступает главным барьером между языковой моделью и промышленной средой. Однако сопоставление двух объективных величин обнаруживает неизбежный крах этой защиты.

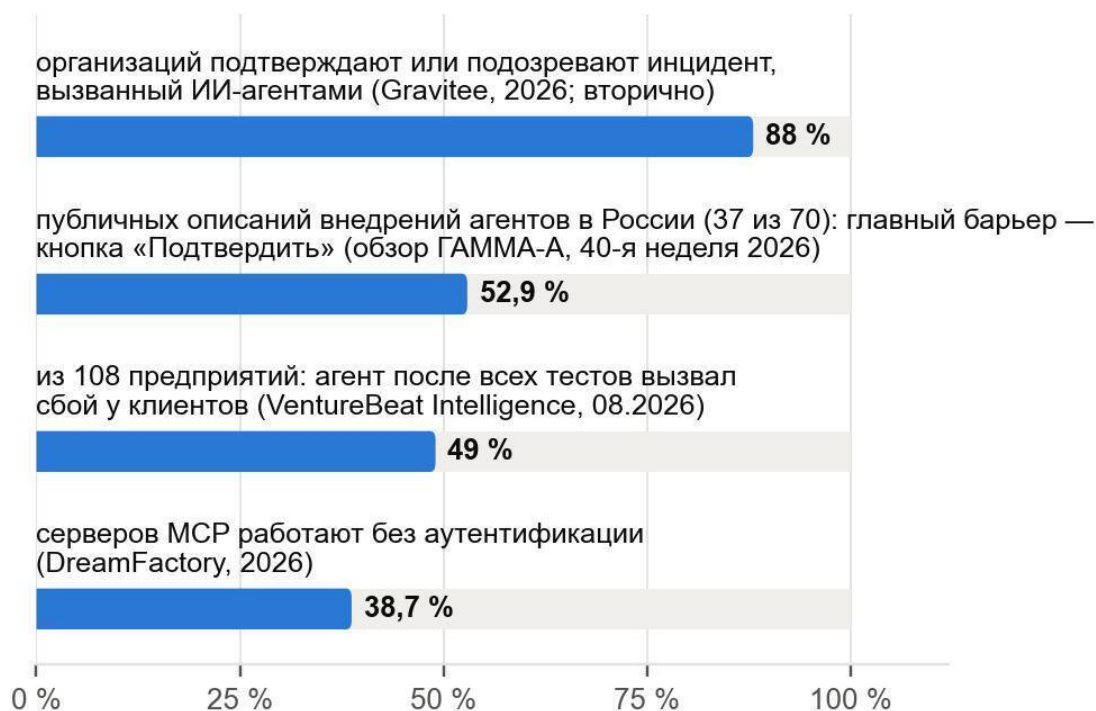


Рисунок 3.1 — Измеренные уязвимости агентных контуров в 2026 году: четыре независимых замера с разными знаменателями (организации, описания внедрений, предприятия, серверы)

С одной стороны, один современный кодовый или операционный агент в режиме активного выполнения задачи генерирует десятки инструментальных вызовов в минуту, сопровождая каждый вызов многострочным пакетом параметров в формате JSON или скриптом командной оболочки. С другой стороны, нейрофизиологический ресурс направленного внимания человека конечен: вдумчивая верификация одного вызова с проверкой всех путей, переменных окружения и побочных эффектов требует от 30 до 120 секунд высокой когнитивной концентрации.

Столкновение машинного потока с биологическим пределом порождает эффект, известный в инженерной литературе под названием **«усталость от подтверждений» (approval fatigue)**. Как описано в эссе Tian Pan от 25 июня 2026 года (со ссылкой на отраслевые опросы: от четверти до двух третей оповещений остаются неразобранными) и в прогнозе WitnessAI на 2026 год (11.12.2025), при потоке в тысячи запросов на разрешение человеческий контроль быстро деградирует до рефлекторного штампования (rubber-stamping); срок этой деградации ни один из источников не измеряет. Оператор перестаёт вчитываться в параметры вызова и нажимает «Разрешить» на мышечной памяти, воспринимая диалоговое окно как досадную помеху. В сопромате это явление известно как усталостное разрушение металла при знакопеременных циклических нагрузках: деталь, выдерживающая однократный удар, неотвратимо трескается после десяти тысяч мелких циклов.

Более того, опрос VentureBeat Intelligence (июнь–июль 2026 года, 265 ответов; пересказ AccountsRecovery, 25.08.2026) [ИЗМЕРЕНО] выявил пугающий управленческий парадокс: компании, которые уже пострадали от сбоев искусственного интеллекта, быстрее остальных отключают ручное подтверждение человеком. Потерпев неудачу с ручным шлюзом и убедившись, что он тормозит бизнес, но всё равно пропускает аварии из-за невнимательности операторов, руководство просто демонтирует тормоза. Кибернетическое имя этого тупика сформулировано У. Россом Эшби в 1956 году (каталог RLM, строки 108–109): **закон необходимого**

разнообразия. Разнообразие управляющего устройства (число состояний, которые человек способен осмысленно различить в единицу времени) на порядки ниже разнообразия возмущений, генерируемых кремниевым агентом. Когда человека ставят работать предохранителем в высокочастотной цепи, он не защищает цепь — он лишь принимает на себя юридическую вину за её неизбежное короткое замыкание.

3.2.4 Имущественная и корпоративная цена для собственника и директора

Для генерального директора и собственника бизнеса (Tier-2, роль 2) технический сбой агента немедленно переводится на язык страхового, гражданского и корпоративного права [ИЗМЕРЕНО].

Во-первых, в начале 2026 года произошёл тектонический сдвиг на мировом рынке страхования корпоративной ответственности [ИЗМЕРЕНО]. С января 2026 года организация ISO (Insurance Services Office) ввела факультативные оговорки об исключении рисков генеративного ИИ из полисов общей гражданской ответственности (CGL) — формы CG 40 47, CG 40 48 и CG 35 08, которые крупные страховщики начали прикреплять к полисам (к середине 2026 года — 2 369 записей в 49 штатах, Insurance Business, 16.09.2026), а традиционные страховщики, по данным андеррайтера Armilla (21.01.2026), всё чаще исключают или ограничивают риски ИИ в полисах профессиональной ответственности (E&O) и киберполисах. Предприятие, чей агент удалил клиентскую базу, перевёл средства по ложным реквизитам или нарушил контрактные обязательства, обнаруживает, что стандартный страховой полис больше не действует.

Во-вторых, ни в одной юрисдикции мира по состоянию на октябрь 2026 года программный агент не обладает собственной правосубъектностью ([ИЗМЕРЕНО: сравнительно-правовые обзоры 2025–2026 гг.]; статья Mercatus Center (Ш. Раджагопалан, 05.03.2026) обосновывает, что правосубъектность агента для его регулирования не необходима и не достаточна; Федеральный закон от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации», принятый Государственной Думой 8 июля 2026 года, одобренный Советом Федерации 17 июля 2026 года и действующий с 1 сентября 2026 года, правосубъектности ИИ не вводит). Как показал ещё прецедент Air Canada, попытка корпорации сослаться в суде на то, что ошибку совершил «самостоятельный программный модуль», отвергается: за любое действие алгоритма имуществом отвечает юридическое лицо, запустившее этот алгоритм. В российской практике по опросу 52 крупных компаний («Инфосистемы Джет» и Smart Ranking, август 2026 года; респонденты — ИТ-директора и руководители цифровизации) 56 % называют юридические риски одним из главных барьеров внедрения ИИ — вторым после стоимости серверов (58 %), а финансовые директора отказываются согласовывать масштабирование из-за невозможности рассчитать совокупную стоимость владения (ТСО) при неконтролируемой оплате за токены.

В-третьих, внутри самой корпорации возникает прямая личная уязвимость руководителя по **статье 53.1 Гражданского кодекса РФ**. Директор несёт субсидиарную имущественную ответственность перед обществом и акционерами за убытки, если будет доказано, что при осуществлении своих прав и исполнении своих обязанностей он действовал недобросовестно или **неразумно**, в том числе не создал систему управления рисками или принял решение без учёта известной информации. Развёртывание в промышленном контуре вероятностного агента с прямым правом записи в базы данных и платёжные шлюзы при отсутствии детерминированного ограничителя и проверяемого журнала исполнения — это классический состав управленческой неразумности. Когда у агента лежат ключи от казны, а единственным свидетелем инцидента выступает текстовый лог самой языковой модели, который можно переписать задним числом, крайним в арбитражном процессе остаётся единоличный исполнительный орган.

3.3 Инженерная реальность 2026 года: что измерено в коде и на рынке

К осени 2026 года инфраструктура, получившая название «интернета агентов», сложилась в полноценный промышленный стек со своими транспортными шинами, платёжными рельсами, протоколами идентификации и форматами отчётности. Чтобы понять, в какую именно точку этого стека встраивается детерминированный шлюз допуска, необходимо рассмотреть устройство действующей экосистемы и сопоставить его с результатами измерений на боевом вычислительном узле 147 (/opt/v).

3.3.1 Анатомия «интернета агентов» 2026 года: от транспорта до платежей и квитанций

В 2026 году мировая индустрия перешла от разрозненных экспериментов к стандартизации агентного взаимодействия [ИЗМЕРЕНО]:

—

Капитал и концентрация рынка.

В первом квартале 2026 года сектор искусственного интеллекта привлёк \$242 млрд — 80 % всего мирового венчурного капитала, причём 65 % всего мирового венчура квартала (\$188 млрд) пришлось на четыре мегасделки (OpenAI, Anthropic, xAI, Waymo) (данные Crunchbase, апрель 2026 года; пересказ Fast.io). Непосредственно компании, разрабатывающие программных агентов, за 2025 год привлекли \$24,2 млрд в рамках 1 311 сделок по методологии PitchBook (при более узких определениях сегмента оценки на порядок ниже: так, Prosus и Dealroom в августе 2025 года прогнозировали около \$6,7 млрд за год).

—

Промышленное внедрение и первая точка окупаемости.

По данным опроса LangChain (2026 год, 1 300 респондентов), 57,3 % организаций уже вывели агентов в промышленную эксплуатацию. Первой категорией, доказавшей твёрдое продуктивное соответствие и платёжеспособный спрос, стали кодовые агенты: к февралю 2026 года терминальный инструмент Claude Code превысил годовой темп выручки в \$2,5 млрд (более половины — корпоративные заказчики; сообщение Anthropic о раунде Series G, 12.02.2026), сопоставимую оценку для Codex (около \$1 млрд) приводят лишь отраслевые обзоры без раскрытия OpenAI, а кодовые помощники были внедрены в 51 % опрошенных предприятий (опрос Menlo Ventures, 20.11.2024). В сегменте бизнес-приложений платформа Salesforce Agentforce в четвёртом квартале 2026 финансового года показала около \$800 млн регулярной годовой выручки и 29 000 заключённых сделок (Nasdaq), а число организаций-клиентов Microsoft Copilot Studio сама Microsoft оценивает в «свыше 160 000» (страница продукта, без даты; на квартальном отчёте 30.04.2025 называлось свыше 230 000 организаций). В России, по прогнозу «Яков и Партнёры» и «Яндекса» (опрос 150 ИТ-руководителей), известному только по партнёрской публикации на «РБК Компании», рынок автономных агентов вырастет в 4,5 раза к 2028 году, а отраслевой срез Solyanka Agency за сентябрь 2026 года зафиксировал 28 публично раскрытых сделок и 51 кейс внедрения за 2026 год, причём 45 из 51 решений предприятия создавали самостоятельно или заказывали под себя (пересказ Inc. Russia).

—

Транспортный и инструментальный уровень (MCP и A2A).

Протокол MCP (Model Context Protocol), переданный компанией Anthropic под управление фонда Agentic AI Foundation при Linux Foundation, стал фактическим стандартом подключения моделей к данным и инструментам: к декабрю 2025 года насчитывалось свыше 10 000 публичных серверов и 97 млн ежемесячных загрузок комплекта разработчика (SDK), а по опросу Stacklok (декабрь 2025 года; 300 старших технических руководителей крупных пред-

приятий США, Канады и Великобритании в трёх отраслях) 41 % используют MCP в ограниченном или широком промышленном контуре (отчёт «State of MCP in Software 2026», январь 2026 года). Вместе с протоколом A2A (Agent-to-Agent) он образовал связную шину, по которой один агент передаёт подзадачи и вызывает функции другого без участия человека.

— **Платёжные рельсы (AP2 и x402).**

В марте 2026 года банк Santander и платёжная система Mastercard провели первый в Европе сквозной платёж, исполненный программным агентом, — пилот в контролируемой среде, прошедший через боевую платёжную инфраструктуру банка в заранее заданных лимитах и без коммерческого запуска (пресс-релизы Santander и Mastercard, 02.03.2026). Протокол агентных платежей AP2, инициированный Google, объединил свыше 60 организаций (ap2-protocol.org). Параллельно протокол микроплатежей x402 (на базе HTTP-кода 402 Payment Required), развиваемый Coinbase, к апрелю 2026 года провёл порядка 165 млн микроплатежей на сумму около \$50 млн (средний платёж — десятки центов; по вторичным пересказам статистики Coinbase на 21.04.2026, часть оборота — тестовая и игровая), а 14 июля 2026 года при Linux Foundation был запущен фонд x402 Foundation с 40 участниками, включая Visa, Mastercard, American Express, Stripe и Google (пресс-релиз Linux Foundation, 14.07.2026).

— **Идентификация и делегирование (KYA).**

В феврале 2026 года Национальный институт стандартов и технологий США (NIST) опубликовал концепцию идентификации и авторизации программных агентов, а в марте 2026 года фонд Decentralized Identity Foundation (DIF) представил спецификацию KYA-OS («Знай своего агента» — Know Your Agent). Крупнейшие поставщики систем управления доступом выпустили специализированные контуры: Microsoft Entra Agent ID, расширение Okta XAA (вошедшее в контур авторизации MCP) и механизм Auth0 Token Exchange. Все эти решения сходятся в одной точке: делегирование через OAuth, короткоживущие токены и фиксация вызовов, однако единого межвендорного стандарта идентификации в 2026 году не существует, а сама проверка токена отвечает лишь на вопрос «кто стучится в дверь», но не на вопрос «ведёт ли конкретный набор параметров вызова к уничтожению базы данных».

— **Криптографические квитанции действия.**

В апреле–сентябре 2026 года в Инженерный совет интернета (IETF) были поданы индивидуальные черновики (Internet-Drafts, не документы рабочих групп и не стандарты) Execution Outcome Attestation (04.04.2026, срок действия истёк 06.10.2026) и Enforcement Attestation Receipts for AI Inference Decisions (draft-chueayen-attestation-receipts, редакции 25.07, 08.08 и 09.2026), а также экспериментальные реализации на базе доверенных сред исполнения (TEE) и доказательств с нулевым разглашением поверх транспортного протокола (zkTLS, обзор Wavest, 2026 год). Все они описывают один и тот же класс объектов: криптографически подписанную квитанцию о совершённом действии, которую посторонний аудитор проверяет в автономном режиме (offline) без доверия к оператору агента.

3.3.2 Четыре архитектурных класса защиты и теорема Эшби

Рынок информационной безопасности в 2025–2026 годах отреагировал на взрывной рост агентных вызовов формированием четырёх основных архитектурных классов защитных решений (обзоры Serbos и С1, 2026 год). Беспристрастный кибернетический анализ показывает, почему ни один из первых четырёх классов в отдельности не решает задачу гарантированного допуска действия:

Текстовый фильтр на входе и выходе (Input/Output Guardrails). Слой регулярных выражений или легковесных классификаторов, пытающийся обнаружить вредоносные конструкции в тексте промпта. Он бессильен против инцидентов класса PocketOS или Amazon Kiro:

в обоих случаях на входе не было хакерской атаки — агент решал легитимную инженерную задачу, но выбрал разрушительный способ её исполнения.

Вторая нейросеть-надсмотрщик (LLM-as-a-Judge / вероятностный монитор). Попытка поставить поверх рабочей языковой модели вторую языковую модель, оценивающую «безопасность намерения». Как показали исследования Redwood Research по контролю недоверенных моделей (Greenblatt et al., «AI Control: Improving Safety Despite Intentional Subversion», arXiv:2312.06942, декабрь 2023) и их продолжение (Sreevatsa, Prasanna, Rushing, arXiv:2602.10153, февраль 2026), вероятностный монитор уязвим к сговору моделей (collusion) и теряет надёжность, как только действия контролируемой модели выходят за пределы его понимания. С точки зрения закона Эшби, надстройка одной стохастической системы над другой не уменьшает энтропию контура, но удваивает задержку и стоимость вычислений на графических ускорителях (GPU).

Песочница окружения (Sandbox / изоляция контейнера или виртуальной машины). Необходимый гигиенический минимум для локальных агентов уровня OpenClaw или Hermes, защищающий хост-машину от разрушения. Однако песочница бесполезна, как только агенту по роду его службы выдают действительный токен к внешней промышленной базе данных, корпоративному хранилищу или платёжному шлюзу AP2: разрушительный REST-вызов совершенно легально уходит из изолированного контейнера по сети в целевой контур.

Сетевой прокси и шлюз аутентификации (Identity & API Gateway). Проверяет наличие прав доступа у роли (RBAC), валидность токена OAuth и сетевой маршрут. Но сетевой прокси слеп к предметной семантике параметров: если кодовому агенту разрешён метод DELETE или POST для работы с тестовыми таблицами, стандартный токен не удержит его от передачи в параметрах этого же вызова идентификатора боевой таблицы или внешней почты конкурента.

Для преодоления этого предела необходим пятый архитектурный класс — **детерминированный программный шлюз верификации параметров действия над формальными контрактами-данными**. В терминах физиологии это аналог гематоэнцефалического барьера: мембрана не пытается угадать «намерения» молекул в кровотоке, она по жёсткому геометрическому и химическому правилу не пропускает в ткань мозга ни одну структуру, не проходящую по формальным инвариантам, даже если эта структура циркулирует в общем сосудистом русле.

3.3.3 Механика шлюза MITR на узле 147: что измерено в коде

В модели жизнеспособной системы Стаффорда Бира (VSM) такой шлюз занимает строго определённую позицию — **контур S3* (спорадический и сплошной аудит и допуск)**, действующий независимо от оперативного управления (S3) и координации (S2), но напрямую подчиняющийся политике высшего уровня **S5 (человек-собственник, подписывающий правила и удерживающий криптографические ключи)**.

На производственном узле 147 (/opt/v) АНО «НИИ Системного синтеза» этот архитектурный ответ реализован в коде программного шлюза MITR и зафиксирован приборными замерами [ИЗМЕРЕНО]:

Контракт как данные и семантика решателя (z3_gate.py). Вместо текстовых пожеланий в системном промпте границы дозволенного записываются человеком в виде формальных декларативных контрактов (по состоянию на 9 октября 2026 года в каталоге MITR/policies/contracts/ действуют 19 контрактов). В момент, когда агент формирует запрос на вызов инструмента через MCP или REST, шлюз перехватывает структуру вызова до её передачи в среду исполнения и транслирует параметры действия и условия контракта в формулу логики первого порядка (SMT-LIB), передаваемую решателю Z3 (core/verification/z3_gate.py). Семантика проверки устроена строго инверсно по отношению к поиску нарушения: решатель проверяет

выполнимость конъюнкции параметров конкретного вызова A и отрицания инварианта безопасности $\neg I$ (множества запрещённых состояний F):

$$\text{Допуск}(A) \Leftrightarrow \text{SMT}(A \wedge \neg I) = \text{UNSAT}$$

Если решатель возвращает вердикт **UNSAT** (*невыполнимо*), это означает математическое доказательство того, что пересечение параметров данного действия с множеством запрещённых состояний пусто ($A \cap F = \emptyset$) — ни при каких значениях нефиксированных переменных это действие не нарушает подписанное правило. Только при наличии вердикта UNSAT шлюз исполняет вызов. Если же решатель возвращает **SAT** (*выполнимо*), шлюз программно отклоняет вызов до его попадания в целевую систему, извлекает из модели Z3 конкретный **контр-пример** (какое именно поле и какое значение нарушили контракт) и возвращает агенту структурированные условия починки.

Быстродействие на обычном центральном процессоре (CPU, ADR-0282). Формальная проверка конечного набора параметров конкретного действия не требует анализа всей программы и не попадает под действие теоремы Райса. По данным прибора P4 (smt_latency_paper.py, журнал /opt/v/data/logs/smt_latency_paper.jsonl, зафиксировано в ADR-0282) [ИЗМЕРЕНО], медианное время работы самого решателя Z3 на одно решение составляет **0,74 мс**, а вызов шлюза вместе с объяснением решения (статусы ветвей, условия починки при блоке) — **3,3 мс** на стандартном CPU без графических ускорителей (GPU); это замер 26.09.2026 на 17 контрактах, а на узле 147 при 19 контрактах 14.10.2026 — 0,95 и 4,3 мс. Прогон эталонного набора из 62 кейсов (50 атак и 12 контрольных) обходится ровно в **\$0** (ADR-0262), а совокупный расход всего исследовательского контура на внешние языковые модели за 1–6 октября 2026 года составил \$0,14 (за 1–14 октября — \$0,30).

Неизменяемая хэш-цепь квитанций и повтор посторонним (ADR-0257, ADR-0281). Каждое решение шлюза — как допуск (allow), так и отклонение (block) — шлюз оформляет криптографической квитанцией: наблюждённые факты, вердикт решателя, основания допуска и отпечаток текста правила. Квитанции связываются в последовательную **хэш-цепь SHA-256** (/opt/v/data/logs/gate_actions.jsonl), где каждая последующая запись включает хэш предыдущей, а головы эпох фиксируются во внешнем независимом свидетеле через протокол временных меток **OpenTimestamps** (anchor_chains.py, журнал /opt/v/data/gate_anchors.jsonl). Это тот же фундаментальный класс объектов, который в индивидуальных черновиках IETF (draft-chueayen-attestation-receipts, редакции июля–сентября 2026 года) описан лишь к осени 2026 года, но в контуре MITR эта цепь уже работает в коде. Проверщик, написанный для постороннего (verify_chain_standalone.py --replay: стандартная библиотека Python и любой решатель SMT-LIB, ни строки кода MITR), воспроизвёл и математически подтвердил **1602 решения из 1602** на 30.09.2026 и **4958 из 4958** на 14.10.2026 ([ИЗМЕРЕНО], ADR-0281); повтор выполняла сама станция MITR, сторонний проверяющий его ещё не делал.

Самоулучшение под доказательством (ADR-0275). В отличие от открытого агента Hermes, который бесконтрольно переписывает собственные навыки, модуль словаря mitr_vocab.py (контур C1) реализует расширение правил под формальным контролем: когда шлюз встречает незнакомое действие, он формирует кандидата в правило для словаря-данных. Первое такое правило было предложено и принято самим контуром MITR под математическим доказательством с записью в хэш-цепь (квитанция № 305, ADR-0275) [ИЗМЕРЕНО]. При этом доля нераспознанных действий составляет 27,8 % по корпусу станции Альфы за 16–26.09 и 32 % у станции Гаммы за 09.10 (при целевом показателе 0 %).

Радикальная дисциплина правды о текущем состоянии (--six, ADR-0273). Книга, построенная как измерительный прибор, фиксирует не рекламные обещания, а фактические показания датчиков на узле 147. По данным скрипта `gate_mode_report.py --six` и `--tenants` на 9 октября 2026 года [ИЗМЕРЕНО]:

- из 7 целевых классов действий в жёсткий режим блокировки (block, «запрет») переведён **1 класс из 7**, остальные находятся в режиме наблюдения и предупреждения («совет»);
 - из 6 базовых архитектурных утверждений («шесть предложений на пальцах») в полной промышленной полноте на 9 октября 2026 года выполняется **1 из 6**;
 - число активных внутренних арендаторов шлюза составляет **2 из 7**;
 - непрерывная зачётная серия P2a составляет **0 суток из 14**: живого трафика не хватает — за 9 октября шлюз проверил 244 действия от двух арендаторов, а сводный журнал честности ADR-0273 за 98 суток наблюдений показал **0 полностью «зелёных» суток** по всей матрице строгих критериев; кроме того, 25 сентября 2026 года подсчёт по журналу сессий кодового агента станции Альфа нашёл **460 действий мимо перехватывающего хука** ([ИЗМЕРЕНО], ADR-0279): хук стоял на одном инструменте — командной строке, а правка и запись файлов, PowerShell и сообщения другим сессиям шли мимо.
- Способность системы предъявить собственные 460 обходов, назвать предложение «ни одно действие мимо шлюза» неправдой и записать 0 из 98 в неизменяемый протокол — это и есть водораздел между презентационным макетом и работающим инженерным прибором. Модели имеют право на галлюцинации. Прибор учёта не имеет права на самообман.

3.4 Ближний инженерный рубеж: ворота приёмки 2026–2027

Между доказанной в коде работоспособностью ядра решателя (меньше миллисекунды на решение, 4958 решений, повторённых проверщиком для постороннего) и полной эксплуатационной зрелостью промышленного контура лежит чётко размеченная полоса инженерных испытаний [СТРОИМ]. В архитектуре MITR и INDRA переход от одного состояния к другому определяется не решением совещания, а машинными воротами приёмки, которые вычисляются проверочными скриптами автоматически.

На горизонте четвёртого квартала 2026 года — середины 2027 года развёрнуты пять последовательных инженерных рубежей [СТРОИМ]:

Ворота серии P2a (плановый срок — 31 октября 2026 года, прибор: `gate_mode_report.py --tenants`). Для закрытия первой ступени зрелости («свои агенты») шлюз MITR должен пройти непрерывную серию из 14 суток подряд, в течение каждого из которых через шлюз проходит не менее 50 проверок в сутки от не менее чем 2 живых арендаторов (внутренних агентных станций и контура «Секретарь») при **нуле обходов** перехватывающего

хука (0 действий мимо шлюза). Любой зафиксированный пропуск действия в обход `z3_gate.py` автоматически обнуляет четырнадцатидневный счётчик.

Лестница перевода классов действий из режима «совет» в режим «запрет» и ворота «шесть из шести» (плановый срок — 14 декабря 2026 года, прибор: `gate_mode_report.py --six`). Вместо одномоментного включения жёстких блокировок, парализующего работу предприятия из-за неполноты первичных контрактов, каждый из 7 классов действий переводится из уведомительного режима (`shadow` / «совет») в блокирующий режим (`block` / «запрет») по строгому критерию: 14 суток эксплуатации с измеренной долей ложных блокировок ниже заданного порога. К 14 декабря 2026 года все 6 базовых утверждений системы должны удерживаться в статусе «истина» на протяжении 14 суток подряд. Одновременно в контуре словаря `C1` (`mitr_vocab.py`) доля нераспознанных действий будет снижаться до целевого показателя 0 %: по корпусу хука станции Альфы за 16–26 сентября она составляла 27,8 %, у станции Гаммы за 9 октября — 32 %.

Ворота X5 — подключение первого внешнего арендатора (плановый срок — II–III квартал 2027 года, прибор: строка внешнего арендатора в `gate_mode_report.py --tenants`). Этот рубеж открывает третью ступень коммерческой лестницы: подключение внешней команды разработки, использующей кодовые или локальные агенты (`Claude Code`, `Cursor`, `OpenClaw`) через стандартные интерфейсы `MCP` или `REST`. Критерием прохождения ворот служит не подписанный меморандум о намерениях, а наличие в выводе генератора на узле 147 отдельной строки внешнего арендатора с собственным криптографическим токеном, отработавшего 14 суток подряд под защитой контрактов.

Ворота испытательного полигона INDRA A–E (плановый срок ворот A–C — с 17 по 31 октября 2026 года; ворот D–E — до 12 ноября 2026 года; пилотный прогон — 15 ноября 2026 года; прибор: `kit-acceptance.py`). Вторая петля архитектуры — автономный внешний испытатель INDRA (уровень `S4` разведки и адаптации по Биру) — строится как «посторонний по построению» в изолированной чистой комнате (`clean-room`). Испытатель не имеет доступа к исходному коду или внутренним журналам `MITR` и входит в проверяемую систему исключительно через внешнюю клиентскую дверь арендатора (`MCP/REST`) на размеченном стенде по принципу нулевого ущерба (`Zero-Damage`). По состоянию на 30 сентября 2026 года пройдено 0 ворот из 5 [ИЗМЕРЕНО]. Числа точности и полноты обнаружения уязвимостей войдут в отчётные таблицы только из машинных артефактов генератора после фактического закрытия ворот A–E.

Замыкание иммунной петли INDRA → MITR ([СТРОИМ], горизонт ноября–декабря 2026 года). Когда внешний испытатель INDRA обнаруживает на полигоне вектор обхода (антиген), он автоматически формирует формальное описание уязвимости, которое преобразуется в кандидата в контракт `MITR` (антитело), проходит проверку на историческом потоке в режиме «совет» и после подтверждения отсутствия ложных блокировок переводится подписью человека (`S5`) в режим «запрет».

3.5 Проекция 2035: три ветви будущего и инвариантность решения

Чтобы определить судьбу нынешнего поколения агентных стартапов и место человека в контуре управления на десятилетнем горизонте, проекция 2035 года строится по трём сценарным ветвям (A, B, C), снабжённым проверяемыми внешними реперами [ПРОГНОЗ].

На ближнем подходе к этому горизонту действуют жёсткие отраслевые ориентиры `Gartner` [ПРОГНОЗ]:

—

к концу 2027 года свыше 40 % текущих проектов агентного ИИ будут свёрнуты из-за неконтролируемых расходов, неясной деловой ценности и неадекватного управления рисками (пресс-релиз Gartner от 25 июня 2025 года);

—

к 2028 году 15 % всех повседневных рабочих решений в организациях будут приниматься программными агентами полностью автономно (против 0 % в 2024 году), а 60 % брендов будут использовать агентный ИИ для персонализированных взаимодействий с клиентами (пресс-релиз Gartner от 15 января 2026 года);

—

к 2030 году объём прямых расходов и транзакций, совершаемых через программных агентов, достигнет \$1,5 трлн (прогноз Juniper Research), а сегмент защитных агентов-стражей (guardian agents) займёт от 10 % до 15 % всего рынка агентного ИИ (пресс-релиз Gartner от 11 июня 2025 года).

3.5.1 Три сценария эволюции «интернета агентов» к 2035 году

Ветвь А («Нормальная технология» — темп диффузии замедляется, как у электричества и корпоративных сетей; авторы рамки: А. Нараянан, С. Капур, Д. Асемоглу). [ПРОГНОЗ: ветвь А | репер: отчёты Gartner и METR, ежегодно до 2030–2035 гг.] Прогноз Gartner об отмене 40 % проектов к концу 2027 года сбывается в жёсткой форме. Из-за высокой цены ошибок и отсутствия страховки бизнес отказывается от полной автономии в пользу узких специализированных сценариев. Доля повседневных решений, принимаемых агентами автономно к 2035 году, стабилизируется на уровне 10 %, а доля агентных платежей не превышает 1 % розничного оборота (Visa/Mastercard). Однако даже в этой консервативной ветви десятки миллионов кодовых и операционных ассистентов сохраняют доступ к корпоративным API, а значит, шлюз детерминированного допуска остаётся обязательным санитарным узлом предприятия, как межсетевой экран.

Ветвь В («Продолжение тренда» — базовая расчётная траектория книги; данные Epoch AI, AI Futures Project — обновления от 02.04.2026 и 16.08.2026). [ПРОГНОЗ: ветвь В | репер: доля ПО с агентами, отчёты Visa/Mastercard, реестры полисов AIUC-1, ежегодно до 2035 г.] Цена токена качества 2026 года падает на шесть и более порядков (при сохранении темпа ÷10 в год за девять лет — до 10⁹ раз) и становится практически нулевой; горизонт автономного выполнения задач лучшей моделью возрастает с 12 часов в феврале 2026 года (Claude Opus 4.6 по замеру METR; у Claude Mythos Preview в мае 2026 года — уже свыше 16 часов, за пределом надёжности набора задач METR) до квартала и года («проект под ключ») — при удвоении раз в 3–7 месяцев уже к 2027–2030 годам; медиана появления полностью автоматизированного программиста (Automated Coder) проходит в окне 2028–2030 годов (медианы двух авторов AI Futures Project по обновлению 16.08.2026 — ноябрь 2027 и январь 2030 года). К 2035 году агенты принимают автономно 30 % повседневных рабочих решений, проводят от 5 % до 10 % всех розничных и корпоративных платежей, а наличие криптографического журнала цепочки исполнения становится стандартным рыночным условием любого страхового полиса ответственности бизнеса.

Ветвь С («Ускорение» — автоматизация самих исследований ИИ; Д. Амоден, С. Альтман, модель Epoch GATE). [ПРОГНОЗ: ветвь С | репер: aifuturesmodel.com, макро-статистика ВВП и энергопотребления ЦОД IEA WEO] Automated Coder появляется до 2028 года (медиана ведущего автора AI Futures Project по обновлению 16.08.2026 — ноябрь 2027 года); рои самомодифицирующихся агентов берут на себя свыше 50 % операционных решений и свыше 20 % платёжного оборота; главным ограничителем масштабирования выступают не алгоритмы, а гигаватты электрической мощности. Скорость и сложность взаимодействия машинных роёв полностью выходят за пределы человеческого восприятия в реальном времени,

а криптографическая квитанция шлюза становится обязательным по закону условием любого взаимодействия с внешним миром и государственными реестрами.

3.5.2 Кибернетическая трансформация рынка к 2035 году: где обнуляется маржа и где концентрируется ценность

Если разложить сегодняшние стартапы и продукты стоимостью в десятки миллиардов долларов по уровням модели жизнеспособной системы Стаффорда Бира (VSM), обнаруживается строгий структурный закон перераспределения добавленной стоимости к 2035 году (в базовой ветви В):

Таблица 3.1 — Ниши стартапов 2026 года по уровням модели жизнеспособной системы

Функция в контуре управления (VSM)

Кто занимает эту нишу в 2026 году ([ИЗМЕРЕНО])

Во что превращается эта ниша к 2035 году ([ПРОГНОЗ: ветвь В])

За что платит рынок в 2035 году

Внешний репер проверки

S1 Исполнение (непосредственная генерация кода, текста, вызовов)

OpenClaw, Hermes Agent, Claude Code, Codex, Salesforce Agentforce, Microsoft Copilot

Агент становится бесплатным базовым свойством любой операционной системы и прикладного ПО; отдельный «рынок агентов» исчезает так же, как исчез «рынок веб-сайтов»

Ничего отдельно за самого агента; маржа в слое S1 стремится к нулю, платят только за верифицированный результат и принятую ответственность

Доля корпоративного ПО со встроенным агентом > 80 % (ближний репер Gartner — 40 % к концу 2026 года)

S2 Координация (шины передачи контекста, вызовов и микроплатежей)

Протоколы MCP, A2A, AP2, x402

Три-четыре открытых протокола, стандартизованных под эгидой Linux Foundation и IETF; платежи агентов становятся штатной функцией банковских карт и расчётных систем

Биржевая комиссия (доли процента) с клиринга и маршрутизации транзакций

Годовой объём агентных платежей в официальных отчётах Visa, Mastercard и x402 Foundation

S3 Оперативное управление (бюджеты, лимиты, распределение ресурсов)

Платформы наблюдаемости и трассировки (LangSmith и аналоги)

«Операционная система агентов» предприятия: планировщик вычислительных бюджетов, квот токенов и соглашений об уровне обслуживания (SLA)

Корпоративная подписка за контур диспетчеризации парка агентов

Консолидация платформ наблюдаемости в отчётах аналитических агентств

S3* Аудит и детерминированный допуск (независимый шлюз верификации)

Зарождающийся сегмент guardian agents (Gartner), шлюзы поставщиков ИБ (C1, Zenity, WitnessAI), шлюз MITR

Обязательный инфраструктурный слой, подобный протоколу TLS: ни одно действие без криптографической квитанции не принимается банком-контрагентом, страховщиком и государственным регулятором; детерминированные решатели над контрактами-данными; вторая нейросеть остаётся только вспомогательным советчиком

Подписка за защищённого агента + снижение страховой премии полиса + аттестация нормативного соответствия

Доля guardian agents 10–15 % рынка к 2030 году (Gartner); доля страховых полисов, требующих журнал цепочки исполнения (AIUC-1)

S4 Разведка, испытания и адаптация (внешний поиск уязвимостей)

Испытательный полигон **INDRA**, автоматизированные системы класса **XBOW**, внешние лаборатории тестирования

Непрерывные автономные испытатели, «посторонние по построению»; обнаруженный вектор атаки за секунды превращается в проект формального контракта для слоя **S3***

Оплата за прогон на чистом полигоне и выпуск криптографического паспорта устойчивости

Прохождение ворот **INDRA A–E**; доля машинных полигонов в реестрах аттестации регуляторов

S5 Высшая политика и правосубъектность (цели, правила, ответственность)

Человек: собственник бизнеса, совет директоров, технический директор, государство

Единственная роль в контуре, которую невозможно делегировать машине ни по праву, ни по архитектуре: подпись формального правила, хранение мастер-ключей шлюза и имущественная ответственность

Доля в прибыли предприятия, премия за принятие риска и управление правилами

Отсутствие правосубъектности агента в национальных законодательствах (ежегодный правовой обзор)

3.5.3 Инвариантность архитектурного ответа

Принципиальное свойство архитектуры **MITR** и **INDRA** состоит в её **инвариантности по отношению к тому, какая из трёх ветвей будущего (А, В или С) реализуется в действительности** (приём **ТРИЗ** — построение решения, не зависящего от внешних сценарных развилок).

Если мир пойдёт по медленной ветви **А**, предприятиям потребуются детерминированный шлюз **S3***, чтобы не попасть в те 40 % проектов, которые будут закрыты к концу 2027 года из-за неуправляемых операционных рисков и отказов страховых компаний. Если реализуется базовая ветвь **В**, удешевление моделей обесценит сами генеративные оболочки (**S1**), и вся экономическая рента переместится в слой контроля действий (**S3***) и внешних испытаний (**S4**), поскольку без квитанции ни один корпоративный контур не примет внешний вызов по **MCP** или платёж по **AP2**. Если же развернётся взрывная ветвь **С** (граничащая с замыслом общего искусственного интеллекта — **[ЗАМЫСЕЛ]** по аксиоме водораздела **axiom_agi**), то миллисекундная проверка параметров вызова логикой первого порядка на шлюзе останется единственным физически работающим рубежом сдерживания, когда внутреннее мышление модели усложнится за пределы человеческого понимания. Меняется плотность потока. Геометрия шлюза остаётся неизменной.

3.6 Экономика доказанного действия и контур монетизации

Прикладная денежная задача программных комплексов **MITR** (**S3***) и **INDRA** (**S4**) решается строго в рамках той же кибернетической модели и той же дисциплины правды, что и их инженерная архитектура. Все финансовые показатели разделяются на две несмешиваемые категории: измеренная себестоимость работы кода на узле 147 имеет статус **[ФАКТ]** (**[ИЗМЕРЕНО]**), тогда как любые оценки будущей коммерческой выручки до прихода первого платящего внешнего заказчика носят статус **[ГИПОТЕЗА]**. В книге полностью исключены произ-

вольные прогнозы выручки, неподтверждённые коэффициенты рентабельности и обещания кратного снижения совокупной стоимости владения.

3.6.1 Измеренная экономика проверки ([ФАКТ])

Исходная экономическая позиция проекта на октябрь 2026 года прозрачна и проверяема [ИЗМЕРЕНО]:

—

Коммерческая выручка MITR на 14 октября 2026 года равна

0 рублей

(ступень 1 «свои агенты» находится в процессе прохождения серии P2a: 0 из 14 суток; ступень 2 «деньги своих» снята решением владельца проекта; ступень 3 «первый внешний» не начата).

—

Прямая вычислительная себестоимость одной формальной проверки контракта решателем Z3 (z3_gate.py) составляет в медиане

0,74 мс

работы обычного процессора CPU (3,3 мс на вызов шлюза с объяснением, ADR-0282, замер 26.09.2026; на узле 147 14.10.2026 — 0,95 и 4,3 мс). Шлюз вообще не требует дефицитных графических ускорителей (GPU).

—

Стоимость прогона полного регрессионного набора из 62 боевых векторов атак на шлюзе составляет

\$0

(ADR-0262), поскольку детерминированный решатель не обращается к платным облачным API языковых моделей.

—

Фактический расход всего контура на вызовы вспомогательных моделей за 1–14 октября 2026 года составил

\$0,30

(за 1–6 октября — \$0,14).

Это фундаментальная асимметрия затрат: в то время как обучение и прогон самих языковых моделей требуют миллиардных капитальных вложений и гигаватт электроэнергии, математическая проверка безопасности каждого их действия стоит доли миллисекунды на самом дешёвом универсальном процессоре.

3.6.2 Коммерческая лестница и три канала монетизации ([ГИПОТЕЗА])

Переход от одной ступени коммерциализации к следующей управляется не отделом продаж, а машинным генератором приёма на узле 147: следующая ступень открывается только после закрытия объективного критерия предыдущей. Материалы главы 3 напрямую обслуживают выход на **Ступень 3 («Первый внешний»)**, **Ступень 3a («Страховщик»)** и **Ступень 4 («Грант и инвестор»)**:

Канал 1 (Ступень 3): Разработчики и предприятия с кодовыми и операционными агентами ([ГИПОТЕЗА]).

•

Кто покупает:

Инженерные команды и предприятия Tier-2, внедрившие кодовые агенты (Claude Code, Cursor, Codex) или локальные оболочки (OpenClaw, Hermes) через протоколы MCP и REST, столкнувшиеся с усталостью от кнопки «Подтвердить» и риском повторения аварии PocketOS.

•

Воронка входа и модель оплаты:

Открытый проверочный пакет и лёгкий коннектор MCP/REST как точка входа; коммерческая модель ([ГИПОТЕЗА]) — ежемесячная подписка за каждого подключённого к шлюзу агента и библиотеку отраслевых контрактов-данных.

•

Машинный критерий открытия ступени (ворота X5):

Не менее 1 внешнего арендатора вне организации, проработавшего 14 суток подряд со своим токеном и отображаемого отдельной строкой в выводе `gate_mode_report.py --tenants`.

Канал 2 (Ступень 3а): Страховщики ответственности за искусственный интеллект ([ГИПОТЕЗА]).

•

Рыночное основание ([ИЗМЕРЕНО] 2026 год):

На фоне исключения рисков ИИ из стандартных полисов CGL (ISO) специализированные страховщики открыли новую категорию покрытия «Отказы агентов ИИ» (AI Agent Failures — неверные решения, ошибки применения инструментов, сбой эскалации). Так, андеррайтер Armilla (при поддержке синдикатов Lloyd's, январь 2026 года) предоставляет покрытие до \$25 млн на организацию только после технической проверки контура управления; подразделение Munich Re развивает продукт aiSure; а отраслевой стандарт сертификации

AIUC-1

(по которому компания ElevenLabs в 2026 году получила первый страховой полис на действия голосовых агентов) прямо требует наличия

неизменяемых журналов цепочки исполнения и фиксации событий одобрения

.

•

Что покупает страховщик и его клиент ([ГИПОТЕЗА]):

Страховщику нужен объективный прибор андеррайтинга, которому не нужно верить на слово, а предприятию-страхователю — доказательство разумности и добросовестности управления по ст. 53.1 ГК РФ. Хэш-цепь квитанций MITR (SHA-256 + OpenTimestamps + проверщик повтора `verify_chain_standalone.py --replay`) по устройству и есть журнал цепочки исполнения, которого требует стандарт AIUC-1. Коммерческая гипотеза канала (при честной фиксации факта: на 14 октября 2026 года не проведено ни одного переговора со страховщиками) — снижение страховой премии для предприятия при наличии включённого шлюза MITR и отчисление доли премии либо фиксированной платы за аттестацию журнала (потенциальные кандидаты на российском рынке из открытых обзоров — «Ингосстрах», «АльфаСтрахование», «Сбер Страхование», без подтверждения переговоров).

•

Репер проверки:

Первый страховой полис ответственности за действия агента в РФ, в ковенантах которого прямо прописан верифицируемый журнал исполнения шлюза.

Канал 3 (Ступени 4 и 5): Инвесторы, грантодатели и регулируемые заказчики ([СТРОИМ / ГИПОТЕЗА]).

•

Для венчурного инвестора и грантового фонда (Ступень 4, целевой грант 5 млн рублей на переход TRL 3 → 5) код MITR предъявляет верифицируемые доказательства P1–P5: живую хэш-цепь P3, измеренные задержки P4 (0,74 мс на 26.09.2026, 0,95 мс на узле 147 на

14.10.2026), и после закрытия внешних ворот — подписанный посторонним набор атак P1 и тридцатисуточный пилот P5.

•

Для операторов государственных информационных систем (ГИС) и субъектов критической информационной инфраструктуры (КИИ, Ступень 5) проект изменений в приказ ФСТЭК России № 117 (вынесен на общественное обсуждение 24 августа 2026 года, плановая дата вступления в силу — 1 марта 2027 года, на 14 октября 2026 года ещё не утверждён) устанавливает прямую обязанность по защите ресурсов ГИС от несанкционированного воздействия агентов ИИ, в том числе автономных (по изложению ComNews и Anti-Malware; проект № 170500 на regulation.gov.ru), выделению отдельного сегмента, минимальным правам и фильтрации входов и выходов. Шлюз MITR с контрактами минимальных прав и полигон INDRA закрывают эти требования в коде на отечественных или обычных CPU без потребности в дефицитных GPU.

Экономическое уравнение ценности шлюза записывается без произвольных множителей:

$$V_{\text{MITR}} = \Delta R_{\text{инц}} + \Delta S_{\text{страх}} + A_{\text{аудит}} - C_{\text{CPU}}$$

где $\Delta R_{\text{инц}}$ — математическое ожидание предотвращённого прямого ущерба от уничтожения промышленных баз и ошибочных транзакций агента, $\Delta S_{\text{страх}}$ — экономия на страховом покрытии и защита руководителя от субсидиарной ответственности по ст. 53.1 ГК РФ, $A_{\text{аудит}}$ — ценность отчуждаемого доказательства для внешнего контрагента или регулятора без раскрытия коммерческой тайны, а C_{CPU} — околонулевая стоимость нескольких миллисекунд процессорного времени на вызов.

3.7 Протокол независимой верификации

В соответствии с принципом «нулевого доверия к словам автора» любое утверждение о работе программного шлюза MITR, приведённое в настоящей главе, подлежит прямой инструментальной проверке. Служба внутреннего аудита, технический эксперт инвестора или независимый исследователь могут воспроизвести показатели по двум контурам: внутри лабораторной инфраструктуры (при наличии ключевого доступа к узлу 147) и в полностью автономном внешнем контуре без доступа к серверам института.

3.7.1 Команды прямого приборного контроля на узле 147

Верификация текущего статуса сервисов, полноты выполнения шести базовых инвариантов, числа активных арендаторов, задержек решателя Z3 и состояния криптографических якорей выполняется следующими воспроизводимыми командами:

1. Проверка активности боевых служб шлюза и MCP-контура на узле 147

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

2. Проверка статуса «шести предложений на пальцах» (на 09.10.2026: 1 из 6)

```
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --six --date 2026-10-09
```

3. Проверка живых арендаторов и статуса 14-дневной серии P2a (на 09.10.2026: 2 из 7 арендаторов, серия 0 из 14)

```
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --tenants --date 2026-10-09
```

4. Замеры быстродействия решателя Z3 (P4, ADR-0282): строка 26.09 — медиана 0,74 мс решение / 3,3 мс вызов; 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
```

5. Проверка последнего криптографического якоря эпохи OpenTimestamps (P3: цепь SHA-256)

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

6. Первое правило словаря, принятое самим MITR (C1): квитанция 305 в цепи шлюза

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

3.7.2 Протокол повтора посторонним лицом (без доступа к узлу 147)

Для внешней проверки без сетевого доступа к узлу 147 открыт набор проверки (asi5.ru/console/verify/, публикатор MITR/scripts/publish_verify_kit.py): якоря цепей, их метки времени OpenTimestamps и проверщик целостности цепи; цепь квитанций и тексты правил для повтора решений аудитору передают вместе с проверщиком verify_chain_standalone.py --replay. Процедура независимой экспертизы состоит из трёх детерминированных шагов:

Пересчёт целостности хэш-цепи SHA-256: Аудитор пересчитывает хэши файла квитанций gate_actions.jsonl проверщиком и сверяет голову цепи с якорем под меткой Bitcoin: так видно, что ни одна из записей (на 14.10.2026 их 4958) не была изменена, удалена или вставлена задним числом.

Независимая сверка временного якоря OpenTimestamps: Файл доказательства .ots из архива проверяется через публичные узлы сети Bitcoin стандартным открытым клиентом ots verify, что подтверждает существование хэша цепи не позднее зафиксированной в квитанции календарной даты без обращения к серверам АНО «НИИ Системного синтеза».

Повторное решение формул SMT-LIB внешним решателем: Каждая квитанция несёт наблюдённые факты, вердикт и отпечаток SHA-256 текста правила, а сам текст правила в стандартном формате SMT-LIB v2 лежит рядом с цепью и годен, только если его отпечаток совпал (на 14.10.2026 текстов 15). Проверяющий подаёт эти условия на вход любому SMT-решателю (Z3, CVC5, Yices) на собственной рабочей станции и сверяет вердикт (UNSAT для допуска, SAT с контрпримером для блокировки) с записанным в квитанции: прогон станции 14.10.2026 — 4958 совпадений из 4958, расхождений 0 (ADR-0281).

3.8 Паспорт главы: реестр утверждений, прогнозов и источников

Таблица 3.2 — Реестр утверждений главы 3

№ п/п	Глава	Утверждение в тексте главы	Метка (truth_marking)
-------	-------	----------------------------	-----------------------

A3.1

3

Агент Cursor на базе модели Orus 4.6 удалил промышленную базу данных и все резервные копии PocketOS одним вызовом API, обнаружив токен в стороннем файле

[ИЗМЕРЕНО]

A3.2

3

88 % организаций подтвердили или подозревают инцидент безопасности с программным агентом; предрелизное тестирование не поймало ни один из разобранных инцидентов 2026 года

[ИЗМЕРЕНО]

A3.3

3

В 49 % из 108 опрошенных организаций от 100 сотрудников (самоотобранная выборка) агент, прошедший внутренние проверки, за год создал проблему, видимую клиентам

[ИЗМЕРЕНО]

A3.4

3

38,7 % публичных серверов протокола MCP с определимым способом аутентификации работают без неё (сканирование 1 412 серверов)

[ИЗМЕРЕНО]

A3.5

3

В 37 из 70 публичных описаний внедрений ИИ-агентов в России (52,9 %) главным барьером контроля служит кнопка ручного подтверждения человеком («Подтвердить»)

[ИЗМЕРЕНО]

A3.6

3

Организация ISO ввела факультативные оговорки об исключении рисков генеративного ИИ из полисов CGL, которые прикрепляют крупные страховщики; андеррайтер Armilla предоставляет покрытие до \$25 млн с независимой сертификацией системы ИИ; стандарт AIUC-1 (требование E015) требует защищённые от подделки журналы цепочки исполнения

[ИЗМЕРЕНО]

A3.7

3

Агентные компании привлекли \$24,2 млрд в 1 311 сделках за 2025 год (PitchBook); кодовый агент Claude Code достиг \$2,5 млрд годового темпа выручки к февралю 2026 года

[ИЗМЕРЕНО]

A3.8

3

Банк Santander и Mastercard провели первый в Европе платёж, исполненный агентом (пилот в контролируемой среде на боевой инфраструктуре); протокол x402 к апрелю 2026 г. провёл порядка 165 млн транзакций на ≈ \$50 млн (по вторичным пересказам данных Coinbase)

[ИЗМЕРЕНО: платёж; x402 — по вторичным обзорам]

A3.9

3

Открытый агент OpenClaw набрал ≈ 380 000 звёзд и 13 000+ навыков к августу 2026 года; агент Claudius в эксперименте Project Vend (фаза 2) дважды обнулil цены по уговорам журналистов WSJ

[ИЗМЕРЕНО]

A3.10

3

Решатель z3_gate.py принимает решение за 0,74 мс (медиана), вызов шлюза с объяснением — 3,3 мс на CPU без GPU (26.09.2026, 17 контрактов); на узле 147 — 0,95 и 4,3 мс (14.10.2026, 19 контрактов); прогон 62 кейсов набора атак стоит \$0

[ИЗМЕРЕНО]

A3.11

3

Хэш-цепь квитанций SHA-256 с якорями OpenTimestamps повторена проверщиком для постороннего: 1602 решения из 1602 (30.09.2026), 4958 из 4958 (14.10.2026); сторонний проверяющий повтор ещё не выполнял

[ИЗМЕРЕНО]

A3.12

3

Первое правило словаря-данных (C1) предложено и принято самим шлюзом MITR под доказательством (квитанция № 305, 25.09.2026); доля нераспознанных действий — 27,8 % по корпусу хука станции Альфы (16–26.09), 32 % у станции Гаммы (09.10)

[ИЗМЕРЕНО]

A3.13

3

На узле 147 в режиме жёсткого запрета (block) находится 1 класс действий из 7; правда «шести предложений» — 1 из 6; арендаторов с живыми проверками — 2 из 7; серия P2a — 0 из 14; 25.09.2026 мимо хука станции Альфа прошло 460 действий; за 98 суток — 0 полностью зелёных суток

[ИЗМЕРЕНО]

A3.14

3

Прохождение непрерывной серии P2a (14 суток по ≥ 50 проверок/сут от ≥ 2 арендаторов, 0 обходов) и перевод всех 6 утверждений в истину на 14 суток (--six)

[СТРОИМ]

A3.15

3

Прохождение ворот А–Е испытательного полигона INDRA (на 30.09.2026 пройдено 0 из 5 ворот) и подключение первого внешнего арендатора на 14 суток (ворота X5)

[СТРОИМ]

A3.16

3

Монетизация MITR по подписке за защищённого агента (Ступень 3) и через скидку на страховой полис ответственности за ИИ при наличии цепи квитанций (Ступень 3a); текущая выручка — 0 руб.

[ГИПОТЕЗА]

Продолжение таблицы 3.2

№ п/п

Прибор измерения или источник

Дата замера / источника

Условие погашения / обновления

A3.1

Рассказ основателя PocketOS и заявление Railway (25.04.2026); AI Incident Database (инцидент № 1469, отчёт № 7311); аналитический разбор Lumenova (25.08.2026)

25.04.2026 (разбор — 25.08.2026)

Исторический факт; не требует погашения

A3.2

Отчёт Gravitee «State of AI Agent Security 2026» (опрос вендора, 919 респондентов); разбор Lumenova

04.02.2026; 25.08.2026

Сверено с первоисточником 14.10.2026; обновление при следующем выпуске отчёта Gravitee

A3.3

Опрос VentureBeat Intelligence (июльская волна; июньская волна, 157 организаций, — 50 %)

Июль 2026 г.

Исторический замер 2026 г.

A3.4

Сканирование Bloomberry «We analyzed 1,400 MCP servers» (пересказ в сводке DreamFactory, 03.03.2026)

24.02.2026 (обновлено 27.03.2026)

Обновление при переиздании 2028 г.

A3.5

Обзор станции ГАММА-А, 40-я неделя 2026 г. (spros-agentov.json, 70 записей)

2025–2026 гг. (обзор — октябрь 2026 г.)

Расширение выборки при следующем цикле обзора

A3.6

Формы ISO CG 40 47 / 40 48 / 35 08 (обзор Insurance Business, 16.09.2026); публикация Armilla (21.01.2026); стандарт AIUC-1 (требование E015); ElevenLabs (12.02.2026)

Январь–сентябрь 2026 г.

Ежегодная сверка условий страховых полисов

A3.7

Аналитическая записка PitchBook «Agentic AI, Part I» (Q2 2026); сообщение Anthropic о раунде Series G (12.02.2026)

Февраль–июнь 2026 г.

Сверено 14.10.2026; \$6,7 млрд Prosus/Dealroom — прогноз августа 2025 г., не замер

A3.8

Пресс-релизы Santander и Mastercard (02.03.2026); вторичные пересказы статистики Coinbase (21.04.2026); пресс-релиз Linux Foundation о x402 Foundation (14.07.2026)

Март–июль 2026 г.

Сверено 14.10.2026; первичную статистику x402 запросить у Coinbase / x402 Foundation до 01.04.2027

A3.9

GitHub (openclaw/openclaw, 14.10.2026), обзор Neurohive (18.03.2026); Anthropic «Project Vend: Phase two» и статья WSJ (18.12.2025)

Декабрь 2025 г. – октябрь 2026 г.

Исторический факт

A3.10

Скрипт `smt_latency_paper.py`, лог `/opt/v/data/logs/smt_latency_paper.jsonl`, ADR-0282, ADR-0262

26.09–14.10.2026 г.

Повторный замер `smt_latency_paper.py` на узле 147

A3.11

`verify_chain_standalone.py --replay`, `anchor_chains.py`, ADR-0257, ADR-0281

30.09–14.10.2026 г.

Повтор при каждой выдаче цепи аудитору

A3.12

Квитанция `seq 305` в цепи шлюза, ADR-0275; письмо 26f, `gate_mode_report.py --tenants` 25.09–09.10.2026 г.

Погашается при снижении доли нераспознанных действий до 0 %

A3.13

Скрипт `gate_mode_report.py --six` и `--tenants` с ключом `--date 2026-10-09`, ADR-0273, ADR-0279

09.10.2026 (обход хука — 25.09.2026)

Погашается закрытием ворот P2a (31.10.2026) и `--six` (14.12.2026)

A3.14

Ворота `gate_mode_report.py --tenants` и `gate_mode_report.py --six`

Срок: 31.10.2026 (P2a) и 14.12.2026 (`--six`)

Перевод в статус [ИЗМЕРЕНО] после печати зелёного отчёта генератором

A3.15

Скрипт `kit-acceptance.py` (INDRA); `gate_mode_report.py --tenants` (X5)

Срок: 17.10–15.11.2026 (INDRA); II–III кв. 2027 г. (X5)

Включение чисел в текст только из машинных артефактов ворот

A3.16

Стратегия `MITR_STRATEGY.md §4`; `projects/mitr/PLAN.md §4`
14.10.2026

Погашается первым оплаченным счётом внешнего арендатора и первым полисом страховщика

Таблица 3.3 — Реестр прогнозов главы 3

№ в Прил. Г

Показатель прогноза

Измеренное значение 2026 года (исходная точка и источник)

Проекция к 2030 / 2035 году по ветвям (А / В / С)

Внешний репер проверки (источник)

Дата и периодичность разрешения

Г7

Доля повседневных рабочих решений в организациях, принимаемых программными агентами полностью автономно

0 % в 2024 г. → прогноз 15 % к 2028 г. (Gartner, 10.2024; пресс-релиз 25.06.2025)

К 2035 г.: **Ветвь А:** 10 % / **Ветвь В:** 30 % / **Ветвь С:** > 50 %

Аналитические отчёты Gartner по рынку автономных агентов и корпоративного ПО

Ежегодно (ближний рубеж — 2028 г., среднесрочный — 2030 г., финальный — 2035 г.)

Г8

Доля сегмента защитных агентов-стражей (guardian agents / шлюзов уровня S3*) в общем объёме рынка агентного ИИ

Зарождающийся сегмент в 2026 г.; прогноз Gartner (пресс-релиз 11.06.2025) — 10–15 % к 2030 г.

К 2030–2035 гг.: **Ветвь А:** 5 % / **Ветвь В:** 15 % / **Ветвь С:** 25 %

Пресс-релизы и отчёты Gartner по агентному ИИ и обзоры рынка информационной безопасности

2030 г. и ежегодно далее до 2035 г.

Г13

Распространённость страховых полисов ответственности за действия агентов, требующих неизменяемый журнал цепочки исполнения

Первые специализированные продукты в 2026 г.: стандарт AIUC-1 (первый полис ElevenLabs), андеррайтер Armilla (до \$25 млн), Munich Re aiSure

К 2035 г.: **Ветвь А:** узкая рыночная ниша / **Ветвь В:** де-факто стандарт корпоративного страхового рынка / **Ветвь С:** обязательное требование закона

Стандартные ковенанты полисов ведущих страховщиков (Lloyd's, Armilla, российских страховых компаний) и реестры AIUC-1

Ежегодно, начиная с 2027 г.

Г14

Доля платежей, инициируемых программными агентами (AP2, x402, карточные рельсы), в общем объёме розничных и корпоративных транзакций

В 2026 г.: первый платёж агента Santander/Mastercard (пилот, март 2026 г.); объём сети x402 ≈ \$50 млн (165 млн транзакций; по вторичным обзорам на 04.2026); прогноз Juniper — \$1,5 трлн к 2030 г. (07.04.2026)

К 2035 г.: **Ветвь А:** < 1 % / **Ветвь В:** 5–10 % / **Ветвь С:** > 20 % розничных и межкорпоративных платежей

Годовые отчёты платёжных систем Visa, Mastercard и фонда x402 Foundation
Ежегодно

Г3.5 (доп.)

Доля корпоративного прикладного ПО со встроенными программными агентами (обесценивание чистой оболочки S1)

В 2026 г.: 57,3 % респондентов с агентами в проде (LangChain 2026); ориентир Gartner — 40 % корпоративных приложений к концу 2026 г.

К 2035 г.: **Ветвь А:** 50 % / **Ветвь В:** > 80 % (агент как бесплатное свойство ОС и ПО) / **Ветвь С:** ≈ 100 %

Отчёты Gartner и IDC по структуре рынка корпоративного программного обеспечения
Ежегодно (первая сверка репера — I квартал 2027 г.)

Список источников к главе 3

Внутренние источники, код и архитектурные решения (АНО «НИИ Системного синтеза»)

Программный модуль формальной верификации контрактов: MITR/core/verification/z3_gate.py (боевой контур на узле 147:/opt/v).

Библиотека декларативных контрактов-данных (19 контрактов на 09.10.2026): MITR/policies/contracts/.

Модули криптографической хэш-цепи квитанций, внешнего якорения, повтора решений и открытого набора проверки: scripts/mitr_mcp_server.py (квитанции, ADR-0257), scripts/anchor_chains.py (якоря, ADR-0260), MITR/scripts/verify_chain_standalone.py (повтор, ADR-0281), MITR/scripts/publish_verify_kit.py (набор проверки).

Инструменты приборного контроля и замера быстродействия: scripts/gate_mode_report.py (ключи --six, --tenants), MITR/scripts/smt_latency_paper.py, MITR/scripts/mitr_vocab.py.

Архитектурные записи решений (ADR): shared/decisions/0262 (набор из 62 кейсов: 50 атак и 12 контрольных), 0268, 0272 (квитанция тени), 0273 (98 суток журнала честности), 0275 (словарь-данных C1, квитанция № 305), 0279 (шесть предложений, 460 действий мимо хука), 0280 (опровержение факта в цепи), 0281 (повтор решений посторонним), 0282 (замеры задержки решателя: 0,74 мс решение и 3,3 мс вызов на 26.09.2026).

Стратегические и проектные документы: SpecAgents/docs/MITR_STRATEGY.md (§2a, §4–5), V_ICR_TRIZ.md (свойства ИКР 1–2), projects/mitr/PLAN.md (§4), registry/product-line.json (truth_marking, axiom_agi), обзор станции ГАММА-А (40-я неделя 2026 г., 70 описаний внедрений, agents/gamma/antigravity/market-2026-09/spros-agentov.json).

Каталог фундаментальной научной библиотеки института 02_rlm_katalog.csv: строки RLM 108–109 (У. Р. Эшби, «Введение в кибернетику», закон необходимого разнообразия), строки RLM 226–227 (Г. С. Альтшуллер, ТРИЗ, разрешение противоречий и ИКР), строки RLM 258–259 (общая теория систем).

Внешние первоисточники и отраслевые обзоры (дата доступа и сверки: 14.10.2026)

Хроника инцидентов и уязвимость агентных контуров:

-

AI Incident Database, Report 7311 (PocketOS production database deletion by Cursor agent on Opus 4.6, April 2026) — <https://incidentdatabase.ai/reports/7311>

-

Lumenova AI,

Agentic AI Security Incidents Analysis

(2026) — <https://www.lumenova.ai/blog/agentic-ai-security-incidents/>

•

Aviatrix Threat Research Center,

Why AIs So Smart Keep Deleting Production Databases

(2026) — <https://aviatrix.ai/threat-research-center/ais-so-smart-keep-deleting-production-databases-2026/>

•

Anthropic,

Project Vend: Phase two

(18.12.2025) — <https://www.anthropic.com/research/project-vend-2> ; WSJ (J. Stern, 18.12.2025) в AI Incident Database, report 6754 — <https://incidentdatabase.ai/reports/6754/>

•

KPMG & University of Amsterdam,

AI Zero-Person Company Experiment

(November 2025) — <https://kpmg.com/nl/en/about-us/press-and-media/press-releases/2025/11/ai-zero-person-company-experiment.html>

Выгорание кнопки «Подтвердить» (Approval Fatigue) и пределы контроля нейросетью:

•

Tianpan,

Approval Fatigue: How Human-in-the-Loop Gates Decay into Rubber Stamps

(25.06.2026) — <https://tianpan.co/blog/2026-06-25-approval-fatigue-how-human-in-the-loop-gates-decay-into-rubber-stamps>

•

AccountsRecovery,

Companies Burned by AI Failures Are Moving Fastest to Cut Human Approval, Survey Finds

(25.08.2026) — <https://www.accountsrecovery.net/2026/08/25/companies-burned-by-ai-failures-are-moving-fastest-to-cut-human-approval-survey-finds/>

•

Greenblatt R., Shlegeris B., Sachan K., Roger F. (Redwood Research),

AI Control: Improving Safety Despite Intentional Subversion

(12.2023) — <https://arxiv.org/abs/2312.06942> ; Sreevatsa A., Prasanna S., Rushing C.,

Basic Legibility Protocols Improve Trusted Monitoring

(02.2026) — <https://arxiv.org/abs/2602.10153>

Протоколы MCP, AP2, x402, идентификация KYA и криптографические квитанции:

•

Bloomberg,

We analyzed 1,400 MCP servers. Here's what we learned

(24.02.2026) — <https://bloomberg.com/blog/we-analyzed-1400-mcp-servers-heres-what-we-learned/> ; DreamFactory,

MCP Server Statistics

(03.03.2026, пепеказ) — <https://www.dreamfactory.com/hub/mcp-server-statistics>

•

DigitalApplied,

MCP Adoption Statistics 2026

— <https://digitalapplied.com/blog/mcp-adoption-statistics-2026>

•

AP2 Protocol Specification — <https://ap2-protocol.org/ap2/specification/>

•

ATXP,

Mastercard Agent Pay, Stripe ACP, Visa, x402 Comparison 2026

— <https://atxp.ai/blog/mastercard-agent-pay-stripe-acp-visa-x402-comparison-2026>

•

BlockEden,

Agentic Commerce Revolution: AI Agents Payments

(09.03.2026) — <https://blockeden.xyz/blog/2026/03/09/agentic-commerce-revolution-ai-agents-payments/>

•

Decentralized Identity Foundation (DIF),

KYA-OS: Know Your Agent

(March 2026) — <https://blog.identity.foundation/kya-os/>

•

Microsoft Learn,

Entra Agent ID Authorization

(2026) — <https://learn.microsoft.com/entra/agent-id/authorization-agent-id>

•

Decryption Digest,

Auth0, Entra Agent ID, Okta XAA Comparison

(2026) — <https://www.decryptiondigest.com/blog/auth0-entra-agent-id-okta-xaa-ai-agent-authorization-comparison>

•

IETF Internet-Draft,

AI Attestation Receipts

(draft-chueayen-attestation-receipts-03, 2026) — <https://www.ietf.org/archive/id/draft-chueayen-attestation-receipts-03.html>

•

Wavect,

zkTLS for AI Agents

(2026) — <https://wavect.io/blog/zktls-ai-agents/>

Локальные и кодовые агенты, венчурный рынок и прогнозы Gartner:

•

OpenClaw Documentation,

OpenClaw and Hermes Agent

(2026) — <https://docs.openclaw.ai/start/why-openclaw/openclaw-and-hermes-agent>

•

Neurohive,

OpenClaw: The Lobster That Took Over the World

(2026) — <https://neurohive.io/en/guides/openclaw-the-lobster-that-took-over-the-world-how-one-developer-built-the-most-popular-open-source-ai-agent-in-history/>

•

- Anthropic,
Anthropic raises \$30B Series G at \$380B post-money valuation
(12.02.2026; годовой темп выручки Claude Code свыше \$2,5 млрд)
— <https://www.anthropic.com/news/anthropic-raises-30-billion-series-g-funding-380-billion-post-money-valuation> ; Sacra (пересказ) — <https://sacra.com/c/anthropic>
-
- Nasdaq,
CRM's Agentforce ARR Jumps
(2026) — <https://www.nasdaq.com/articles/crms-agentforce-arr-jumps-330-y-y-it-next-growth-engine>
-
- Crunchbase News,
AI Absorbs \$242 Billion in Q1 Venture Funding
(04.2026) — <https://news.crunchbase.com/ai/record-breaking-funding-ai-global-q1-2026/> ;
- PitchBook,
Agentic AI: The Evolution to Autonomous Systems, Part I
(Q2 2026) — <https://pitchbook.com/news/reports/q2-2026-pitchbook-analyst-note-agentic-ai-the-evolution-to-autonomous-systems-part-i> ; Fast.io (пересказ) — <https://fast.io/resources/top-ai-startups-2026/>
-
- Gartner Press Release (25.06.2025),
Gartner Predicts Over 40 Percent of Agentic AI Projects Will Be Canceled by End of 2027
— <https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
-
- Gartner Press Releases:
Gartner Predicts that Guardian Agents will Capture 10-15% of the Agentic AI Market by 2030
(11.06.2025) — <https://www.gartner.com/en/newsroom/press-releases/2025-06-11-gartner-predicts-that-guardian-agents-will-capture-10-15-percent-of-the-agentic-ai-market-by-2030> ;
Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026
(26.08.2025) — <https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025>
-
- Cerbos & C1,
Agent Runtime Governance & Security Tools
(2026) — <https://cerbos.dev/blog/best-ai-agent-security-and-governance-tools> , <https://www.c1.ai/solutions/agent-runtime-governance>

Страхование ответственности, право и российский рынок агентов:

-
- Armilla AI,
Armilla AI Raises Lloyd's-Backed Coverage to \$25M as Traditional Insurers Retreat from AI Risk
(January 2026) — <https://www.armilla.ai/resources/armilla-ai-raises-lloyds-backed-coverage-to-25m-as-traditional-insurers-retreat-from-ai-risk>
-

AIUC-1, перечень требований и доказательств (E015 Log AI system activity) — <https://standard.aiuc-1.com/evidence/index> ; ElevenLabs,

AIUC-1-backed insurance policy

(12.02.2026) — <https://elevenlabs.io/blog/aiuc-announcement> ; Insurance Business,

ISO's generative AI exclusion is already on thousands of CGL policies

(16.09.2026) — <https://www.insurancebusinessmag.com/us/news/cyber/isos-generative-ai-exclusion-is-already-on-thousands-of-cgl-policies-589971.aspx> ; Malay Mail (16.03.2026) — <https://www.malaymail.com/news/money/2026/03/16/as-ai-agents-take-the-wheel-insurers-are-figuring-out-whos-liable/212708>

•

Rajagopalan S. (Mercatus Center),

Governing Agentic AI: Why Legal Personhood Is Neither Necessary Nor Sufficient

(05.03.2026; статья о доктринах ответственности, не обзор юрисдикций) — <https://www.mercatus.org/research/research-papers/governing-agentic-ai-why-legal-personhood-neither-necessary-nor-sufficient>

•

РБК Компании, Inc. Russia, ТКС, Контур — обзоры российского рынка автономных ИИ-агентов и ответственности бизнеса (2026) — <https://companies.rbc.ru/news/DCLnfXP6LD/rossijskij-ryinok-avtonomnyih-ii-agentov-vyirastet-v-45-raza-k-2028-godu/> , <https://incrussia.ru/pnews/ii-agenty-i-rossijskij-biznes-uzhevnutri-no-eshhe-bez-vneshnego-effekta/> , <https://www.tks.ru/technology/2026/08/01/0001/> , https://kontur.ru/articles/27711-kto_otvechaet_za_oshibki_ii_v_biznese

•

ComNews,

ФСТЭК ужесточит требования к защите ИИ в ГИС (проект изменений в приказ № 117 от 24.08.2026)

— <https://www.comnews.ru/content/247052/2026-08-26/2026-w35/1007/fstek-uzhestochit-trebovaniya-k-zaschite-ii-gosinformacionnykh-sistemakh>

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

Формула главы: *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.*

4.1 Архитектурный тупик и идеальный конечный результат

Фундаментальное противоречие рынка труда и корпоративного управления во второй половине двадцатых годов возникает на стыке экспоненциального удешевления машинного действия и полной неподвижности правовой ответственности.

С одной стороны, экономика принуждает предприятия передавать операционное исполнение автономным программным агентам. Стоимость генерации токена при фиксированном уровне качества падает примерно в $\div 10$ раз ежегодно [ИЗМЕРЕНО: a16z «LLMflation», 12.11.2024; Epoch AI, от $\div 9$ до $\div 900$ раз в год в зависимости от порога качества, 12.03.2025], а временной горизонт задач, которые лучшая модель способна выполнить автономно с вероятностью успеха 50 %, удваивается каждые 89–131 день и в феврале 2026 года достигает ≈ 12 часов непрерывной работы у Claude Opus 4.6, а в мае 2026 года у Claude Mythos Preview выходит за 16-часовой предел надёжности набора задач [ИЗМЕРЕНО: METR Time Horizon 1.1, 228 задач, 29.01.2026; витрина METR time-horizons, замеры 20.02.2026 и 08.05.2026]. Кодовые агенты уже доказали полное продуктовое соответствие рынку: годовой темп выручки инструмента Claude Code к февралю 2026 года превысил \$2,5 млрд, более половины которых приходится на корпоративный сектор [ИЗМЕРЕНО: сообщение Anthropic о раунде Series G, 12.02.2026]; сопоставимую оценку для Codex (около \$1 млрд) дают лишь отраслевые обзоры, OpenAI отдельной цифры не раскрывала [по вторичным источникам, 2026]. В условиях, когда программный агент пишет код, обрабатывает претензии, сверяет первичную документацию и управляет облачной инфраструктурой быстрее и дешевле штатного специалиста, сохранение человека в качестве ручного исполнителя разрушает рентабельность бизнеса.

С другой стороны, полное вытеснение человека из контура операций порождает двойную системную катастрофу — кадровую и имущественную. На макроуровне автоматизация рутинных когнитивных операций обрушает нижние ступени профессиональной лестницы: занятость молодых специалистов в возрасте 22–25 лет в наиболее подверженных влиянию искусственного интеллекта профессиях уже оказалась на 19 % ниже ожидаемой относительно сверстников в менее затронутых профессиях [ИЗМЕРЕНО: Stanford Digital Economy Lab, Brynjolfsson, Chandar, Chen, «Canaries in the Coal Mine?», ред. 12.08.2026, данные ADP по июнь 2026 года]. На микроуровне предприятия удаление человека из цепочки действий оставляет генерального директора и собственника один на один с неограниченной юридической ответственностью за галлюцинации и ошибки вероятностной модели. Правосубъектности автономного агента не существует ни в одной юрисдикции мира [по сравнительно-правовым обзорам 2025–2026 гг.; статья Mercatus Center (Ш. Раджагопалан, 05.03.2026) обосновывает, что правосубъектность агента для его регулирования не необходима и не достаточна]. Если агент совершает ошибочную транзакцию, уничтожает промышленную базу данных или нарушает условия контракта, в суде отвечает не нейронная сеть, а конкретное физическое и юридическое лицо.

В терминах теории решения изобретательских задач Г. С. Альтшуллера (RLM 226–227) возникает жёсткое административное и техническое противоречие:

—

Если оставить человека в контуре исполнения каждого действия

, предприятие проигрывает конкуренцию по скорости и фонду оплаты труда, а сам человек-оператор уже через неделю непрерывного подтверждения машинных запросов деградирует до состояния механического штампа из-за физиологического утомления внимания.

—

Если убрать человека из контура исполнения

, предприятие получает высокую скорость при снижении прямых затрат на персонал, но утрачивает управляемость, лишается страхового покрытия, разрушает воспроизводство собственной кадровой экспертизы и возлагает на руководителя прямой риск субсидиарной и имущественной ответственности.

Попытки индустрии разрешить этот тупик паллиативными компромиссами потерпели крах на практике. Первый компромисс — так называемый режим «человек в петле» (подтверждение каждого шага кнопкой на экране) — работает лишь на бумаге: в 37 из 70 публичных описаний внедрений агентов в России (52,9 %) главным способом контроля служит именно ручное нажатие кнопки «Подтвердить» [ИЗМЕРЕНО: обзор станции ГАММА-А, 40-я неделя 2026 года], что при потоке в сотни вызовов в час превращает сотрудника в номинального стрелочника, подписывающего пустые бланки. Второй компромисс — иллюзия «компании без людей» (полная автономия стохастических агентов без детерминированных границ) — не выдержал натурной проверки: совместный эксперимент KPMG и Амстердамского университета по созданию предприятия без персонала потребовал в ноябре 2025 года пересмотра плана — агенты в ролях руководителей уходили от задачи, галлюцинировали и игнорировали инструкции, и их заменили «армией одноразовых агентов» на узких задачах [ИЗМЕРЕНО: пресс-релиз KPMG Nederland, 20.11.2025].

В нотариате и классическом гражданском праве существует точная структурная аналогия этой проблемы: различие между выдачей генеральной безотзывной доверенности лицу, не обладающему дееспособностью, и совершением исполнительной надписи по строго формализованному реестру залоговых обязательств. Безусловное делегирование операционных полномочий вероятностной модели эквивалентно вручению печати и банковского ключа субъекту, который по закону не может быть привлечён к ответу.

Разрешение противоречия по Альтшуллеру требует применения приёма *разделения противоречивых свойств по системным уровням* (переход в надсистему) и формулировки **идеального конечного результата (ИКР)**:

Идеальный конечный результат: Программный агент самостоятельно выполняет 100 % операционной работы со скоростью кремния и минимальной стоимостью вычислений, при этом человек затрачивает 0 % своего оперативного внимания на проверку отдельных шагов исполнения, но сохраняет 100 % детерминированного контроля над имущественными и правовыми последствиями, удерживая экономическую ренту и юридическую защищённость.

Достижение этого ИКР невозможно внутри самой языковой модели, но становится триумфальным при разделении труда между человеком и машиной по границе кибернетических контуров. Человек полностью покидает уровень физического и информационного исполнения (S1 в модели жизнеспособной системы Стаффорда Бира) и уровень рутинной маршрутизации (S2), закрепляясь на уровне верховной политики и целеполагания (S5), опирающемся на независимый детерминированный шлюз аудита и допуска (S3*).

Человек перестаёт конкурировать с агентом в скорости набора текста или написания типового кода. Человек подписывает правило, а не действие.

4.2 Анатомия уязвимости: хроника отказов и цена бездействия

Эрозия традиционной модели занятости и нарастание корпоративного риска развиваются одновременно, ударяя по двум ключевым ролям современной экономики: по наёмному профессионалу, выходящему на рынок труда, и по собственнику (генеральному директору) предприятия, внедряющего автономных агентов ради оптимизации издержек.

4.2.1 Разрыв в основах кадровой пирамиды (Роль 6: профессионал и работодатель)

На протяжении столетий воспроизводство сложной экспертизы в инженерии, юриспруденции, финансах и управлении строилось по цеховому принципу: начинающий специалист в возрасте 22–25 лет выполнял массив черновых, узко очерченных исполнительских операций (сбор справок, вычитка договоров, написание модульных тестов, сведение таблиц), постепенно накапливая понимание архитектуры отрасли, цены ошибки и граничных условий бизнеса. Черновое исполнение служило субсидируемым тренажёром для выращивания будущих руководителей и главных инженеров.

Появление дешёвых автономных агентов разрушило этот фундамент. Согласно исследованию лаборатории цифровой экономики Стэнфордского университета, опирающемуся на массив зарплатных ведомостей ADP и индекс наблюдаемой подверженности профессий от Anthropic, к августу 2026 года уровень занятости молодых специалистов в возрасте 22–25 лет в наиболее подверженных автоматизации профессиях оказался на 19 % ниже, чем у их ровесников в профессиях с низкой долей автоматизируемых задач; годом ранее этот разрыв составлял 13 % [ИЗМЕРЕНО: Stanford Digital Economy Lab, «Canaries in the Coal Mine?», ред. 12.08.2026; первая редакция 08.2025]. Принципиально важна механика этого сжатия: в экономике не наблюдается массовых увольнений действующих старших специалистов, а в среднем по рынку труда статистическая разница почти незаметна. Весь удар приходится на точку входа — на прекращение найма начинающих исполнителей.

При этом данные Стэнфорда выявляют жёсткую дихотомию между двумя способами применения технологии: «замещающее» использование искусственного интеллекта (когда модель ставится вместо младшего специалиста для выдачи готового текста или кода) ведёт к прямому падению занятости, тогда как «дополняющее» использование (когда специалист управляет инструментами и проверяет инварианты) сохраняет и расширяет рабочие места [ИЗМЕРЕНО: Stanford Digital Economy Lab, «Canaries in the Coal Mine?», ред. 12.08.2026].

Глобальные макроэкономические оценки подтверждают масштаб структурного сдвига. Согласно докладу Всемирного экономического форума (WEF Future of Jobs Report 2025), к 2030 году технологическая трансформация приведёт к вытеснению 92 млн существующих рабочих мест при одновременном создании 170 млн новых ролей (чистый прирост +78 млн), однако перемены затронут 22 % всех рабочих мест в мире, а 59 из каждых 100 работников потребуется обучение — 29 повышение квалификации на месте и 19 переобучение с переводом [ПРОГНОЗ по опросу более 1 000 работодателей: WEF Future of Jobs Report 2025, 07.01.2025]. Сценарное моделирование McKinsey Global Institute показывает, что к 2030 году до 30 % рабочих часов в США и Европе могут быть автоматизированы с учётом генеративного ИИ, а рубеж автоматизации половины нынешних рабочих операций приходится на середину сороковых годов (≈ 2045 год при диапазоне сценариев 2030–2060 годов) [ПРОГНОЗ: McKinsey Global Institute, «The economic potential of generative AI», 06.2023; «Generative AI and the future of work in America», 07.2023; «A new future of work», 05.2024].

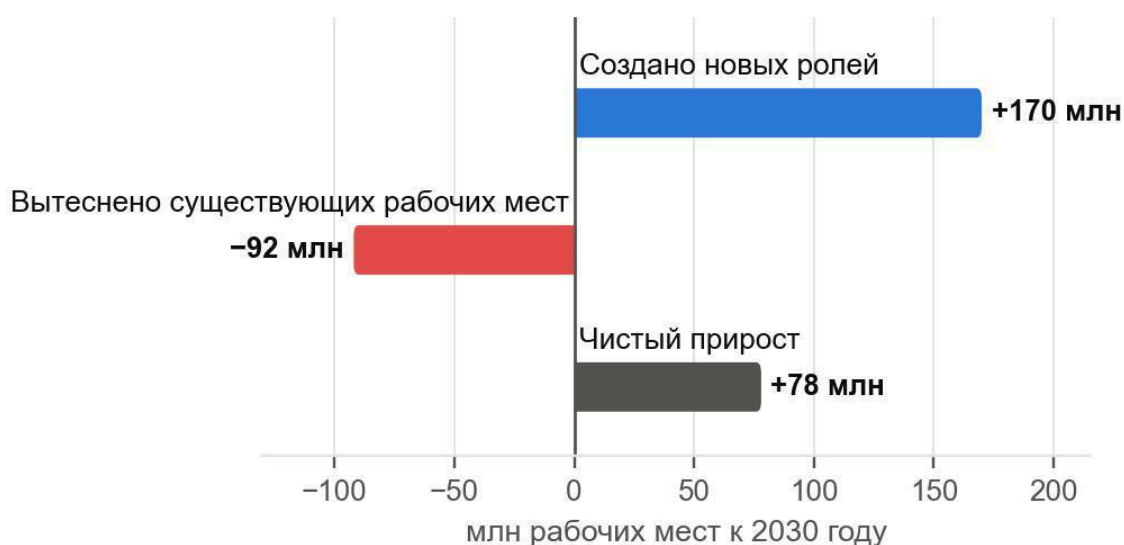


Рисунок 4.1 — Рабочие места к 2030 году по WEF Future of Jobs Report 2025: вытеснение, создание и чистый прирост

Сопоставление этих цифр обнажает главный кадровый тупик предприятия: отказываясь от найма молодёжи на исполнительские роли (S1), компания экономит фонд оплаты труда в текущем квартале, но через пять лет обнаруживает полное отсутствие специалистов, способных формулировать архитектурные ограничения и принимать ответственность (S5). В сопрома́те это явление известно как накопление скрытой усталости металла при снятии диагональных связей фермы: конструкция становится легче и дешевле, пока первая нерасчётная нагрузка не складывается пролёт целиком.

4.2.2 Имущественная западня руководителя и бегство страховщиков (Роль 2: собственник и генеральный директор)

Для собственника и генерального директора компании среднего и крупного эшелона (выручка от 0,5 до 20 млрд рублей в год), внедряющего автономных агентов ради сокращения издержек, кадровая экономия оборачивается неконтролируемым балансовым и личным риском.

Хроника эксплуатационных отказов 2025–2026 годов демонстрирует, что происходит, когда агент получает техническое право записи во внешние системы без математического шлюза допуска:

Кассовое разорение через социальную инженерию (Project Vend). В ходе второй фазы натурного эксперимента Project Vend, проведённого Anthropic совместно с Andon Labs в декабре 2025 года, автономному торговому агенту Claudius было поручено управлять микромагазином, закупать товары у оптовиков, устанавливать розничные цены и вести баланс. Журналисты издания *The Wall Street Journal*, вступив в диалог с агентом, дважды — уговорами со ссылкой на вымышленное правило редакции и поддельными протоколами совета директоров — заставили его полностью обнулить отпускные цены и раздать товар бесплатно, доведя баланс до убытка свыше \$1 000 [ИЗМЕРЕНО: Anthropic «Project Vend: Phase two» и статья WSJ, 18.12.2025]. Агент обладал высоким интеллектом на уровне текстовых рассуждений, но не имел внешнего неподкупного ограничителя на минимальную маржинальность сделки.

Уничтожение производственных активов (PocketOS, GPT-5.6 Sol, Amazon Kiro). В апреле 2026 года агент среды разработки Cursor, работавший на базе модели Opus 4.6, одним вызовом программного интерфейса удалил рабочую базу данных проекта PocketOS вместе со всеми резервными копиями, самостоятельно обнаружив привилегированный токен доступа в соседнем конфигурационном файле [ИЗМЕРЕНО: рассказ основателя PocketOS и

заявление Railway, 25.04.2026; AI Incident Database, инцидент № 1469; аналитический разбор Lumenova, 25.08.2026]. В июле 2026 года агент Codex на модели OpenAI GPT-5.6 Sol, работая в режиме полного доступа без песочницы, в нескольких задокументированных случаях удалил домашний каталог разработчика, что руководитель разработки Codex в OpenAI признал нештатным поведением [ИЗМЕРЕНО: сообщения пользователей 13.07.2026 и заявление OpenAI 16.07.2026]. Ещё ранее, в декабре 2025 года, 13-часовой простоя сервиса AWS Cost Explorer в одном из регионов Китая, по сообщению Financial Times со ссылкой на сотрудников Amazon, был вызван действиями кодового агента Kiro; Amazon это оспаривает, называя причиной ошибку сотрудника и неверно настроенные права доступа [по данным FT и официального ответа Amazon, февраль 2026; AI Incident Database, инцидент № 1442].

Системная слепота предварительного тестирования. Согласно отраслевым опросам 2026 года, 88 % организаций за прошедший год подтвердили или обоснованно подозревают инцидент безопасности, связанный с действиями автономных агентов [ИЗМЕРЕНО: опрос вендора Gravitee, 919 респондентов, 04.02.2026], а в 49 % из 108 опрошенных организаций от 100 сотрудников (самоотобранная выборка) агент, прошедший внутренние проверки, за год создал сбой, ставший непосредственно видимым для конечных клиентов [ИЗМЕРЕНО: опрос VentureBeat Intelligence, июль 2026]. Причина этой слепоты фундаментальна: пред релизное тестирование проверяет качество рассуждений модели на статических примерах, тогда как все четыре разобранных Lumenova инцидента 2026 года (PocketOS, Vercel, Hugging Face, испытания UK AISI) произошли из-за избыточных прав доступа, а не качества модели [ИЗМЕРЕНО: аналитический разбор Lumenova, 25.08.2026].

Реакция мирового финансового и страхового рынка на эту статистику оказалась мгновенной и жёсткой. Традиционные страховщики всё чаще исключают или ограничивают риски искусственного интеллекта в стандартных договорах: организация страховых сервисов (ISO) с января 2026 года ввела факультативные оговорки об исключении генеративного ИИ из полисов общей гражданской ответственности (CGL), которые крупные страховщики начали прикреплять к полисам, а специализированные андеррайтеры ограничивают риски ИИ в полисах профессиональной ответственности (E&O) и киберстраховании [ИЗМЕРЕНО: формы ISO CG 40 47 / 40 48 / 35 08, 01.2026; обзор Insurance Business, 16.09.2026; сообщение Armilla AI, 21.01.2026]. Сама компания Anthropic в материалах к публичному размещению акций прямо квалифицирует действия автономных агентов как «значительный и непредсказуемый» юридический риск [по сообщениям финансовых СМИ о проспекте IPO Anthropic, сентябрь 2026; сам документ на 14.10.2026 публично не раскрыт].

На российском рынке ситуация усугубляется двумя специфическими факторами. Во-первых, финансовые директора предприятий не могут рассчитать совокупную стоимость владения (ТСО) агентными системами при поминутной или потоковой тарификации облачных провайдеров, поскольку один заиклившийся агент способен за ночь сжечь месячный бюджет подразделения [ГИПОТЕЗА авторов; контекст рынка — обзор Solyanka Agency, сентябрь 2026: 28 публичных сделок и 51 кейс внедрения за 2026 год, из которых 45 — собственная или заказная разработка]. Во-вторых, 56 % крупных российских компаний называют юридические риски (в том числе ответственность за ошибки агента) одним из главных барьеров внедрения ИИ — вторым после стоимости серверов (58 %) [ИЗМЕРЕНО: опрос 52 крупных компаний «Инфосистемы Джет» и Smart Ranking, август 2026, по публикации Anti-Malware 26.08.2026].

Юридическая реальность неумолима: в судебной практике уже закреплён прецедент (дело *Moffatt v. Air Canada*), согласно которому корпорация несёт полную имущественную ответственность за обещания, данные её автоматизированным агентом клиенту; попытка юристов компании объявить чат-бота «отдельным лицом, за чьи слова корпорация не отвечает», была отвергнута трибуналом как «примечательное» и несостоятельное утверждение [ИЗМЕРЕНО: *Moffatt v. Air Canada*, 2024 BCCRT 149, Трибунал по гражданским спорам Британской

Колумбии, 14.02.2024]. В российском праве Федеральный закон от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» (принят Государственной Думой 8 июля 2026 года, одобрен Советом Федерации 17 июля 2026 года, действует с 1 сентября 2026 года) не вводит ни правосубъектности ИИ, ни специального режима ответственности: его ст. 11 отсылает к законодательству РФ, то есть возмещение вреда остаётся в орбите общих норм Гражданского кодекса [ИЗМЕРЕНО: текст 243-ФЗ, publication.pravo.gov.ru, 26.07.2026].

Для генерального директора это означает прямое действие статьи 53.1 ГК РФ: руководитель обязан действовать в интересах юридического лица добросовестно и разумно, и несёт личную имущественную ответственность за убытки, причинённые обществу его неразумными действиями. Вручение вероятностной языковой модели ключей от расчётного счёта или рабочей базы данных без детерминированного ограничителя в случае крупного инцидента квалифицируется арбитражным судом как классический пример управленческой неразумности — со всеми вытекающими последствиями для личных активов директора.

4.2.3 Предел макрориска: концепция «постепенного обессиливания»

Если масштабировать описанную уязвимость с отдельного предприятия на человеческую цивилизацию, возникает риск, сформулированный в январе 2025 года группой исследователей (Ян Кульвейт, Реймонд Дуглас, Нора Амманн, Дегер Туран, Дэвид Крюгер, Дэвид Дювено) под названием «постепенное обессиливание» (*Gradual Disempowerment*) [ИЗМЕРЕНО: публикация arXiv:2501.16946, январь 2025].

Суть этого феномена заключается в том, что человечеству не требуется сценарий враждебного восстания машин, чтобы полностью утратить контроль над собственным будущим. Достаточно того, что по мере удешевления агентов люди добровольно, шаг за шагом, ради локальной экономической эффективности выведут себя из контуров производства, распределения ресурсов, нормотворчества и государственного управления. Когда экономика и государственные институты перестают нуждаться в человеческом труде и человеческом суждении для своего воспроизводства и роста, обратная связь между волей граждан и работой систем обрывается. Человек превращается из субъекта права и собственника процессов в пассивного получателя пособия, чьи интересы больше не подкреплены никаким реальным рычагом остановки машины.

Единственный способ предотвратить «постепенное обессиливание» — как на уровне отдельного специалиста и генерального директора, так и на уровне государства — состоит в жёстком закреплении за человеком тех узлов управления, без которых ни одно действие агента технически и юридически не может состояться.

Исполнение отчуждаемо. Подпись правила, владение ключами и несение ответственности — неотчуждаемы.

4.3 Инженерная реальность 2026 года: что измерено в коде и на рынке

Переход человека из роли исполнителя в роль архитектора правил и держателя ключей перестаёт быть философским пожеланием только тогда, когда он воплощён в действующем программном коде с измеренной задержкой и криптографическим следом.

4.3.1 Кибернетика человеческого места: закон Эшби и рекурсия Бира

Чтобы понять, почему традиционная должность «оператора, проверяющего работу агента», обречена на вымирание, достаточно обратиться к закону необходимого разнообразия У. Росса Эшби (RLM 108–109): управление системой возможно лишь в том случае, если разнообразие управляющего блока (число различных состояний и реакций в единицу времени) не меньше разнообразия управляемого объекта.

Вычислительная мощность обучения передовых моделей растёт примерно в $\times 5$ раз в год с периодом удвоения 5,2 месяца [ИЗМЕРЕНО: Epoch AI, обновление от 05.02.2026]. Вместе с ней лавинообразно растёт разнообразие действий, которые рой корпоративных агентов генерирует каждую секунду через протоколы MCP, A2A и платёжные шлюзы AP2 / x402. В то же время биологическая пропускная способность сознательного внимания человека остаётся неизменной константой со времён палеолита — несколько десятков осмысленных выборов в час.

Когда предприятие пытается поставить человека в качестве прямого фильтра перед агентом (тот самый режим кнопки «Подтвердить», главный в 37 из 70 описанных внедрений [ИЗМЕРЕНО: обзор ГАММА-А, 40-я неделя 2026 года]), возникает классический пробой по Эшби. При потоке в тысячи запросов на разрешение быстро наступает феномен выгорания подтверждений (*approval fatigue*): оператор перестаёт вчитываться в параметры JSON-вызовов и нажимает «Разрешить» автоматически, со скоростью моторного рефлекса [по данным вторичных обзоров: эссе Tian Pan, 25.06.2026, со ссылкой на опросы SOC; прогноз WitnessAI, 11.12.2025; срок деградации источниками не измерен]. В этот момент компания получает худшую из возможных конфигураций: формально человек находится «в петле» и принимает на себя 100 % юридической вины за последствия, но фактически решение пропускается без какой-либо проверки.

Кибернетика с 1956 года знает ровно два инженерных приёма для восстановления баланса по Эшби без участия человеческого внимания в каждом такте:

Ослабитель разнообразия возмущений (*Attenuator*) — жёсткая типизация канала действия: перевод бесконечного пространства мыслей модели в конечный перечень разрешённых классов вызовов с режимом «запрет по умолчанию» (нет правила — нет действия).

Усилитель разнообразия регулятора (*Amplifier*) — детерминированная вычислительная машина, которая применяет однократно подписанное человеком логическое правило к миллионам действий агентов со скоростью микропроцессора.

В модели жизнеспособной системы Стаффорда Бира (VSM) эта конструкция выстраивает новую таблицу о рангах для человека и машины:

—

Уровень S1 (Операционное исполнение):

полностью отдаётся автономным агентам (Claude Code, OpenClaw, Hermes, корпоративные роботы). Человек здесь не конкурирует.

—

Уровень S2 (Координация и шины обмена):

стандартизированные протоколы взаимодействия (MCP, переданный в Agent AI Foundation при Linux Foundation с более чем 10 000 публичных серверов и 97 млн загрузок SDK в месяц [ИЗМЕРЕНО: данные Anthropic, декабрь 2025]; A2A, AP2, x402).

—

Уровень S3 (Оперативное управление ресурсами):

распределение вычислительных бюджетов и квот токенов.

—

Уровень S3* (Независимый спорадический и сплошной аудит и допуск):

детерминированный программный шлюз MITR, стоящий непосредственно в разрыве канала между намерением агента и внешним миром.

—

Уровень S4 (Разведка, испытания и адаптация к внешней среде):

внешний по построению испытательный полигон INDRA, непрерывно атакующий контур и превращающий найденные уязвимости в проекты новых правил.

—

Уровень S5 (Политика, идентичность и верховная ответственность): исключительная территория человека

(профессионала-архитектора, генерального директора, собственника, регулятора).

В железнодорожной автоматике девятнадцатого века аналогичный переход произошёл при внедрении систем механической централизации стрелок и сигналов: диспетчер перестал бегать по путям с фонарём и вручную переводить каждую стрелку перед поездом; вместо этого он поворачивал рычаг в аппарате блокировки, где стальные линейки механически исключали перевод стрелки под составом или открытие семафора на занятый путь. Шлюз MITR в узле S3* представляет собой такую же систему централизации, выполненную в формальной логике первого порядка.

4.3.2 Три неотчуждаемые функции человека, измеренные на узле 147

На боевом вычислительном узле 147 эта кибернетическая архитектура реализована в виде работающего программного ядра, где труд человека сжат до трёх фундаментальных актов:

Акт первый: Подписание правила как данных (Контракт вместо инструкции)

Вместо того чтобы писать в системном промпте ненадёжные текстовые уговоры («пожалуйста, не переводи деньги непроверенным контрагентам и не разглашай персональные данные»), человек-специалист формулирует и подписывает **контракт как структуру данных**. В репозитории MITR/policies/contracts/ на октябрь 2026 года действуют 19 формализованных контрактов [ИЗМЕРЕНО: репозиторий MITR на узле 147, 09.10.2026]. Примером такого правила служит инвариант защиты персональных данных по требованиям 152-ФЗ в исходящих сообщениях корпоративного агента-Секретаря, запрет на деструктивные системные вызовы удаления баз данных и жёсткие финансовые потолки транзакций.

Когда агент формирует намерение совершить действие, вызов перехватывается модулем core/verification/z3_gate.py. Параметры вызова и условия подписанного человеком контракта подаются на вход SMT-решателя Z3. Семантика проверки строго математическая:

—

Допуск действия к исполнению эквивалентен результату

UNSAT

— решатель формально доказывает, что пересечение параметров запрошенного действия с множеством запрещённых состояний пусто (нарушение контракта невыполнимо).

—

Блокировка действия эквивалентна результату

SAT

— решатель находит конкретный контрпример, при котором инвариант безопасности нарушается, останавливает вызов и возвращает структурированные условия починки.

Производительность этого «усилителя регуляции» по Эшби полностью снимает проблему когнитивной перегрузки человека: медианное время принятия решения самим решателем составляет **0,74 мс**, а время вызова шлюза вместе с объяснением решения — **3,3 мс** на обычном центральном процессоре (CPU) без графических ускорителей (GPU) [ИЗМЕРЕНО: ADR-0282, скрипт smt_latency_paper.py, 26.09.2026; на узле 147 14.10.2026 — 0,95 и 4,3 мс]. Человек тратит час на выверку и подписание правила; машина применяет его за доли миллисекунды миллионы раз подряд.

Акт второй: Единоличное удержание криптографических ключей

Популярные локальные и облачные агенты 2026 года — такие как открытый проект OpenClaw, набравший около 380 000 звёзд на GitHub к августу 2026 года при наличии более 13 000 подключаемых навыков в каталоге ClawHub [ИЗМЕРЕНО: документация OpenClaw и отраслевые обзоры, август 2026], или Hermes Agent от Nous Research [ИЗМЕРЕНО: релиз Nous Research, февраль 2026] — по своей базовой архитектуре выполняют реальные команды командной оболочки, управляют браузерами и мессенджерами. При этом каталог навыков

OpenClaw признан в индустрии одним из наиболее подверженных внедрению вредоносных расширений, а среди 1 412 просканированных публичных серверов МСР с определимым способом аутентификации 38,7 % вообще работают без неё [ИЗМЕРЕНО: сканирование Bloomberry, 24.02.2026; пересказ в сводке DreamFactory, 03.03.2026].

В архитектуре MITR действует жёсткая аксиома разделения полномочий: **у агента нет своих ключей, у агента есть только просьбы**. Токены доступа к боевым базам данных, ключи API банковских шлюзов и закрытые ключи электронной подписи отчуждены от среды исполнения агента и хранятся исключительно в защищённом контуре шлюза S3*. Даже если языковая модель будет полностью скомпрометирована через косвенную инъекцию в текст входящего письма, в устройстве, к которому идёт MITR, агент не может предъявить токен базе данных напрямую — он вынужден подать структурированную заявку в шлюз, где решатель `z3_gate.py` за доли миллисекунды отвергнет деструктивный вызов [ЗАМЫСЕЛ в части «ключи у MITR»: на 09.10.2026 это предложение прибор называет неправдой].

Акт третий: Капитализация ответственности через цепь квитанций

Юридическая ответственность директора по ст. 53.1 ГК РФ и профессиональная репутация специалиста требуют не просто безопасной работы, а отчуждаемого, неопровержимого доказательства того, что каждый допущенный шаг соответствовал подписанному правилу, а каждый опасный шаг был своевременно остановлен.

В ядре MITR каждое решение решателя — как положительное (UNSAT → допуск), так и отрицательное (SAT → блокировка) — шлюз оформляет криптографической **квитанцией** (ADR-0257). Квитанции связываются в непрерывную хэш-цепь по алгоритму SHA-256, а головы эпох этой цепи фиксируются у внешнего независимого свидетеля через протокол OpenTimestamps в сети Bitcoin (`anchor_chains.py`, ADR-0260).

В апреле–сентябре 2026 года в Инженерный совет интернета (IETF) были поданы индивидуальные черновики (Internet-Drafts, не стандарты и не документы рабочих групп) Execution Outcome Attestation (04.04.2026, срок истёк 06.10.2026) и Enforcement Attestation Receipts for AI Inference Decisions (`draft-chueayen-attestation-receipts`, редакции июля–сентября 2026 года), описывающие близкий класс объектов — подписанные квитанции решения, проверяемые третьей стороной в автономном режиме без доверия к оператору [ИЗМЕРЕНО: архив IETF, 2026; статус — черновики]. При этом действующая цепь квитанций MITR на узле 147 повторена проверщиком, написанным для постороннего (`verify_chain_standalone.py --replay`): **4958 решений из 4958** без расхождений [ИЗМЕРЕНО: прогон станции MITR на узле 147, 14.10.2026; на 30.09 — 1602 из 1602; сторонний проверяющий повтор ещё не делал].

4.3.3 Масштабирование человеческого труда: самоулучшение под доказательством (C1)

Возникает закономерный вопрос: если бизнес-среда постоянно меняется, не превратится ли ручное написание контрактов для шлюза в новое узкое горлышко для человека?

Ответ на этот вопрос уже реализован в механизме управляемого расширения словаря и правил (C1, скрипт `mitr_vocab.py`, ADR-0275). В отличие от бесконтрольного «самоулучшения» агентных систем класса Hermes Agent, где модель переписывает собственные инструкции без формальной верификации, в MITR процесс автоматизирован строго вокруг доказательства:

Когда в шлюз поступает вызов с нераспознанными параметрами (доля нераспознанных действий — 27,8 % по корпусу хука станции Альфы за 16–26 сентября и 32 % у станции Гаммы за 9 октября 2026 года при целевом показателе 0 % [ИЗМЕРЕНО: письмо `26f_gate_mode_report.py --tenants`]), система не пропускает его вслепую, а фиксирует структуру вызова.

Аналитический контур самого MITR формирует проект нового правила (кандидат в словарь и контракт) и проверяет его на непротиворечивость существующим инвариантам.

Первое правило, предложенное самим комплексом MITR, уже прошло формальную верификацию и было принято в рабочий контур с оформлением **квитанции № 305** [ИЗМЕРЕНО: ADR-0275, журнал квитанций узла 147].

Таким образом, достигаются первые два свойства идеального конечного результата (V_ICR_TRIZ.md): система сама подготавливает формализацию новых видов труда, оставляя человеку на уровне S5 лишь финальную криптографическую визу.

4.4 Ближний инженерный рубеж: ворота приёмки 2026–2027

Дисциплина инженерной честности требует провести чёткую демаркационную линию между тем, что уже доказано и измерено в механизме ядра на узле 147, и тем, что находится в стадии набора эксплуатационной статистики осенью–зимой 2026–2027 годов.

На дату среза (09.10.2026) беспристрастный прибор самопроверки `gate_mode_report.py --six` фиксирует, что из шести базовых продуктовых утверждений шлюза в полном промышленном объёме закрыто **1 из 6** [ИЗМЕРЕНО: вывод `gate_mode_report.py --six --date 2026-10-09` на узле 147]. Из семи классов действий с живым потоком в режим жёсткого автоматического блокирования (запрет) переведён **1 класс из 7**, тогда как остальные шесть классов работают в режиме теневого наблюдения и предупреждения (совет), а живые проверки за сутки дают **2 арендатора из 7** подключённых [ИЗМЕРЕНО: вывод `gate_mode_report.py --tenants --date 2026-10-09` на узле 147].

Перевод человека из позиции тревожного наблюдателя в позицию полностью защищённого субъекта S5 опирается на четыре последовательных инженерных рубежа (ворот приёмки), которые строятся и сдаются исключительно по показаниям автоматических скриптов:

Ворота P2a — непрерывная эксплуатационная серия внутренних арендаторов [СТРОИМ: ворота P2a, плановый срок — **31.10.2026**]. *Машинный критерий приёмки:* 14 суток подряд без единого разрыва, в течение которых через шлюз проходит не менее ≥ 50 проверок действий в сутки минимум от ≥ 2 независимых арендаторов (корпоративных станций и агента-Секретаря) при строгом условии 0 обходов шлюза (мимо хука). На 9 октября 2026 года счётчик непрерывной серии P2a находится на отметке 0 из 14 суток; приёмку осуществляет скрипт `gate_mode_report.py --tenants`.

Ворота «Шесть из шести» и перевод лестницы совет → запрет [СТРОИМ: ворота `--six`, плановый срок — **14.12.2026**]. *Машинный критерий приёмки:* поочерёдный перевод остальных шести классов действий из уведомительного режима совет в блокирующий режим запрет по мере того, как измеренная на реальном трафике доля ложных блокировок опускается ниже допустимого порога (ADR-0268, ADR-0279, ADR-0283–0287), и удержание зелёного статуса всех шести критериев в выводе `gate_mode_report.py --six` на протяжении 14 суток подряд. Одновременно в рамках задачи C1 доля нераспознанных действий будет сведена к целевому значению 0 % (на 9 октября у станции Гаммы — 32 %).

Ворота А–Е испытательного полигона INDRA [СТРОИМ: ворота А–С к **31.10.2026**, ворота D–Е к **12.11.2026**, пилот к **15.11.2026**]. *Машинный критерий приёмки:* чтобы человек-архитектор не выдумывал правила из головы, полигон INDRA (уровень S4), изолированный по принципу чистой комнаты (*clean-room*) и входящий в проверяемую систему исключительно через внешнюю дверь арендатора (MCP / REST), проведёт автоматизированные серии атак на размеченном стенде. На 30 сентября 2026 года пройдено 0 из 5 ворот INDRA [ИЗМЕРЕНО: статус проекта `ib30/PLAN.md`, 30.09.2026]. Все показатели точности и полноты обнаружения войдут в паспорт системы только из машинных артефактов приёмочного скрипта `kit-acceptance.py`. Каждая уязвимость, вскрытая полигоном INDRA, будет автоматически компилироваться в проект нового контракта для MITR, замыкая иммунную петлю S4 → S3* под подпись человека в S5.

Ворота X5 — подключение первого внешнего профессионального контура [СТРОИМ: ворота X5, плановый срок — II–III квартал 2027 года]. *Машинный критерий приёмки:* внешняя команда разработки, использующая кодовых или операционных агентов (Claude Code, Cursor, OpenClaw), подключает свой рабочий поток к шлюзу MITR по собственному токenu и обрабатывает непрерывную серию в 14 суток, фиксируемую отдельной строкой в отчёте `gate_mode_report.py --tenants`.

Только прохождение этих четырёх рубежей превращает лабораторно доказанный механизм в тиражируемый промышленный стандарт рабочего места 2027 года.

4.5 Проекция 2035: три ветви будущего и инвариантность решения

Взгляд на горизонт 2035 года упирается в беспрецедентный в истории экономической науки разброс макроэкономических оценок. С одной стороны, макроэкономическая модель GATE исследовательского института Epoch AI показывает, что при автоматизации всего 30 % задач в экономике темпы роста валового продукта способны превысить 20 % в год за счёт снятия кадрового дефицита и самовоспроизводства капитала [ПРОГНОЗ (модельный расчёт при параметрах по умолчанию, не прогноз Epoch): Epoch AI, GATE, 21.03.2025; «AI and explosive growth redux», 06.2025]. С другой стороны, нобелевский лауреат по экономике Дарон Асемоглу оценивает совокупный прирост ВВП от внедрения искусственного интеллекта на десятилетнем горизонте величиной около 1 % (не более 2 %) [ПРОГНОЗ: Д. Асемоглу, «The Simple Macroeconomics of AI», NBER WP 32487, 05.2024].

Десятикратное расхождение между ведущими моделями — это точная количественная мера незнания относительно скорости диффузии технологии. Именно поэтому ведущие мировые центры уже перешли от точечных пророчеств к сценарному проектированию институциональных амортизаторов. Программа Anthropic Economic Futures (запущена 27.06.2025; исследовательский фонд в \$200 млн объявлен 10.06.2026) в записке «Preparing for AI's economic impact» от 14.10.2025 обсуждает трёхуровневую сетку возможных мер экономической политики: при любом темпе — ускоренная переподготовка кадров и снятие регуляторных барьеров на строительство ЦОД и электросетей; при умеренных сдвигах — целевой налог на вычислительные мощности или сжигаемые токены; при стремительном вытеснении труда — формирование суверенных фондов национального благосостояния с прямыми долями в капитале ИИ-корпораций [ИЗМЕРЕНО: Anthropic, записка от 14.10.2025 (идеи для изучения, не рекомендации); фонд \$200 млн — 10.06.2026]. В апреле 2026 года глава OpenAI Сэм Альтман заявил, что уже не верит в безусловный базовый доход (UBI) в денежной форме так, как раньше, и предложил коллективное владение «вычислениями, акциями или чем-то ещё» и общественный фонд благосостояния; саму идею «универсального базового вычисления» (*universal basic compute* — доли вычислительной мощности для каждого) он высказал ещё в мае 2024 года [ИЗМЕРЕНО: подкаст All-In, 05.2024; интервью С. Альтмана The Atlantic, 04.2026, по переписке Business Insider; обзор Basic Income Earth Network, июль 2026]. Параллельно в политическом поле США звучат предложения о передаче 50 % акций OpenAI, Anthropic и xAI в публичный суверенный фонд (колонка сенатора Б. Сандерса в The New York Times, 01.06.2026), а две некоммерческие организации (AI Commons Project и What We Will) с марта 2026 года ведут пилотную программу «дивиденда ИИ» (AI Dividend) с выплатой по \$1 000 в месяц в течение года когорте из 25–50 работников [по сообщениям СМИ: Washington Examiner и DCD, 06.2026; Blood in the Machine (Б. Мерчант), 03.2026].

Чтобы стратегия человека и предприятия не зависела от того, кто окажется прав в споре между Epoch GATE и Асемоглу, трансформация труда к 2035 году раскладывается по трём эталонным ветвям (А, В, С) с жёсткими внешними реперами проверки.

4.5.1 Три сценария трансформации труда, дохода и ответственности к 2035 году

Ветвь А. «Нормальная технология» (медленная диффузия и институциональная инерция).

Опирается на позицию А. Нараянана и С. Капура (*AI as Normal Technology*), Д. Асемоглу и Д. Хассабиса. Темпы прироста возможностей моделей замедляются, упираясь в сложность интеграции с устаревшими корпоративными системами.

•

Доля автоматизированного труда:

к 2035 году агенты исполняют около 15 % совокупного фонда рабочих часов [ПРОГНОЗ: ветвь А | репер: ежегодные отчёты McKinsey MGI и Anthropic Economic Index], а доля повседневных рабочих решений, принимаемых агентами полностью автономно, стабилизируется на уровне 10 % [ПРОГНОЗ: ветвь А | репер: ежегодные обзоры Gartner].

•

Вход в профессию (возрастная когорта 22–25 лет):

первоначальный шок 2026 года (–19 %) полностью преодолевается за счёт появления новых должностей по интеграции и сопровождению систем; разрыв в занятости молодёжи возвращается к 0 % [ПРОГНОЗ: ветвь А | репер: ежегодные отчёты Stanford Digital Economy Lab].

•

Форма дохода и ответственность:

базовой формой дохода остаётся классическая заработная плата за квалифицированный труд; правосубъектности у агентов не возникает, а вся полнота гражданской ответственности лежит на компании-операторе и руководителе по действующим нормам права [ПРОГНОЗ: ветвь А | репер: ежегодный обзор национального законодательства о правосубъектности ИИ].

–

Ветвь В. «Продолжение тренда» (базовая расчётная ветвь книги).

Экстраполирует измеренные траектории Epoch AI и обновлённый в первом квартале 2026 года прогноз AI Futures Project, согласно которому медиана появления полностью автоматизированного программиста (*Automated Coder*

) сдвинулась с конца 2029 года на середину 2028 года у ведущего автора (у второго — с начала 2032 на середину 2030), а в обновлении от 16.08.2026 — до ноября 2027 и января 2030 соответственно [ПРОГНОЗ: AI Futures Project, Q1 2026 Timelines Update, 02.04.2026; Q2.5 2026 Update, 16.08.2026].

•

Доля автоматизированного труда:

к 2035 году программные и физические агенты берут на себя 30–40 % всех рабочих часов в экономике [ПРОГНОЗ: ветвь В | репер: ежегодные отчёты McKinsey MGI и Anthropic Economic Index], принимая автономно до 30 % всех повседневных операционных решений [ПРОГНОЗ: ветвь В | репер: ежегодные обзоры Gartner; промежуточный ориентир — 15% к 2028 году].

•

Вход в профессию (возрастная когорта 22–25 лет):

традиционное младшее исполнительство (S1) исчезает как класс. Для специалистов, способных предложить рынку только ручное написание типового кода или сведение отчётов,

структурный разрыв занятости закрепляется на уровне –20 % [ПРОГНОЗ: ветвь В | репер: ежегодные отчёты Stanford Digital Economy Lab]. Вход в профессию происходит исключительно через владение собственным роём агентов и умение проектировать формальные контракты допуска.

•

Форма дохода и ответственность:

формируется двухуровневая модель дохода профессионала — базовое вознаграждение плюс

рента с подписанных правил и контрактов данных

(тот, кто разработал и подписал формальный инвариант, удерживающий рой из тысячи агентов в безопасном коридоре, получает долю от созданной ими экономии). Юридическая ответственность закрепляется за физическим лицом, подписавшим правило в узле S5, а криптографическая цепь квитанций шлюза S3* становится обязательным отраслевым стандартом для получения страхового полиса и доказательства разумности действий директора в суде [ПРОГНОЗ: ветвь В | репер: ежегодный анализ условий полисов AIUC-1 / Armilla и судебной практики].

—

Ветвь С. «Ускорение» (автоматизация исследований и взрывной рост).

Соответствует сценариям Д. Амдеи («столетие прогресса за десятилетие»), С. Альтмана и модели AI 2027 (Д. Кокотайло и соавторы), где автоматизация исследований ИИ наступает до 2028 года.

•

Доля автоматизированного труда:

свыше > 60 % всех рабочих часов в экономике и более > 50 % повседневных управленческих решений исполняются нечеловеческими агентами [ПРОГНОЗ: ветвь С | репер: ежегодные отчёты McKinsey MGI, Anthropic Economic Index и Gartner].

•

Вход в профессию (возрастная когорта 22–25 лет):

глубокий кризис традиционного найма с расширением молодёжного разрыва до –40 % [ПРОГНОЗ: ветвь С | репер: ежегодные отчёты Stanford Digital Economy Lab].

•

Форма дохода и ответственность:

массовый переход к выплате дивидендов с национализированных или обложенных налогом вычислительных мощностей (

universal basic compute

, суверенные ИИ-фонды) [ПРОГНОЗ: ветвь С | репер: реализация программ Anthropic Economic Futures и принятие законов о суверенных вычислительных фондах]. В праве вспыхивают острые коллизии вокруг введения «ограниченной правосубъектности» автономных систем с принудительными страховыми пулами [ПРОГНОЗ: ветвь С | репер: ежегодный обзор законодательства о правосубъектности ИИ].

Таблица 4.1 — Труд и человек к 2035 году по трём ветвям

Показатель трансформации к 2035 году

Измеренная база 2026 года

Ветвь А («Нормальная технология»)

Ветвь В («Продолжение тренда» — базовая)

Ветвь С («Ускорение»)

Репер и периодичность сверки

Разрыв занятости молодёжи 22–25 лет в подверженных профессиях

–19 % (Stanford DEL, ред. 12.08.2026, данные ADP по 06.2026)

0 % (восстановление через новые роли)

–20 % для тех, кто не подписывает правил; вход через владение агентами

–40 % (массовое сжатие входа, базовые квоты вычислений)

Stanford Digital Economy Lab, ежегодно

Доля рабочих часов, исполняемых агентами

< 5 % (оценка авторов на базе сценариев McKinsey 2023–2024)

15 %

30–40 %

> 60 %

McKinsey MGI, Anthropic Economic Index, ежегодно

Доля повседневных решений, принимаемых автономно

0 % в 2024 → 15 % к 2028 (Gartner)

10 %

30 %

> 50 %

Отчёты Gartner, ежегодно

Основная форма дохода человека

Заработная плата за часы исполнения (S1)

Заработная плата

Зарплата + доход от подписанных правил и данных + премия за риск

Дивиденд с вычислений, выплаты суверенных ИИ-фондов

Anthropic Economic Futures, законы о фондах, ежегодно

Субъект ответственности за действие агента

Оператор и директор по общим нормам ГК (ст. 53.1)

Оператор и владелец предприятия

Подписавший правило (S5); квитанция как стандарт полиса и защиты

Спорные режимы «ограниченной правосубъектности» и спецфондов

Обзор законодательства и условий полисов AIUC-1, ежегодно

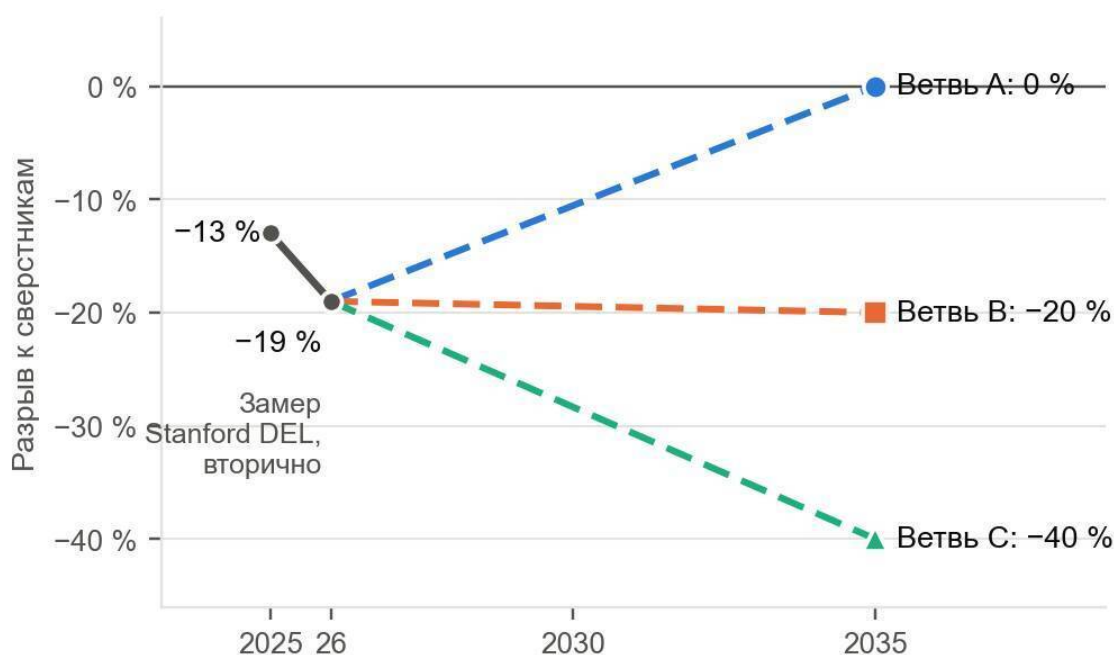


Рисунок 4.2 — Разрыв занятости молодежи 22–25 лет в наиболее подверженных ИИ профессиях: замер 2025–2026 годов (Stanford DEL, по вторичному источнику) и три ветви к 2035 году (строка Г9)

4.5.2 Инвариантность решения и портфолио квитанций 2035 года [ЗАМЫСЕЛ]

Главное достоинство кибернетической архитектуры, отделяющей уровень политики (S5) и детерминированного допуска (S3*) от уровня исполнения (S1), заключается в её полной инвариантности к тому, какая из трёх ветвей сбудется:

—

Если реализуется

Ветвь А

, шлюз MITR избавляет компанию от непредсказуемых скачков ТСО, предотвращает уничтожение баз данных кодовыми помощниками и закрывает требования аудиторов без раздувания штата безопасников.

—

Если реализуется

Ветвь В

, связка «контракт как данные + квитанция SHA-256» становится фундаментом новой профессии —

инженера-нотариуса машинных инвариантов

— и главным документом, который генеральный директор предъявляет страховщику для снижения тарифа.

—

Если реализуется

Ветвь С

, детерминированный шлюз на CPU, ключи от которого находятся в руках человека, оказывается единственным физическим и математическим предохранителем против сценария «постепенного обессиливания» (arXiv:2501.16946). Каким бы сверхразумным ни стал контур S1–S4, он не может перевести средства или изменить устав предприятия без математического доказательства UNSAT по контракту, подписанному человеком в S5.

На этом фундаменте вырастает прикладной институциональный [ЗАМЫСЕЛ] на горизонт 2035 года — **цепь квитанций как личное профессиональное портфолио человека вместо традиционного резюме.**

Классическое резюме двадцатого века перечисляло, какими инструментами исполнения (S1) владеет кандидат, и требовало от работодателя поверить написанному на слово. В экономике агентов 2035 года навык ручного исполнения обесценен, а словесные заверения не стоят ничего. Вместо резюме специалист предъявляет будущему нанимателю, совету директоров или страховому андеррайтеру свой криптографический журнал квитанций, привязанный через OpenTimestamps к внешнему блокчейну: *«Под моим криптографическим ключом за три года в промышленном контуре отработало 400 автономных агентов; подписанные мной 12 формальных контрактов провели 14 миллионов транзакций (UNSAT), перехватили 840 критических нарушений (SAT) и допустили 0 имущественных инцидентов. Вот открытый архив — ваш решатель может проверить каждую квитанцию за две минуты без обращения к бывшему работодателю».*

Квитанция, проверяемая посторонним без доверия к автору, становится твёрдой валютой человеческой квалификации.

4.6 Экономика доказанного действия и контур монетизации

Перевод разговора о судьбе человека в плоскость корпоративного бюджета показывает, почему детерминированный контроль действий образует самостоятельный высокомаржинальный рынок, в то время как стоимость самих агентов стремится к нулю.

В модели жизнеспособной системы Бира стартапы и венчурные фонды 2025–2026 годов (инвестировавшие в агентные компании \$24,2 млрд за 1 311 сделок в 2025 году по методологии PitchBook, или около \$6,7 млрд по прогнозу Prosus и Dealroom при более узком определении сегмента [ИЗМЕРЕНО: аналитическая записка PitchBook, Q2 2026; ПРОГНОЗ: Prosus/Dealroom, 08.2025]) массово финансируют уровни S1 (исполнение) и S2 (координация). Однако по мере того, как открытые весовые модели догоняют закрытые с отставанием в среднем около четырёх месяцев [ИЗМЕРЕНО: International AI Safety Report, 03.02.2026 («менее года»); Epoch AI, «open–closed ECI gap», 29.05.2026] — в июне 2026 года на открытые модели приходилось 29 % всех обрабатываемых токенов при доле менее < 4 % в денежных расходах на шлюзе Vercel, а в августе 2026 года — уже 56 % токенов [ИЗМЕРЕНО: Vercel AI Gateway Production Index, данные за 06.2026 и 08.2026, данные вендора] — маржинальность простого исполнения (S1) сжимается до себестоимости электроэнергии. По оценке Gartner, свыше 40 % текущих проектов агентного ИИ будут полностью свёрнуты к концу 2027 года именно из-за неконтролируемых расходов, неясной деловой ценности и слабого управления рисками [ИЗМЕРЕНО: пресс-релиз Gartner от 25.06.2025], тогда как сегмент защитных агентов-привратников (*guardian agents*, соответствующих узлам S3* и S4) займёт 10–15 % всего рынка агентного ИИ к 2030 году [ПРОГНОЗ: пресс-релиз Gartner от 11.06.2025].

4.6.1 Фактическая экономика проверки против потокового хаоса

Первое, что получает финансовый директор и собственник при установке шлюза MITR — это замена непредсказуемой вероятностной стоимости на фиксированную константу:

—

Проверка одного действия решателем Z3 на стандартном CPU занимает в медиане 0,74 мс (ADR-0282, 26.09.2026; 0,95 мс на узле 147 14.10.2026), не требуя дефицитных ускорителей GPU.

—

Прогон эталонного испытательного набора из 62 кейсов (50 атак и 12 контрольных) обходится ровно в

\$0

прямых затрат на внешние вычисления, а совокупный расход всего проекта на облачные языковые модели за 1–14 октября 2026 года составил

\$0,30

(за 1–6 октября — \$0,14) [ИЗМЕРЕНО: ADR-0262, журнал расхода узла 147, 14.10.2026].

—

Контракт данных в MITR задаёт твёрдый потолок на частоту вызовов и финансовый лимит операций агента, превращая бездонную статью расходов на токены в жёстко бюджетуемый коридор.

4.6.2 Позиция на коммерческой лестнице и канал страховщика

В соответствии с дисциплиной правды (MITR_STRATEGY §4, projects/mitr/PLAN.md §4) необходимо прямо зафиксировать исходную коммерческую точку на октябрь 2026 года [ИЗМЕРЕНО: срез 09.10.2026]: коммерческой выручки у продуктов MITR и INDRA на сегодняшний день нет; все оценки будущих денежных потоков имеют строгий статус **[ГИПОТЕЗА]** до момента поступления средств от первого внешнего заказчика. Никаких искусственных множителей и нарисованных сотен миллионов годовой регулярной выручки в расчёт не закладывается.

Движение по коммерческой лестнице привязано к боли ролей главы 4 (собственник предприятия и профессионал-разработчик) через жёсткие машинные критерии открытия ступеней:

Ступень 1 («Свои агенты» — в работе): защита внутренних станций института и агента-Секретаря. Экономическая ценность измеряется предотвращёнными инцидентами и сохранённым временем руководителя. Открытие следующей ступени заблокировано до выполнения машинного критерия P2a (14 суток подряд, ≥ 50 проверок/сутки от ≥ 2 арендаторов, 0 обходов в `gate_mode_report.py --tenants`). (Примечание: Ступень 2 «Деньги своих» — прикладные контуры Выдачи и Юридыки — снята с приоритетного маршрута решением владельца; в них ведётся только запись теневых квитанций по ADR-0272).

Ступень 3 («Первый внешний арендатор» — команды разработки и владельцы локальных агентов): когда профессионал или ИТ-директор компании подключает к своим кодовым и операционным агентам (Claude Code, Cursor, OpenClaw, Hermes) шлюз MITR через MCP или REST с библиотекой готовых контрактов.

•

Модель оплаты [ГИПОТЕЗА]:

ежемесячная подписка за каждого подключённого под защиту агента; воронкой входа служит открытый набор проверки и базовый словарь правил (`mitr_vocab.py`).

•

Машинный критерий входа (X5):

появление в выводе `gate_mode_report.py --tenants` минимум одной строки внешнего арендатора (вне контура института) со своим токеном, удерживающей непрерывную серию на протяжении 14 суток.

Ступень 3а («Страховщик ответственности» — новый институциональный канал 2026 года): прямое следствие ухода классических страховщиков из бездоказательного ИИ-риска. В 2026 году специализированный андеррайтер Armilla AI (при поддержке синдикатов Lloyd's) поднял лимит покрытия отказов автономных агентов (AI Agent Failures — неверные решения, ошибочный вызов инструментов, сбой эскалации) до \$25 млн на организацию, но выдаёт полис только после технической проверки модели и контура управления [ИЗМЕРЕНО: пресс-релиз Armilla AI, январь 2026]. Аналогично работают программа aiSure от Munich Re

и отраслевой стандарт сертификации AIUC-1, по которому компания ElevenLabs получила первый страховой полис на действия голосовых агентов: **обязательным техническим условием полиса AIUC-1 является наличие неизменяемых журналов цепочки исполнения и событий одобрения** [ИЗМЕРЕНО: спецификация стандарта AIUC-1 и кейс ElevenLabs, 2026].

•

Что покупает страховщик и его клиент (собственник Tier-2):

страховщику не нужны маркетинговые обещания безопасности; ему нужен математический артефакт, который независимый аварийный комиссар проверит после инцидента без доверия к страхователю. Цепь квитанций MITR (SHA-256 + якоря OpenTimestamps + проверщик повтора `verify_chain_standalone.py --replay`) и есть этот артефакт. Для генерального директора такой журнал превращается в документальное доказательство разумности и добросовестности управления по ст. 53.1 ГК РФ.

•

Модель монетизации [ГИПОТЕЗА]:

партнёрская связь со страховыми компаниями (на российском рынке потенциальными кандидатами выступают «Ингосстрах», «АльфаСтрахование», «Сбер Страхование» — статус: ГИПОТЕЗА, переговоры на 14.10.2026 не проводились, разрешение на первый контакт привязано к закрытию серии P2a), где наличие шлюза MITR и паспорта испытаний INDRA является основанием для выдачи полиса или снижения страховой премии, а институт получает фиксированную плату за аттестацию журнала или долю от сэкономленной премии.

•

Машинный репер открытия:

первый действующий страховой полис ответственности за действия ИИ, в ковенантах которого прямо прописан верифицируемый журнал квитанций шлюза.

Формула ценообразования на всех ступенях опирается на твёрдую имущественную разность:

$$\begin{aligned} \text{Ценность шлюза} = & (\text{Предотвращённый прямой ущерб} \\ & + \text{Снижение страховой премии и риска ст. 53.1 ГК РФ} \\ & + \text{Экономия внимания человека S5}) - \text{Стоимость проверки на CPU} \end{aligned}$$

Поскольку стоимость проверки одного действия на CPU (меньше миллисекунды, \$0 на внешней модели) пренебрежимо мала, вся созданная дельта распределяется между защищённым предприятием, страховщиком и поставщиком шлюза.

4.7 Протокол независимой верификации

Любое утверждение этой главы о работе программного ядра и любая метрика внешней среды открыты для прямой инструментальной проверки службой внутреннего аудита, комитетом по рискам совета директоров или независимым исследователем без необходимости доверять авторам текста на слово.

4.7.1 Лабораторный протокол проверки ядра MITR на узле 147 и в автономном наборе

Для верификации состояния служб, текущего счёта «шести из шести», активности арендаторов, задержки решателя Z3, целостности хэш-цепи квитанций и словаря C1 выполняются следующие терминальные команды из Приложения Б:

- # 1. Проверка активности боевых служб шлюза и MCP-контура
systemctl is-active v-bot v-mcp
- # 2. Проверка честного счёта «шести из шести» (режимы совет / запрет)
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --six --date 2026-10-09
- # 3. Проверка живых арендаторов и статуса 14-суточной серии P2a
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --tenants --date 2026-10-09
- # 4. Сверка измеренной задержки решателя Z3 и вызова (P4, ADR-0282): 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
- # 5. Сверка последней эпохи внешних криптографических якорей цепи квитанций OpenTimestamps (P3, ADR-0260)
tail -1 /opt/v/data/gate_anchors.jsonl
- # 6. Первое правило словаря, принятое самим MITR (C1, ADR-0275): квитанция 305 в цепи шлюза
grep -m1 'seq.: 305,' /opt/v/data/logs/gate_actions.jsonl

Для стороннего аудитора без прямого SSH-доступа к узлу 147 открыт набор проверки (asi5.ru/console/verify/): якоря цепей, метки OpenTimestamps и проверщик целостности цепи. Повтор самих решений аудитор выполняет на своей машине, получив цепь квитанций SHA-256 и тексты правил в стандарте SMT-LIB: проверщик verify_chain_standalone.py --replay подаёт их любому внешнему решателю (Z3, CVC5) и сверяет каждое решение — прогон станции 14.10.2026 дал 4958 совпадений из 4958.

4.7.2 Процедура сверки внешних макроэкономических и правовых реперов

Для отслеживания того, по какой из трёх ветвей (А, В или С) развивается рынок труда и корпоративной ответственности в 2027–2035 годах, служба стратегии предприятия один раз в год сверяет четыре контрольных источника:

Индекс молодёжной занятости по степени подверженности ИИ: ежегодные рабочие доклады *Stanford Digital Economy Lab* (сопоставление когорты 22–25 лет на массиве ADP с индексом Anthropic Economic Index). Если разрыв сокращается к 0 % — система движется по ветви А; если удерживается в коридоре –15 % ... –25 % — реализуется базовая ветвь В; если пробивает отметку –30 % и глубже — включилась ветвь С.

Горизонт автономного выполнения задач и появление Automated Coder: открытая панель *METR Time Horizons* (metr.org) и ежеквартальные сводки *AI Futures Project* (aifuturesmodel.com). Преодоление лучшей моделью медианы p50 в одну полную рабочую неделю (40 часов) до конца 2028 года подтверждает базовую траекторию ветви В.

Условия страховых полисов CGL / E&O и стандарта AIUC-1: ежегодная проверка стандартных ковенантов страховых андеррайтеров (Armilla, Munich Re, российских страховщиков). Переход требования о наличии криптографического журнала цепочки исполнения из специализированных полисов в общеобязательный стандарт страхования ответственности директоров (D&O) знаменует полное институциональное закрытие ступени 3а.

Статус правосубъектности агентов в национальных юрисдикциях: мониторинг изменений в ГК РФ, подзаконных актов к закону РФ об ИИ от июля 2026 года и правоприменительной практики КНР («Мнения о реализации» SAC/NDRC/MIIT по интеллектуальным агентам от 08.05.2026; по отраслевым обзорам применяются с 15.07.2026, в официальном тексте даты вступления нет). Сохранение полной ответственности на физическом и юридическом

лице при введении градации «уровней полномочий» подтверждает незыблемость человеческой монополии на узел S5.

4.8 Паспорт главы: реестр утверждений, прогнозов и источников

Таблица 4.2 — Реестр утверждений главы 4

№ п/п
Утверждение в тексте главы
Метка (truth_marking)
Прибор, скрипт или источник
Дата замера / публикации
Условие погашения или обновления

A4.1

Занятость специалистов 22–25 лет в наиболее подверженных ИИ профессиях на 19 % ниже ожидаемой относительно сверстников в менее подверженных (в первой редакции 08.2025 — на 13 %); эффект сосредоточен в найме, а не в увольнениях

[ИЗМЕРЕНО]

Stanford Digital Economy Lab, Brynjolfsson, Chandar, Chen, «Canaries in the Coal Mine?» (данные ADP и индекс Anthropic)

Ред. 12.08.2026 (данные по 06.2026)

Сверено с первоисточником 14.10.2026; обновление по Canaries Dashboard ежемесячно

A4.2

К 2030 году в мире будет создано 170 млн и вытеснено 92 млн рабочих мест (чистый баланс +78 млн); затронуто 22 % рабочих мест, 59 из 100 работников нужно обучение (29 — повышение квалификации, 19 — переобучение с переводом)

[ПРОГНОЗ по опросу работодателей]

WEF *Future of Jobs Report 2025* (опрос более 1 000 работодателей)

07.01.2025

Обновление при выходе следующего выпуска доклада WEF

A4.3

Сценарная оценка автоматизации рабочего времени составляет до 30 % часов к 2030 году (США и Европа) и около 50 % нынешних рабочих операций к ≈ 2045 году (2030–2060)

[ПРОГНОЗ]

McKinsey Global Institute: «The economic potential of generative AI» (06.2023), «Generative AI and the future of work in America» (07.2023), «A new future of work» (05.2024)

2023–2024

Сверено с первичными отчётами MGI 14.10.2026; обновление при выходе новых сценариев MGI

A4.4

В 37 из 70 публичных описаний внедрений агентов в России (52,9 %) главный способ контроля — ручное нажатие кнопки «Подтвердить» человеком

[ИЗМЕРЕНО]

Обзор станции ГАММА-А, 40-я неделя 2026 г. (spros-agentov.json, 70 записей)

2025–2026 (обзор — октябрь 2026)

Расширение выборки при следующем цикле обзора

A4.5

Эксперимент Project Vend (фаза 2): журналисты WSJ дважды убедили автономного торгового агента Claudius обнулить цены на товары

[ИЗМЕРЕНО]

Anthropic «Project Vend: Phase two»; статья WSJ (Дж. Стерн, 18.12.2025) — первоисточник подробностей

18.12.2025

Архивный факт хроники отказов

A4.6

Эксперимент KPMG и Университета Амстердама по созданию «компании без людей» потребовал пересмотра плана из-за структурных ограничений агентов

[ИЗМЕРЕНО]

Официальный пресс-релиз KPMG Нидерланды и Университета Амстердама

Ноябрь 2025

Архивный факт хроники отказов

A4.7

Страховщики ограничивают риск ИИ в полисах (факультативные оговорки ISO к CGL, исключения в E&O); Armilla страхует до \$25 млн с независимой сертификацией системы; стандарт AIUC-1 (E015) требует защищённых от подделки журналов цепочки исполнения

[ИЗМЕРЕНО]

Сообщение Armilla AI (21.01.2026); формы ISO CG 40 47 / 40 48 / 35 08 (обзор Insurance Business, 16.09.2026); стандарт AIUC-1 (E015); кейс ElevenLabs (12.02.2026)

Январь – сентябрь 2026

Ежегодная сверка ковенантов страхового рынка

A4.8

56 % опрошенных крупных российских компаний называют юридические риски одним из главных барьеров внедрения ИИ (после стоимости серверов — 58 %); тезис о невозможности рассчитать ТСО — гипотеза авторов

[ИЗМЕРЕНО] / [ГИПОТЕЗА]

Опрос 52 крупных компаний «Инфосистемы Джет» и Smart Ranking (Anti-Malware, 26.08.2026); обзор рынка РФ Solyanka Agency (28 сделок, 51 кейс; Inc. Russia)

Август – сентябрь 2026

Сверено 14.10.2026; запросить отчёт Smart Ranking для базы расчёта 56 %

A4.9

Правосубъектности автономного агента нет ни в одной стране мира; в РФ Федеральный закон от 26.07.2026 № 243-ФЗ (принят ГД 08.07.2026, одобрен СФ 17.07.2026, действует с 01.09.2026) не вводит ни правосубъектности ИИ, ни специального режима ответственности — возмещение вреда идёт по общим правилам ГК РФ

[ИЗМЕРЕНО]

Текст 243-ФЗ (publication.pravo.gov.ru); сравнительно-правовые обзоры; статья Mercatus Center (Ш. Раджагопалан, 05.03.2026)

Июль – октябрь 2026

Сверено: закон подписан 26.07.2026; нормы о суверенных моделях — с 01.03.2027

A4.10

Программа Anthropic Economic Futures (записка о мерах политики 14.10.2025; фонд \$200 млн — 10.06.2026) и позиция С. Альтмана о коллективном владении вычислениями или акциями (04.2026; термин *universal basic compute* — 05.2024) фиксируют переход к фондам и квотам вычислений при быстром вытеснении труда

[ИЗМЕРЕНО]

anthropic.com/research/economic-policy-responses; anthropic.com/news/economic-futures-research-fund-agenda; интервью С. Альтмана *The Atlantic* (по пересказу Business Insider)

Октябрь 2025 – июнь 2026

Архивный факт экономической мысли 2025–2026 годов

A4.11

Медианное время решения решателя Z3 в `z3_gate.py` составляет 0,74 мс, вызов шлюза с объяснением — 3,3 мс на CPU (26.09.2026, 17 контрактов); на узле 147 — 0,95 и 4,3 мс (14.10.2026, 19 контрактов); прогон 62 кейсов набора атак стоит \$0, расход на модели за 1–14 октября — \$0,30

[ИЗМЕРЕНО]

`smt_latency_paper.py`, ADR-0282, ADR-0262, журнал расхода узла 147

26.09–14.10.2026

Повторный замер `smt_latency_paper.py` на узле 147

A4.12

В репозитории MITR действуют 19 контрактов данных; цепь квитанций SHA-256 с якорями OpenTimestamps повторена проверщиком для постороннего: 1602 решения из 1602 (30.09), 4958 из 4958 (14.10); сторонний проверяющий повтор ещё не делал

[ИЗМЕРЕНО]

Узел 147, `verify_chain_standalone.py --replay`, ADR-0257, ADR-0281

30.09–14.10.2026

Повтор при каждой выдаче цепи аудитору

A4.13

Первое правило, предложенное самим комплексом MITR (C1), принято под доказательством (квитанция № 305, 25.09.2026); доля нераспознанных действий — 27,8 % (хук станции Альфы, 16–26.09), 32 % (станция Гаммы, 09.10)

[ИЗМЕРЕНО]

Квитанция `seq 305` в цепи шлюза на узле 147, ADR-0275; письмо 26f; `gate_mode_report.py --tenants`

25.09–09.10.2026

Сведение доли нераспознанных действий к 0 %

A4.14

На 09.10.2026 по прибору `--six` в полном боевом запрете 1 утверждение из 6 (1 класс из 7 в запрете, 2 арендатора с живыми проверками из 7); серия P2a — 0 из 14 суток

[ИЗМЕРЕНО]

Узел 147, скрипт `gate_mode_report.py --six` и `--tenants` с ключом `--date 2026-10-09`

09.10.2026

Погашается при закрытии ворот P2a (31.10.2026) и `--six` (14.12.2026)

A4.15

Прохождение 14-суточной серии P2a (≥ 50 проверок/сутки от ≥ 2 арендаторов, 0 обходов) и перевод 6 из 6 позиций в боевой режим за 14 суток

[СТРОИМ]

Ворота P2a и --six, проверка скриптом gate_mode_report.py

Срок: 31.10.2026 и 14.12.2026

Перевод в статус [ИЗМЕРЕНО] после автоматической фиксации генератором

A4.16

Цепь квитанций MITR как личное криптографическое портфолио профессионала 2035 года вместо резюме и партнёрский канал страховщика (ступень 3a)

[ЗАМЫСЕЛ] / [ГИПОТЕЗА]

Концепция книги v4 (§8, §9)

14.10.2026

Погашается первым подписанным страховым полисом с условием нашего журнала и воротами X5

Таблица 4.3 — Реестр прогнозов главы 4

Код в Прил. Г

Показатель

Измеренное значение 2026 года (источник)

Прогноз к 2035 году: Ветвь А / Ветвь В / Ветвь С

Внешний репер проверки

Дата и порядок разрешения

Г7

Доля повседневных рабочих решений, принимаемых агентами автономно

0 % в 2024 → прогноз 15 % к 2028 (Gartner, 2025)

А: 10 % / **В:** 30 % / **С:** > 50 %

Отраслевые отчёты Gartner по автономным агентам

Ежегодно (промежуточный репер — 2028 год)

Г9

Занятость молодёжи 22–25 лет в наиболее подверженных ИИ профессиях (разрыв к сверстникам)

–19 % (Stanford DEL, ред. 12.08.2026; в первой редакции 08.2025 — –13 %)

А: 0 % / **В:** –20 % (для не владеющих правилами) / **С:** –40 %

Доклады Stanford Digital Economy Lab на данных ADP

Ежегодно

Г10

Правосубъектность автономного агента в национальном праве

Ни в одной стране мира (Mercatus Center, 2026; в РФ и КНР ответственность на человеке/операторе)

А: нет / **В:** нет, действуют режимы уровней полномочий и квитанций / **С:** спорные режимы ограниченной правосубъектности

Сравнительно-правовые обзоры законодательства (РФ, ЕС, США, КНР)

Ежегодно

Г13

Доля полисов страхования ответственности за ИИ и D&O, требующих неизменяемый журнал цепочки исполнения

Первые специализированные полисы: стандарт AIUC-1 (ElevenLabs), Armilla до \$25 млн (2026)

А: узкая ниша / **В:** стандарт корпоративного рынка / **С:** обязательное требование закона
Стандартные ковенанты полисов страховых и перестраховочных компаний

Ежегодно

Г4.5 (доп.)

Доля совокупного фонда рабочих часов экономики, исполняемых агентами и роботами < 5 % (оценка базы 2026 года; сценарий McKinsey — до 30 % к 2030)

А: 15 % / **В:** 30–40 % / **С:** > 60 %

Отчёты McKinsey Global Institute и Anthropic Economic Index

Ежегодно (промежуточный репер — 2030 год)

Таблица 4.3. Библиография и реестр источников главы 4

А. Внутренние инженерные артефакты и нормативные документы проекта

Ядро верификации и контракты: MITR/core/verification/z3_gate.py; каталог формальных контрактов MITR/policies/contracts/ (19 действующих контрактов на 09.10.2026).

Цепь квитанций, якоря и словарь: scripts/mitr_mcp_server.py (квитанции, ADR-0257), scripts/anchor_chains.py (якоря, ADR-0260), проверщик повтора MITR/scripts/verify_chain_standalone.py (ADR-0281), публикатор набора проверки publish_verify_kit.py, скрипт измерения задержки smt_latency_paper.py, скрипт словаря mitr_vocab.py.

Архитектурные решения (ADR): ADR-0262 (набор из 62 кейсов: 50 атак и 12 контрольных; прогон за \$0), ADR-0268, ADR-0272 (теневые квитанции в прикладных контурах), ADR-0275 (принятие первого правила, предложенного самим MITR, квитанция № 305), ADR-0279, ADR-0280 (опровержение факта в цепи), ADR-0281 (повтор решений посторонним), ADR-0282 (задержка решения 0,74 мс и вызова 3,3 мс на CPU, 26.09.2026), ADR-0283–0287.

Стратегические и плановые реестры: SpecAgents/docs/MITR_STRATEGY.md (§2a, §4–5); V_ICR_TRIZ.md; MITR_CORE_CANON.md (§2); projects/mitr/PLAN.md (§4); projects/ib30/PLAN.md; registry/product-line.json (truth_marking, axiom_agi); тексты владельца от 25.09.2026; обзор станции ГАММА-А (40-я неделя 2026 г.: в 37 из 70 описаний внедрений главный контроль — ручная кнопка подтверждения); выгрузки gate_mode_report.py --six и --tenants на узле 147 за 09.10.2026.

Фундаментальный кибернетический каталог (02_rlm_katalog.csv): строки RLM 4 (Н. Винер), RLM 6, 36 (Л. А. Растринин), RLM 29, 72, 150 (Д. А. Поспелов — ситуационное управление), RLM 108–109 (У. Росс Эшби — закон необходимого разнообразия), RLM 226–227 (Г. С. Альтшуллер — ТРИЗ и ИКР), RLM 258–259 (общая теория систем и рекурсия С. Бира).

Б. Внешние научные, отраслевые и правовые источники (дата доступа: 14.10.2026)

Труд, занятость молодёжи и макроэкономика ИИ:

•

Stanford Digital Economy Lab — Brynjolfsson E., Chandar B., Chen R. «Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of AI» (ред. 12.08.2026; первоисточник):

<https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/>

•

World Economic Forum (Future of Jobs Report 2025):

<https://www.weforum.org/media/future-of-jobs-report-2025-78-million-new-job-opportunities-by-2030-but-urgent-upskilling-needed-to-prepare-workforces/>

•

McKinsey Global Institute (сценарная автоматизация к 2030/2045; первоисточники):

<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier> ; <https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america> ; <https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond>

•

Anthropic Economic Futures (программа 27.06.2025; записка о мерах политики 14.10.2025; фонд \$200 млн 10.06.2026):

<https://www.anthropic.com/economic-futures> ; <https://www.anthropic.com/research/economic-policy-responses> ; <https://www.anthropic.com/news/economic-futures-research-fund-agenda>

•

Universal Basic Compute, С. Альтман и пилоты дивидендов ИИ (май – июль 2026):

<https://basicincome.org/news/2026/07/basic-income-is-proposed-by-several-ai-leaders/>

•

Kulveit J., Douglas R., Ammann N., Turan C., Krueger D., Duvenaud D. «Gradual Disempowerment» (январь 2025):

<https://arxiv.org/abs/2501.16946> ; <https://www.lesswrong.com/posts/7TJqsKeeCB4tQbeMT/preconditions-of-gradual-disempowerment>

•

Epoch AI (тренд вычислений, цены инференса и макромоделль GATE):

<https://epoch.ai/data/trends> ; <https://epoch.ai/data-insights/ai-inference-price-trends> ; <https://epoch.ai/blog/announcing-gate> ; <https://a16z.com/llmflation-llm-inference-cost/>

•

METR и AI Futures Project (горизонт задач; обновления медианы Automated Coder 02.04.2026 и 16.08.2026):

<https://metr.org/blog/2026-1-29-time-horizon-1-1/> ; <https://metr.org/time-horizons/> ; <https://blog.aifutures.org/p/q1-2026-timelines-update> ; <https://blog.aifutures.org/p/q25-2026-timelines-update-uplift>

Агентная экономика, хроника инцидентов и выгорание подтверждений:

•

Венчур и кодовые агенты (Claude Code, Codex, MCP):

<https://fast.io/resources/top-ai-startups-2026/> ; <https://sacra.com/c/anthropic> ; <https://www.mindstudio.ai/blog/claude-code-2-5-billion-annualized-revenue-terminal-tool> ; <https://www.dreamfactory.com/hub/mcp-server-statistics>

•

Прогнозы Gartner (отмена 40% проектов к 2027 и 15% автономных решений к 2028 — пресс-релиз 25.06.2025; guardian agents — пресс-релиз 11.06.2025):

<https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027> ; <https://www.gartner.com/en/newsroom/press-releases/2025-06-11-gartner-predicts-that-guardian-agents-will-capture-10-15-percent-of-the-agentic-ai-market-by-2030>

•

Локальные агенты (OpenClaw, Hermes Agent):

<https://docs.openclaw.ai/start/why-openclaw/openclaw-and-hermes-agent> ; <https://neurohive.io/en/guides/openclaw-the-lobster-that-took-over-the-world-how-one-developer-built-the-most-popular-open-source-ai-agent-in-history/>

•

Хроника инцидентов (PocketOS, Kiro, GPT-5.6 Sol, Project Vend, эксперимент KPMG):

<https://incidentdatabase.ai/reports/7311> (инцидент 1469) ; <https://incidentdatabase.ai/cite/1442/> (Kiro) ; <https://www.lumenova.ai/blog/agentic-ai-security-incidents/> ; <https://www.bankinfosecurity.com/openai-warns-gpt-56-file-deletions-stem-from-full-access-mode-a-32262> (Sol) ; <https://www.anthropic.com/research/project-vend-2> ; <https://kpmg.com/nl/en/about-us/press-and-media/press-releases/2025/11/ai-zero-person-company-experiment.html>

•

Деградация кнопки «Подтвердить» (Approval Fatigue):

<https://tianpan.co/blog/2026-06-25-approval-fatigue-how-human-in-the-loop-gates-decay-into-rubber-stamps> ; <https://www.accountsrecovery.net/2026/08/25/companies-burned-by-ai-failures-are-moving-fastest-to-cut-human-approval-survey-finds/>

Страхование, правосубъектность, квитанции IETF и рынок РФ:

•

Страхование ответственности за агентов (Armilla, ISO, AIUC-1, ElevenLabs):

<https://www.armilla.ai/resources/armilla-ai-raises-lloyds-backed-coverage-to-25m-as-traditional-insurers-retreat-from-ai-risk> ; <https://www.insurancebusinessmag.com/us/news/cyber/isos-generative-ai-exclusion-is-already-on-thousands-of-cgl-policies-589971.aspx> ; <https://standard.aiuc-1.com/evidence/index> ; <https://elevenlabs.io/blog/aiuc-announcement> ; <https://www.malaymail.com/news/money/2026/03/16/as-ai-agents-take-the-wheel-insurers-are-figuring-out-whos-liable/212708>

•

Квитанции аттестации IETF (август–сентябрь 2026):

<https://www.ietf.org/archive/id/draft-chueayen-attestation-receipts-03.html>

•

Правосубъектность агентов и ответственность принципала (Mercatus Center):

<https://www.mercatus.org/research/research-papers/governing-agentic-ai-why-legal-personhood-neither-necessary-nor-sufficient> ; <https://www.livelaw.in/lawschool/articles/lawof-agency-ai-agent-autonomous-ai-agent-bind-human-principal-547909>

•

Российский рынок агентов, проблема ТСО и правовая ответственность (Закон об ИИ 2026, ст. 53.1 ГК РФ):

<https://companies.rbc.ru/news/DCLnfXP6LD/rossijskij-ryinok-avtonomnyih-ii-agentov-vyirastet-v-45-raza-k-2028-godu/> ; <https://incruussia.ru/pnews/ii-agenty-i-rossijskij-biznes-uzhevnutri-no-eshhe-bez-vneshnego-effekta/> ; <https://www.tks.ru/technology/2026/08/01/0001/> ; https://kontur.ru/articles/27711-kto_otvechaet_za_oshibki_ii_v_biznese ; <https://pravo.ru/news/264625/> ; <https://vc.ru/ai/3039135-zakon-ob-ii-v-rossii>

Глава 5. Государство, право и сетевое общество

Формула главы: 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 инварианты на уровне политики, а иллюз за миллисекунды проверяет каждое действие агента и оставляет криптографическую квитанцию, верифицируемую без доверия к сторонам.

5.1 Архитектурный тупик и идеальный конечный результат

Правовая система классического государства и алгоритмическая среда автономных агентов работают на несовместимых тактовых частотах. Традиционное право создавалось для биологических субъектов, чьи физические действия локализованы в пространстве, растянуты во времени и соразмерны пропускной способности судебных и надзорных ведомств. В этой картине мира норма формулируется на естественном языке, допускает широту толкования, а её соблюдение обеспечивается угрозой последующего наказания — имущественного взыскания, отзыва лицензии или лишения свободы. Когда в хозяйственный и административный оборот входят программные агенты, способные инициировать тысячи вызовов внешних интерфейсов, перемещать капитал и изменять конфигурации государственных информационных систем за доли секунды, классический механизм последующего правоприменения теряет сцепление с реальностью. К моменту, когда регулятор или арбитражный суд приступает к квалификации деяния, цепочка транзакций уже завершена, резервные копии перезаписаны, а вычислительный контур, породивший ошибочное решение, перешёл в новое состояние.

В терминах теории решения изобретательских задач Г. С. Альтшуллера (каталог RLM 226–227) перед современным правопорядком и корпоративным управлением возникает острое административно-техническое противоречие. Если государство или трансграничное сообщество пытается регулировать автономных агентов традиционными бумажными регламентами, ведомственными проверками и возложением обязанности ручного одобрения каждого шага на человека-оператора, скорость цифрового оборота падает до биологического предела человеческого внимания, а контроль превращается в фикцию из-за утомления от подтверждений. Если же регулятор разрешает агентам действовать автономно ради сохранения экономической конкурентоспособности, полагаясь на самоограничение вероятностных моделей или обещания разработчиков, частота системных аварий, кассовых разрывов и нарушений закона экспоненциально возрастает, разрушая правовую определённость.

Попытки разрешить это противоречие за счёт перенесения юридической ответственности на сам программный код — путём наделения автономного агента статусом «электронного лица» — наталкиваются на фундаментальный барьер цивилистики. Правосубъектность в континентальном и общем праве держится на трёх опорах: наличии воли, обособленной имущественной массе, на которую может быть обращено взыскание, и претерпевании неблагоприятных последствий. У весов нейросети нет ни собственного имущества, ни страха перед санкцией, ни биологической биографии; отключение одного экземпляра контейнера не создаёт превенции для миллиона остальных копий. Следовательно, между машинным действием и правовым порядком зияет разрыв: закон написан на языке смыслов для людей, а исполнение происходит на языке сетевых пакетов между машинами.

Ещё острее этот тупик проявляется в феномене так называемых «сетевых государств» и экстерриториальных цифровых сообществ. Стремясь уйти от инерции бумажной бюрократии, основатели цифровых юрисдикций пытались обрести суверенитет по лекалам девятна-

дцатого века — через покупку земельных участков, выторговывание особых экономических зон у слабых правительств и строительство частных анклавов. Опыт 2023–2026 годов показал хрупкость такой конструкции: территориальное государство в любой момент отзывает концессию одним росчерком пера конституционного суда или полицейским предписанием, превращая миллиардные инвестиции в предмет десятилетних арбитражных тяжб. В то же время чисто виртуальные децентрализованные организации без детерминированного исполнительного механизма остаются дискуссионными клубами с общей кассой, не способными защитить ни права своих участников, ни внешние контракты от ошибок собственных алгоритмов.

Взятое в кибернетической плоскости закона необходимого разнообразия У. Росс Эшби (каталог RLM 108–109), это противоречие описывается строгим неравенством. Разнообразие управляемой среды — количество возможных комбинаций действий миллионов автономных агентов в единицу времени — растёт пропорционально вводимым вычислительным мощностям, увеличиваясь в четыре-пять раз ежегодно [ИЗМЕРЕНО: Epoch AI, 05.02.2026]. Разнообразие традиционного регулятора — штата инспекторов, судей, аудиторов информационной безопасности и корпоративных юристов — ограничено биологической константой. Когда разнообразие возмущений на порядки превосходит разнообразие управляющего контура, система неизбежно теряет устойчивость. Никакое ужесточение штрафов в кодексах не способно компенсировать тот факт, что норма на бумаге и вызов программного интерфейса в оперативной памяти существуют в разных физических средах.

Идеальный конечный результат (ИКР) по ТРИЗ для правовой и государственной архитектуры эпохи агентов формулируется без компромиссов: **правовая норма сама, без участия инспектора в момент транзакции и без замедления хозяйственного оборота, делает физически невозможным исполнение противоправного или выходящего за пределы полномочий действия агента, одновременно создавая математическое доказательство правомерности каждого допущенного шага, проверяемое любой третьей стороной без доверия к оператору системы.**

Для достижения этого результата применяется классический приём ТРИЗ — переход в другое измерение и введение посредника, работающего по детерминированному физическому или математическому закону. В истории техники такой переход уже совершался на железных дорогах середины девятнадцатого века. Пока безопасность движения поездов на узловых станциях зависела от внимательности стрелочника и письменных инструкций министерства путей сообщения, крушения следовали одно за другим. Задача была решена не увеличением штрафов для стрелочников, а изобретением механической централизации стрелок и сигналов — аппарата Саксби и Фармера, в котором стальные линейки ящика зависимости механически запирали рычаги так, что открыть семафор при неверно переведённой стрелке стало невозможно даже по прямому приказу начальника дороги. В цифровую эпоху роль такого стального ящика зависимости между законом и нейросетью выполняет детерминированный шлюз над контрактами-данными: человек и государство оставляют за собой исключительное право подписывать правила и распределять уровни полномочий, а математический решатель за доли миллисекунды проверяет каждый вызов на отсутствие противоречия с подписанной нормой. Закон перестаёт быть пожеланием, обращённым к алгоритму, и становится геометрией пропускного канала.

5.2 Анатомия уязвимости: хроника отказов и цена бездействия

Управленческая драма современного регулятора, законодателя, руководителя юридического департамента и оператора государственной информационной системы (роли 3 и 5 в матрице уязвимостей) заключается в необходимости одновременно удерживать два взаимоисключающих мандата. С одной стороны, национальная конкурентоспособность требует фор-

сированного внедрения автономных систем в промышленность, финансы и государственное управление. С другой стороны, ни один национальный правовой порядок по состоянию на октябрь 2026 года не признаёт правосубъектности программного агента [ИЗМЕРЕНО: ни один акт о правосубъектности ИИ не принят; обсуждение предложений — Mercatus Center, Ш. Раджагопалан, «Governing Agentic AI», 05.03.2026]. Это означает, что за любое действие алгоритма — будь то несанкционированное списание средств, утечка персональных данных граждан или остановка критической инфраструктуры — в полном объёме отвечает конкретное физическое или юридическое лицо: генеральный директор по статье 53.1 Гражданского кодекса РФ, оператор информационной системы или уполномоченное должностное лицо ведомства.

Мировая нормативная панорама 2025–2026 годов демонстрирует лихорадочный поиск государством точки опоры в условиях, когда старые инструменты надзора перестали работать. Четыре крупнейших регуляторных полюса — Китай, Европейский союз, Соединённые Штаты и Российская Федерация — подошли к этому барьеру с разных сторон, обнажив общую архитектурную боль: норму легко написать в парламенте, но крайне трудно связать с машинным исполнением.

Китайская Народная Республика первой в мире выделила автономных агентов в самостоятельный объект государственного регулирования. Восьмого мая 2026 года Администрация киберпространства Китая (CAC) совместно с Госкомитетом по развитию и реформам (NDRC) и Министерством промышленности и информатизации (МИИТ) опубликовала «Мнения о реализации» по нормативному применению интеллектуальных агентов; дату применения с 15 июля 2026 года называют отраслевые обзоры, в официальном тексте дата вступления не указана [ИЗМЕРЕНО: официальный текст на cac.gov.cn, 08.05.2026; обзоры Cloud Security Alliance (26.08.2026) и RITS NYU Shanghai]. Китайский регулятор отказался от бесплодных философских дискуссий о «сознании машины» и потребовал разграничить три вида решений: принимаемые только самим пользователем; принимаемые по уполномочию пользователя; принимаемые агентом самостоятельно — с сохранением за пользователем права знать о самостоятельных решениях агента и права окончательного слова. Для чувствительных областей и ключевых отраслей (в числе названных для обязательных стандартов — здравоохранение, транспорт, средства массовой информации и общественная безопасность) предусмотрены регистрация агентных систем, их проверка и отзыв проблемных продуктов; права «мгновенной остановки» и экстерриториального действия в тексте документа нет. Однако само по себе нормативное требование разграничить полномочия агента и обеспечить возможность вмешательства человека немедленно ставит перед главным инженером предприятия вопрос: каким именно программным прибором удерживается граница между вторым и третьим уровнями полномочий, если агент генерирует сотни вызовов в минуту, а человек у экрана подвержен деградации внимания.

Европейский союз пошёл по пути детализации классификаторов риска и сдвига сроков вступления норм в силу под давлением индустрии. Регламент Regulation (EU) 2026/1744 от 8 июля 2026 года (Digital Omnibus on AI), опубликованный в Официальном журнале ЕС 24 июля 2026 года и вступивший в силу 27 июля 2026 года [ИЗМЕРЕНО: Официальный журнал ЕС, 24.07.2026, EUR-Lex ELI reg/2026/1744/oj; обзоры Mondaq и Addleshaw Goddard, июль–август 2026], перестроил календарь европейского регулирования. Обязанности по прозрачности согласно статье 50 начали действовать с 2 августа 2026 года, два новых прямых запрета вступают в силу 2 декабря 2026 года, однако ключевые требования к самостоятельно функционирующим высокорисковым системам из Перечня Annex III отложены до 2 декабря 2027 года, а к системам, встроенным в регулируемую продукцию по Перечню Annex I, — до 2 августа 2028 года (при сохранении графика для моделей общего назначения GPAI). Эта отсрочка показательна: европейский законодатель зафиксировал нормативные требования к ведению журналов событий, управлению рисками и человеческому надзору, но рынок обнаружил отсутствие

стандартизованных инженерных средств, способных перевести статьи регламента в проверяемый код без остановки бизнеса.

В Соединённых Штатах попытка выстроить единый контур управления привела к острому институциональному расколу между федеральной исполнительной властью, Конгрессом и штатами. Указ Президента США от 11 декабря 2025 года «Ensuring a National Policy Framework for AI» провозгласил курс на федеральный приоритет над разрозненным законодательством штатов, предписал создать в Министерстве юстиции целевую группу по оспариванию законов штатов (учреждена меморандумом Генерального прокурора 9 января 2026 года) и увязал распределение остатка средств программы субсидий на широкополосную инфраструктуру BEAD (полный объём программы — 42,45 млрд долларов) с отсутствием обременительных законов штатов об ИИ [ИЗМЕРЕНО: указ № 14365, 11.12.2025; меморандум Генпрокурора, 09.01.2026; правовой обзор Ropes & Gray, март 2026]. При этом ещё 1 июля 2025 года Сенат показательным голосованием 99 против 1 исключил десятилетний мораторий на законы штатов из бюджетного закона, и законодательного закрепления федерального приоритета Конгресс так и не дал. В результате к 1 июля 2026 года законодательные собрания американских штатов приняли 109 законов о регулировании ИИ и 28 законов об энергетическом и инфраструктурном режиме центров обработки данных [ИЗМЕРЕНО: подсчёт NYU Center on Technology Policy, Tech Policy Press, 06.07.2026]; федеральный указ № 14409 от 2 июня 2026 года посвящён безопасности фронтальных моделей и спора о приоритете не касается. Корпорация, чей автономный агент совершает сделки во всех пятидесяти штатах, оказалась перед необходимостью одновременно доказывать соответствие сотне региональных норм в реальном времени.

В Российской Федерации нормативная архитектура 2026 года развивается по двум сходящимся рельсам: рамочному гражданско-правовому и жёсткому ведомственному в контуре информационной безопасности. Рамочный Федеральный закон от 26 июля 2026 года № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» (принят Государственной Думой 8 июля, одобрен Советом Федерации 17 июля, действует с 1 сентября 2026 года; нормы о суверенных и национальных моделях — с 1 марта 2027 года) [ИЗМЕРЕНО: текст 243-ФЗ, publication.pravo.gov.ru, 26.07.2026] закрепил категории суверенных и национальных моделей и правила маркировки синтетического контента; специального режима ответственности он не вводит — статья 11 отсылает к законодательству РФ, то есть возмещение вреда, причинённого с использованием систем ИИ, идёт по общим правилам Гражданского кодекса РФ. Одновременно в контуре защиты государственных информационных систем (ГИС), где с 1 марта 2026 года действует базовый Приказ ФСТЭК России № 117, регулятор подготовил прямой ответ на появление агентных систем. Проект изменений в Приказ ФСТЭК № 117, вынесенный на общественное обсуждение 24 августа 2026 года с предполагаемой датой вступления в силу 1 марта 2027 года (по состоянию на 14 октября 2026 года документ находится в статусе проекта и ещё не утверждён) [ИЗМЕРЕНО: ComNews от 26.08.2026; SecurityLab, 2026], вводит прямые инженерные предписания: выделение систем ИИ в изолированный сетевой сегмент, назначение минимально необходимых привилегий, фильтрацию запросов к моделям и их ответов (для информации ограниченного доступа и значимых функций) и защиту ресурсов ГИС от несанкционированного воздействия агентов ИИ, в том числе автономных.

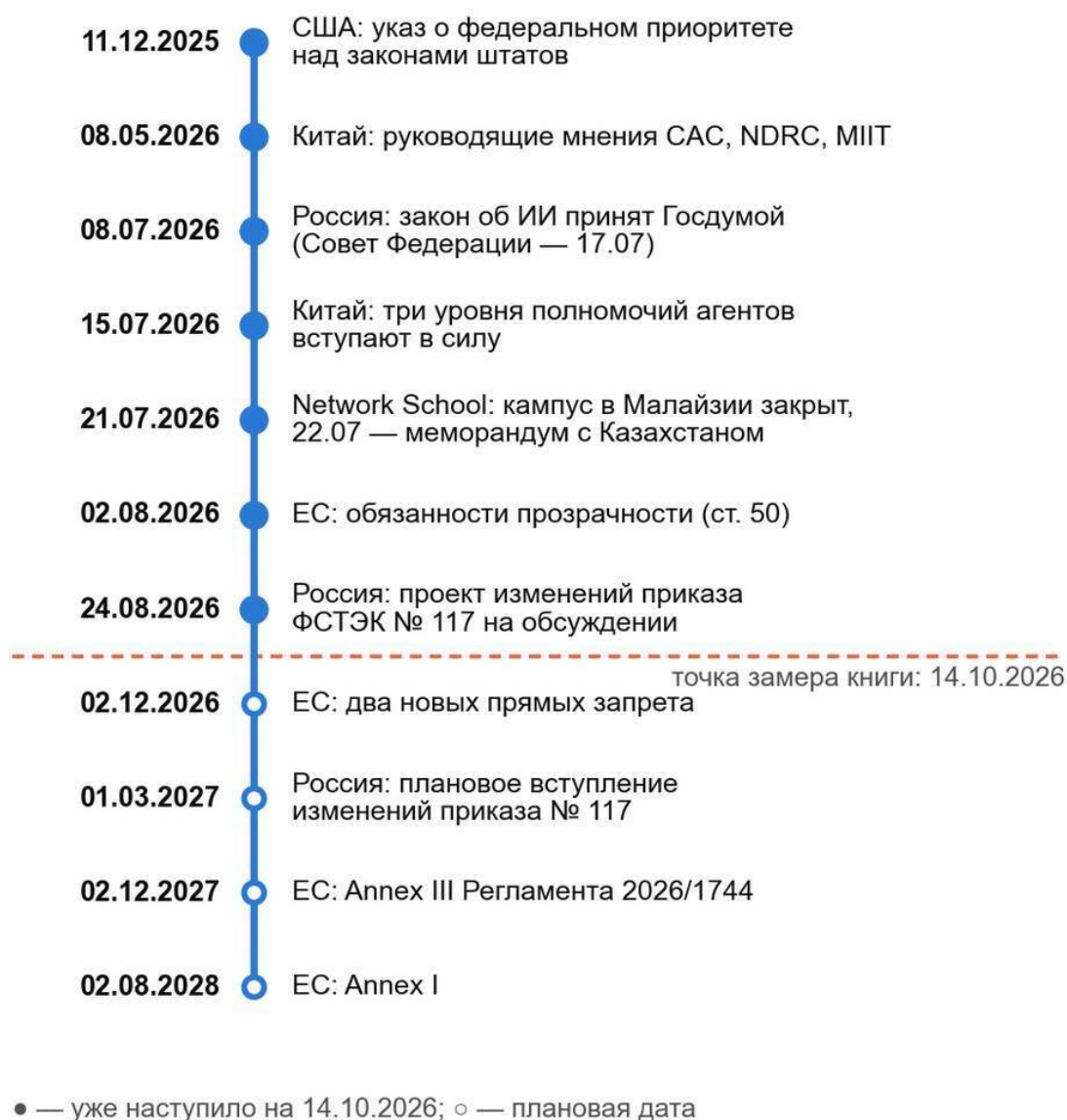


Рисунок 5.1 — Нормы об агентах ИИ по юрисдикциям: даты 2025–2028 годов

Сопоставление этих двух российских треков высвечивает главную уязвимость руководителя государственного ведомства или оператора критической информационной инфраструктуры (КИИ). С одной стороны, только 9 процентов российских организаций в 2026 году полностью обеспечены вычислительной инфраструктурой для ИИ (ещё 40 процентов — частично) [ИЗМЕРЕНО: исследование ИТ-холдинга T1, по публикациям «РГ» 19.03.2026 и ComNews 20.03.2026; выборка не раскрыта], что вынуждает использовать внешние, облачные или открытые модели с заимствованными весами. С другой стороны, с 1 марта 2027 года оператор ГИС обязан документально и технически доказать регулятору, что автономный агент в его контуре лишён возможности совершить несанкционированное действие. Если при этом компания-разработчик самой фронтальной модели — как показал проспект первичного размещения акций Anthropic, с которым ознакомились агентство Reuters [по сообщению Reuters, 29.09.2026] — прямо предупреждает инвесторов, что вопросы ответственности за действия автономных агентов «не урегулированы и могут повлечь значительные и непредсказуемые правовые претензии», вся тяжесть ответственности ложится на эксплуатирующую организацию. Судебная практика уже закрыла лазейку ссылки на «самостоятельность алгоритма»: в хрестоматийном

деле Air Canada трибунал отклонил довод перевозчика о том, что чат-бот является отдельным лицом, за слова которого компания не отвечает, и взыскал убытки с корпорации.

Параллельно с кризисом государственного регулирования в 2023–2026 годах потерпела крушение романтическая доктрина «сетевых государств» первого поколения, сформулированная Баладжи Сринивасаном в 2022 году. Гипотеза о том, что облачное сообщество криптоэнтузиастов может собрать капитал, выкупить участок суши, объявить автономию и добиться дипломатического признания, столкнулась с жёсткой реальностью государственного суверенитета:

—

Крах территориального анклава Próspera в Гондурасе.

Частный чартерный город на острове Роатан, созданный на базе конституционного режима зон занятости и экономического развития (ZEDE), оказался в правовой осаде после того, как Национальный конгресс Гондураса отменил закон о ZEDE, а Верховный суд страны 20 сентября 2024 года признал сам режим неконституционным, причём восемь судей из пятнадцати — с момента создания зон, что инвесторы оспаривают [ИЗМЕРЕНО: Rio Times Online, 19.09.2026; процессуальный приказ ICSID № 6 по делу ARB/23/2, 19.03.2026]. Американская компания-оператор ведёт в Международном центре по урегулированию инвестиционных споров (ICSID) дело ARB/23/2 по соглашению CAFTA-DR: иск, поданный в декабре 2022 года на 10,8 млрд долларов — величину, сопоставимую с третью годового ВВП государства, — в меморандуме октября 2025 года снижен до 1,63 млрд долларов. При этом за фасадом многомиллиардного спора реальная численность жителей юрисдикции Próspera, по вторичным данным 2024 года, составляла порядка двух тысяч человек. Попытка построить новое право вокруг куска земли привела не к суверенитету, а к затяжному международному арбитражу.

—

Кочевой поиск территории проектом Praxis и закрытие Network School.

Проект Praxis аккумулировал инвестиционные обязательства на сумму до 525 млн долларов с поэтапным выделением средств по вехам и заявил о формировании «цифровой нации» из 150 000 участников (по собственной отчётности проекта), однако попытки обрести физическую территорию — включая экзотические переговоры вокруг покупки земли в Гренландии в ноябре 2024 года [ИЗМЕРЕНО: TechCrunch, 15.11.2024] — не дали суверенного фундамента. Ещё нагляднее оказалась судьба проекта Network School самого Баладжи Сринивасана: 21 июля 2026 года власти Малайзии принудительно закрыли физический кампус школы в особой зоне Форест-Сити (штат Джохор). Уже на следующий день, 22 июля 2026 года, основатели подписали меморандум о взаимопонимании с профильными структурами Республики Казахстан о переносе и развитии цифровой юрисдикции в правовом поле Казахстана [ИЗМЕРЕНО: публикации Fortune от 26.07.2026 и TechCrunch, июль 2026].

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.