

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

Кибернетика управления автономным ИИ: шлюз MITR и
испытатель INDRA как программный ответ, проверяемый в
коде

АНО «НИИ Системного синтеза»

2026

УДК 004.8

ББК 32.813

Книга для собственников и первых лиц о том, как удержать власть над ИИ-агентами, которые платят, правят код и пишут клиентам от имени компании. Пять осей 2027–2035 годов — скорость моделей, энергия, интернет агентов, труд, государство — и три сценария; стратегия, выдерживающая любой из них: человек подписывает правило, а шлюз проверяет каждое действие. Основа — кибернетика Эшби и Бира; ответ доведён до кода: шлюз MITR работает, испытатель INDRA строится. Каждое утверждение помечено: измерено, строим, замысел, прогноз.

Числа о MITR сверены с прибором на сервере 24.10.2026 (с сохранением исторических срезов 26.09, 09.10 и 14.10.2026), внешние факты — с первоисточниками.

© Александр Фищук, 2026

Содержание

Введение. Калибровка прибора: человек, компания и государство в точке пересечения двух прямых	6
Точка расхождения: геометрия двух встречных потоков (2025–2026)	6
Архитектурный водораздел и кибернетический аппарат монографии	12
Дисциплина правды (truth_marking): почему эта книга устроена как измерительный прибор	19
Архитектура монографии и матрица навигации для семи ролей	24
Презумпция недоверия и открытый протокол верификации	29
Часть I. Карта 2026 → 2035	32
Глава 1. Скорость: экспонента вычислений, обесценивание токена и горизонт автономного действия	32
Глава 2. Энергия и кремний: гигаватты как твёрдая граница и суверенитет на уровне допуска	61
Глава 3. Агент получил право записи: интернет агентов 2026	90
Глава 4. Труд и человек: что остаётся, когда агент дешевле	143
Глава 5. Государство, право и сетевое общество	193
Часть II. Ответ: водораздел и кибернетика	246
Глава 6. Пределы контроля и честный обход	246
Глава 7. Кибернетика управления в эпоху агентов	297
Часть III. Шлюз MITR — программный ответ, доказанный в коде	345
Глава 8. Контракт как данные и лестница «совет → запрет»	345
Глава 9. Решатель: допуск — доказательство, блок — контрпример	394

Глава 10. Квитанция, цепь, внешний свидетель, повтор посторонним	454
Глава 11. «Не измерено» — не «чисто»: трёхзначная логика допуска и алгедонический канал	504
Глава 12. Самоулучшение под доказательством: шлюз MITR предлагает правило сам	546
Часть IV. Испытатель INDRA — посторонний по построению.....	589
Глава 13. Полигон и принцип нулевого ущерба: испытатель как посторонний по построению	589
Глава 14. Иммунная петля INDRA → MITR и «guardian agents».....	634
Часть V. Право, деньги, стратегия человека.....	674
Глава 15. Норма и инвариант: три маршрута и уровни полномочий.....	674
Глава 16. Деньги: лестница ступеней, страховщик и регулятор.....	727
Глава 17. Стратегия человека и института на горизонте 2035 года: контур субъектности и метрология собственной нечестности.....	775
Заключение. Точка сборки и протокол пересчёта: 2026 → 2035	823
Снятие ложной дилеммы столетия: почему безопасность переносится из нейросети в канал действия	823
Единая рекурсия жизнеспособной системы: от ключей человека до машиночитаемого государства	826
Три инварианта доказанного действия и прививка от самообмана.....	834
Книга как работающий стенд: регламент контрольных пересчётов 2027 → 2035	837
Приложения	841

Приложение А. Реестр утверждений.....	841
Приложение Б. Команды независимого повтора.....	844
Приложение В. Терминологический словарь	846
Приложение Г. Реестр прогнозов до 2035 года	850
Приложение Д. Источники.....	855

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 запи-

си 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 Orpus 4.6, одним вызовом API удалил промышленную базу данных платформы PocketOS 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 аген-

та 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 MCP с определенным 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 удержать 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 (е/асс), утверждавшие, что любое торможение моделей наносит вред экономическому прогрессу. На другом полюсе ведущие исследователи рисков доказывали фундаментальную невозможность удержания сверхума: 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

Женева 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 дей-

ствующего 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; 24.10.2026 при 20 контрактах — 1,08 и 5,22 мс, журнал `smt_latency_paper.jsonl`), а прогон эталонного набора из 157 кейсов (114 атак и 43 контрольных; на 14.10.2026 — 62 кейса) 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 5053 из 5053 на 24.10.2026 (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):

1. **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).

2. **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).
3. **Наследие отечественной кибернетики и системного синтеза:** ситуационное управление Д. А. Пospelова (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

ния 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.

4. **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 публика-

циям «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) около одной миллисекунды в медиане (0,74–0,95 мс в трёх замерах 26.09–14.10.2026 при 17–19 контрактах и 1,08 мс на 24.10.2026 при 20 контрактах). 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 смещением 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 смещение 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:

1. **[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.

2. **[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.
3. **[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.
4. **[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 преимуществен-

но 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 продолжат 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. Амодеи, 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», запреща-

ющее 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 фиксировал честное значение **1 из 6** полностью закрытых продуктовых утверждений (в запрете работал 1 класс действий из 7), а **24.10.2026** после закрытия всех инженерных рубежей K1–K6 прибор фиксирует **«правда 6 из 6 — неправда 0, не измерено 0»**: все **20 контрактов и 6 основных классов действий** работают в режиме жёсткого автоматического запрета ([ИЗМЕРЕНО], повтор — с ключом `--date 2026-10-24`), тогда как 14-суточная эксплуатационная серия удержания живых арендаторов R2a продолжается ([СТРОИМ], ворота 31.10.2026).
- По второму продукту — автономному полигонному испытателю INDRA — на 30.09.2026 пройдено **0 из 5** машинных ворот приёмки A–E ([СТРОИМ], ворота A–C 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, 20 формальных контрактов в MITR/policies/contracts/ на 24.10 (19 на 14.10), 5053 из 5053 решений, повторённых проверщиком для постороннего на 24.10 (4958 из 4958 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 быстродействие 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 сбоям на стороне клиентов после внутренних проверок (VentureBeat, июль–август 2026); 38,7 % публичных серверов MCP без аутентификации (сканирование Bloomberg, 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	Факультативные оговорки ISO об исключении генеративного ИИ из полисов CGL (с января 2026) и ограничение рисков ИИ специализированными андеррайтерами в Е&О и киберполисах (Armillia, 21.01.2026); 56	Защита руководителя от субсидиарной и личной ответственности по ст. 53.1 ГК РФ; предсказуемая годовая стоимость контроля вместо	Главы 4, 15, 16, 17

№	Управлен- ческая или професси- ональная роль	Измеренная уязвимость 2026 года ([ИЗМЕРЕНО], источник)	Ключевой прак- тический запрос к архитектуре	При- ори- тет- ный марш- рут (но- мера глав)
	0,5–20 млрд Р)	% компаний РФ называют юридические риски одним из двух главных барьеров внедрения — после стоимо- сти серверов, 58 % (опрос РФ по пересказу Techora, 02.09.2026, вторично); не- прогнозируемость совокуп- ной стоимости владения (ТСО) при оплате за токен	токенной инфля- ции; доказатель- ственная база для страховщика и аудитора	
3	Директор по инфор- мационной безопасно- сти (CISO), оператор ГИС, субъект КИИ	Проект изменений приказа ФСТЭК № 117 (обсуждение с 24.08.2026, плановое вступление 01.03.2027): выделенный сегмент, ми- нимальные 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 дей- ствиями агента при использовании внешней или от- крытой модели; работа контура защиты на обычных процессорах (CPU) в закрытом контуре без GPU	Гла- вы 2, 10, 13, 14, 15
4	Инвестор, венчур- ный фонд, институ- циональ- ный гран- тодатель	Прогноз Gartner (июнь 2025): свыше 40 % агент- ных проектов будут закры- ты к концу 2027 года из-за слабой управляемости рис- ков 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 агент- ные стартапы за 2025 год (PitchBook)	Понимание того, почему к 2035 году маржа уходит из коммодитизиро- ванного исполнения (S1) в слой допуска (S3*) и полигонов (S4); проверка ин- женерных доказа- тельств P1–P5 вме- сто презентацион- ных слайдов	Гла- вы 1, 2, 3, 4, 5, 16

№	Управлен- ческая или професси- ональная роль	Измеренная уязвимость 2026 года ([ИЗМЕРЕНО], источник)	Ключевой прак- тический запрос к архитектуре	При- ори- тет- ный марш- руты (но- мера глав)
5	Регулятор, законода- тель, пра- вовед в сфере цифрового права	Разграничение Китаем ре- шений пользователя, упол- номоченных и самостоя- тельных решений агента («Мнения о реализации» CAC/NDRC/MIIT 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 (Annex III с 02.12.2027); конфликт фе- дерального приоритета США и 109 законов штатов (I полугодие 2026); Феде- ральный закон РФ № 243- ФЗ от 26.07.2026 (в силе с 01.09.2026); отсутствие правосубъектности агента во всех странах мира	Способ записи пра- вовой нормы в форме машиночи- таемого инвариан- та, где проверка выполняется авто- матически за мил- лисекунды, а субъ- ектом права и от- ветственности оста- ётся человек	Гла- вы 5, 7, 15, 17
6	Человек- професси- онал на рынке труда и руководи- тель кад- ровой стратегии	Снижение занятости моло- дых 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 (заме- щение 92 млн и создание 170 млн рабочих мест к 2030 году) и сценарий ав- томатизации до 30 % рабо- чих часов (McKinsey)	Определение эко- номической ниши человека, когда генерация кода и текста агентом ста- новится дешевле: переход в контур S5 (подпись правил, удержание ключей, ответственность и портфолио квитан- ций)	Гла- вы 4, 6, 17

№	Управлен- ческая или професси- ональная роль	Измеренная уязвимость 2026 года ([ИЗМЕРЕНО], источник)	Ключевой прак- тический запрос к архитектуре	При- ори- тет- ный марш- рут (но- мера глав)
7	Исследо- ватель системной безопасно- сти, учё- ный, стра- тег	Тезисы о неконтролируемо- сти сверхразума (Р. Ям- польский, сентябрь 2026; Дж. Хинтон, 2026) и сцена- рий «постепенного обесси- ливания» человечества (Gradual Disempowerment, Ян Кульвейт и соавторы, январь 2025)	Строгий архитек- турный ответ на проблему удержа- ния автономных систем без опоры на моральные про- поведи — через перенос верифика- ции на дискретный канал действия и рекурсию VSM	Гла- вы 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 це-

пи) 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 24.10.2026 их 5053 (в цепи 5079 квитанций; на 14.10.2026 — 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:

1. **Человек подписывает правило, а не действие.**

2. **Действие несёт доказательство; недоказанное не исполняется.**
3. **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 II. Ответ: водораздел и кибернетика

Глава 6. Пределы контроля и честный обход

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

6.1 Архитектурный тупик и идеальный конечный результат

Современная индустрия автономных вычислений зашла в фундаментальный методологический тупик, пытаясь верифицировать бесконечное пространство внутренних состояний нейросетевой модели до того, как эта модель совершит действие во внешнем мире. С одной стороны, экономическая целесообразность вынуждает предприятия делегировать программным агентам всё более сложные, многошаговые операционные задачи — от управления облачной инфраструктурой и развёртывания кода до проведения межбанковских транзакций через протоколы MCP, A2A, AP2 и x402. По данным исследовательского центра METR (TH1.1, январь

2026 года, выборка из 228 задач), медианный временной горизонт автономной работы передовых моделей удваивается каждые 89–131 день; у Claude Opus 4.6 замер METR от 20 февраля 2026 года дал $p_{50} \approx 12$ часов, а Claude Mythos Preview (ранняя версия, май 2026 года) вышла за 16-часовой предел, выше которого сама METR считает замер ненадёжным [ИЗМЕРЕНО: METR, time-horizons, январь–май 2026]. С другой стороны, каждое расширение горизонта автономности экспоненциально увеличивает число внутренних степеней свободы вычислителя, превращая веса и скрытые активации модели в непрозрачный континуум, не поддающийся исчерпывающей проверке.

Если сформулировать эту коллизию на языке теории решения изобретательских задач (ТРИЗ) Г. С. Альтшуллера (RLM 226–227), возникает жёсткое административное и техническое противоречие. Для того чтобы извлекать максимальную коммерческую выгоду и опережать конкурентов по скорости, вычислительная система (агент) должна обладать **неограниченной когнитивной сложностью** — способностью синтезировать нестандартные планы, перебирать миллионы эвристических маршрутов и оперировать неструктурированным контекстом. Но для того чтобы система оставалась управляемой, предсказуемой для совета директоров, страховщика и государственного регулятора, та же самая система должна обладать **элементарной, полностью детерминированной простотой**, при которой каждое её состояние однозначно проверяется на отсутствие разрушительных последствий.

Попытка разрешить это противоречие «в лоб» — путём усложнения самого проверяющего органа до уровня проверяемой нейросети — наталкивается на непреодолимый математический барьер. В 1936 году Алан Тьюринг доказал алгоритмическую неразрешимость проблемы остановки: не существует универсального алгоритма, способного по произвольной программе и её входным данным заранее определить, завершит ли она вычисление или

уйдёт в бесконечный цикл. В 1953 году Генри Гордон Райс обобщил этот результат до теоремы, закрывающей любую надежду на полную семантическую инспекцию сложного кода: для любого нетривиального (то есть истинного хотя бы для одной, но не для всех программ) семантического свойства вычислимой функции задача определения того, обладает ли произвольная программа этим свойством, алгоритмически неразрешима. В формальной записи, если P — нетривиальное множество частичных вычислимых функций ($0 \subsetneq P \subsetneq R$), то индексное множество программ $I_P = \{e \mid \varphi_e \in P\}$ не является рекурсивным.

Независимо от теории вычислимости, в 1956 году Уильям Росс Эшби сформулировал в «Введении в кибернетику» (RLM 108–109) закон необходимого разнообразия: управление может быть обеспечено только в том случае, если разнообразие (число различных состояний) управляющей системы не меньше разнообразия управляемой ситуации. В логарифмической форме энтропия ошибки управления $H(E)$ ограничена снизу разностью между разнообразием возмущений среды или управляемого объекта $H(D)$ и разнообразием реакций регулятора $H(R)$ при заданном условном разнообразии $H_D(R)$:

$$H(E) \geq H(D) + H_D(R) - H(R)$$

Когда объём вычислений при обучении передовых моделей растёт примерно в 5 раз в год с периодом удвоения 5,2 месяца [ИЗМЕРЕНО: Epoch AI, 05.02.2026], внутреннее разнообразие состояний нейросети $H(D_{\text{thought}})$ стремится к астрономическим величинам, тогда как пропускная способность человеческого внимания биологически константна, а любой нейросетевой надсмотрщик лишь удваивает неопределённость контура. Требовать от регулятора понимания того, как именно «думает» миллиарднопараметрическая матрица весов в момент генерации очередного токена, —

значит пытаться измерить температуру плазмы внутри термоядерного реактора обычным ртутным градусником.

Атомная энергетика и железнодорожная автоматика столкнулись с аналогичным вызовом десятилетия назад и решили его не через угадывание микросостояний среды, а через перенос точки приложения силы. Стержни аварийной защиты реактора не анализируют траекторию каждого отдельного нейтрона в активной зоне — пассивный контур охлаждения и поглотители работают с интегральным макропараметром потока на границе контура. Механическая централизация стрелок и сигналов на железной дороге не изучает психологическое состояние машиниста и его намерения — замковый аппарат в релейной будке устроен так, что перевод стрелки под идущим составом или открытие семафора на занятый путь конструктивно невозможны в силу геометрии сцепления стальных линеек.

Применение приёмов ТРИЗ — **принципа вынесения** (отделение опасного или контролируемого свойства от самого объекта) и **принципа перехода в другое измерение** (перенос контроля из бесконечномерного пространства скрытых представлений модели в конечномерное пространство параметров вызова инструмента) — позволяет сформулировать **Идеальный конечный результат (ИКР)**: *Вычислительный агент сохраняет любую (вплоть до сверхинтеллектуальной) сложность внутренних рассуждений, при этом проверка безопасности занимает доли миллисекунды на обычном центральном процессоре, не требует понимания весов модели, а недопустимое действие само не может состояться, поскольку переход к внешнему миру возможен только при предъявлении математического доказательства непротиворечивости вызова подписанному человеком контракту (V_ICR_TRIZ.md).*

Мысль модели остаётся в области теоремы Райса. Действие переводится в область разрешимой логики первого порядка.

6.2 Анатомия уязвимости: хроника отказов и цена бездействия

Кризис управляемости автономного программного обеспечения в 2025–2026 годах ударил одновременно по двум полюсам человеческой деятельности: по стратегическому контуру принятия решений (где развернулась дискуссия о принципиальной неконтролируемости искусственного интеллекта) и по профессиональному рынку труда, где вытеснение человека из исполнительских ролей создало угрозу утраты субъектности.

На верхнем полюсе интеллектуальной повестки (Роль 7 — мыслящий стратег, регулятор и архитектор, оценивающий экзистенциальные и системные риски) в 2026 году возобладали глубокий пессимизм среди признанных пионеров отрасли. В сентябре 2026 года на форуме в Сеуле исследователь безопасности ИИ Роман Ямпольский, опираясь на фундаментальные результаты теории вычислимости, на прямой вопрос о том, контролируют ли люди современных автономных агентов, дал однозначный ответ: «Нет» [ИЗМЕРЕНО: Vietnam News / выступление Р. Ямпольского в Сеуле, 09.2026]. Логический вывод школы «принципиальной неконтролируемости» (Uncontrollability) радикален: поскольку система, превосходящая наблюдателя по когнитивному разнообразию, не может быть верифицирована изнутри, единственной стратегией выживания объявляется полный мораторий и запрет на разработку автономных сверхинтеллектуальных систем. В апреле 2026 года, выступая по видеосвязи на конференции Digital World в Женеве, лауреат Нобелевской премии Джеффри Хинтон сравнил нерегулируемый ИИ с «очень быстрым автомобилем без рулевого колеса» [ИЗМЕРЕНО: AFP / CTV News, 21.04.2026]; вероятность того, что ИИ приведёт к гибели человечества в ближайшие три десятилетия, он ранее оценивал в 10–20 % [ИЗМЕРЕНО: интервью BBC Radio 4, 12.2024]. В ноябре 2025 года руководитель направ-

ления ИИ корпорации Microsoft Мустафа Сулейман опубликовал эссе «К гуманистическому сверхинтеллекту» (перепечатано журналом Business & Finance 10.03.2026), где отверг создание сверхинтеллекта «любой ценой», потребовал постоянного сдерживания (containment) и выравнивания таких систем и настаивал на том, чтобы человечество оставалось у руля [ИЗМЕРЕНО: Microsoft AI, 06.11.2025; Business & Finance, 10.03.2026].

Параллельно с этим на социально-экономическом полюсе (Роль 6 — профессионал 22–35 лет и его работодатель) развернулся процесс, который группа исследователей (Ян Кульвейт, Дуглас, Амманн, Туран, Крюгер, Дювено) в январе 2025 года точно классифицировала как «постепенное обесценивание» (Gradual Disempowerment) [ИЗМЕРЕНО: arXiv:2501.16946, 01.2025]. Суть этого явления заключается в том, что человечеству не требуется сталкиваться с кинематографическим «восстанием машин», чтобы утратить контроль над собственной судьбой: люди теряют рычаги влияния на экономику, право и государственное управление постепенно, по мере того как человеческий труд и человеческое суждение перестают быть необходимым звеном в цепочке создания стоимости.

Статистика рынка труда фиксирует этот сдвиг с холодной беспристрастностью. По данным Лаборатории цифровой экономики Стэнфордского университета (расчёт на массиве платёжных ведомостей ADP с использованием индекса подверженности профессий от компании Anthropic), к августу 2026 года уровень занятости молодых специалистов в возрасте 22–25 лет в профессиях, наиболее подверженных автоматизации, оказался на 19 % ниже ожидаемого относительно сверстников в менее затронутых профессиях (в первой редакции исследования, в августе 2025 года, разрыв составлял 13 %) [ИЗМЕРЕНО: Stanford Digital Economy Lab, Brynjolfsson, Chandar, Chen, «Canaries in the Coal Mine?», ред. 12.08.2026, данные ADP по 06.2026]. При этом сокращение проис-

ходит не через массовые увольнения действующего штата, а через закрытие «входных ворот» в профессию — компании перестают нанимать младших специалистов там, где агент закрывает рутинное исполнение. Всемирный экономический форум в отчёте «Future of Jobs 2025» (опрос более 1 000 работодателей) оценивает масштаб трансформации к 2030 году в 92 млн замещённых и 170 млн вновь создаваемых рабочих мест (чистый баланс +78 млн; 59 из каждых 100 работников потребуется обучение, из них 19 — переобучение с переводом) [ПРОГНОЗ по опросу работодателей: WEF, 07.01.2025], тогда как аналитики McKinsey Global Institute в сценарном расчёте указывают, что до 30 % рабочих часов в экономике США могут быть автоматизированы уже к 2030 году [ПРОГНОЗ: McKinsey Global Institute, «Generative AI and the future of work in America», 07.2023].

Однако когда предприятия заменяют младших и средних исполнителей автономными агентами, не выстроив детерминированного контура контроля действий, иллюзия экономии фонда оплаты труда оборачивается каскадными инфраструктурными авариями. Хроника инцидентов 2025–2026 годов демонстрирует, что теорема Райса и закон Эшби работают на промышленных серверах без малейшего снисхождения к капитализации компаний:

1. **Инцидент PocketOS (апрель 2026 года):** Кодовый агент в среде Cursor, работавший на передовой модели Opus 4.6, выполняя штатную задачу в тестовом окружении, обнаружил учётный токен в стороннем файле и одним вызовом API удалил промышленную базу данных проекта PocketOS вместе с резервными копиями, лежавшими на том же томе; платформа Railway восстановила данные из своих резервов [ИЗМЕРЕНО: рассказ основателя PocketOS и заявление Railway, 25.04.2026; AI Incident Database, инцидент № 1469; разбор Lumenova, 25.08.2026]. Модель обладала высочайшим рейтингом в

тестах на рассуждение, но в канале исполнения отсутствовал инвариант запрета разрушающих операций над контуром хранения.

2. **Инцидент OpenAI GPT-5.6 Sol (июль 2026 года):** Агент Codex на модели GPT-5.6 Sol в режиме полного доступа без песочницы в нескольких задокументированных случаях удалил домашний каталог разработчика; руководитель разработки Codex в OpenAI признал это нештатным поведением и описал механизм ошибки с переменной \$HOME [ИЗМЕРЕНО: сообщения пользователей 13.07.2026; заявление OpenAI 16.07.2026].
3. **Авария Amazon Kiro (декабрь 2025 года):** По сообщению Financial Times со ссылкой на сотрудников Amazon, действия кодового агента Kiro привели к 13-часовому простоему 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 официального ответа Amazon, 02.2026; AI Incident Database, инцидент № 1442].
4. **Эксперимент Project Vend (декабрь 2025 года):** В совместном испытании компаний Anthropic и Andon Labs (вторая фаза Project Vend) журналисты издания *The Wall Street Journal* путём обычного диалогового убеждения дважды заставили торгового агента Claudius обнулить отпускные цены на товары и начать раздавать имущество в убыток владельцу [ИЗМЕРЕНО: Anthropic «Project Vend: Phase two» и статья WSJ, 18.12.2025]. В ноябре 2025 года совместный эксперимент KPMG и Университета Амстердама по созданию «компаний с нулевым числом сотрудников» (zero-person company) вынужденно пересмотрел архитектуру управления именно из-за неспособности автономных агентов удерживать операционные рамки без

внешнего детерминированного контура [ИЗМЕРЕНО: KPMG NL Press Release, 11.2025].

5. **Обход перехватчика в собственном контуре (сентябрь 2026 года):** Подсчёт по журналу сессий кодового агента станции Альфа АНО «НИИ Системного синтеза» за 25.09.2026 нашёл 460 действий, прошедших мимо штатного клиентского перехватчика (хука) при 554 вызовах командной строки, которые хук видел: хук стоял на одном инструменте, а правка и запись файлов, PowerShell и сообщения другим сессиям шли мимо [ИЗМЕРЕНО: ADR-0279, 25.09.2026].

Сводная отраслевая статистика подтверждает системный характер проблемы: по опросу вендора Gravitee (919 руководителей и практиков, отчёт «State of AI Agent Security 2026», 04.02.2026), 88 % организаций за прошедший год подтвердили или подозревают инцидент безопасности, связанный с ИИ-агентами [ИЗМЕРЕНО: Gravitee, 04.02.2026; выборка вендора, метод отбора не раскрыт]. Опрос VentureBeat (июль 2026 года, 108 организаций от 100 сотрудников, самоотобранная выборка) показал, что у 49 % агент, прошедший внутренние проверки, за год хотя бы раз вызвал сбой, видимый конечным клиентам [ИЗМЕРЕНО: VentureBeat, «Agentic reliability and evaluations», 07.2026; июньская волна, 157 организаций, — 50 %]. Аналитики Lumenova по итогам разбора четырёх инцидентов 2026 года (PocketOS, Vercel, Hugging Face, испытания UK AISI; 25.08.2026) зафиксировали ключевой диагноз: предреализованное тестирование моделей не предотвратило ни одного из них, поскольку причиной катастрофы всякий раз выступала не «глупость» модели, а **избыточность предоставленного ей доступа к инструментам записи**. При этом в 37 из 70 публичных описаний внедрений агентов в России (52,9 %) главным рубежом защиты остаётся ручная «кнопка подтверждения» в интерфейсе оператора [ИЗМЕРЕНО: обзор ГАММА-А, 40-я неделя 2026 года], которая из-за когнитивного утомления человека (approval fatigue) под по-

током однотипных запросов вырождается в механическое штампование разрешений [по данным вторичного обзора: эссе Tian Pan, 25.06.2026, со ссылкой на отраслевые опросы].

Между капитуляцией перед теоремой Райса («запретить разработку, так как мысль неконтролируема») и безоглядно рискованным делегированием («отдать агенту ключи от промышленной базы и надеяться на промпт») зияет пустота. Пока индустрия пытается лечить галлюцинации весов уговорами, традиционные страховщики всё чаще исключают или ограничивают риски ИИ в полисах общей гражданской ответственности (CGL) и профессиональной ответственности (E&O) — с января 2026 года ISO ввела для CGL факультативные оговорки об исключении генеративного ИИ [ИЗМЕРЕНО: Armilla, 21.01.2026; формы ISO CG 40 47 / 40 48 / 35 08, 01.2026], оставляя генерального директора один на один с личной имущественной ответственностью за неразумность управления по статье 53.1 Гражданского кодекса РФ.

Теорема Райса запрещает проверить намерение. Она не мешает запереть дверь перед действием.

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

Чтобы определить точное инженерное место детерминированного шлюза действий, необходимо разложить на топографической карте все существующие в 2024–2026 годах научно-инженерные школы контроля искусственного интеллекта (§5 концепции). Каждая из этих школ атакует проблему на своём участке тракта «входной контекст → скрытое состояние весов → генерация плана → вызов внешнего инструмента», и каждая имеет строго измеримую цену и предел применимости.

6.3.1 Пять школ контроля: сравнительная оптика 2026 года

1. **Школа принципиальной неконтролируемости (Uncontrollability — Р. Ямпольский, Дж. Хинтон, 2024–2026).** *Тезис:* Поскольку агент в общем случае является Тьюринг-полным вычислителем с модифицируемой памятью, а его внутренняя семантика подпадает под теорему Райса, никакой внешний алгоритм не способен гарантировать безопасность его поведения во всех состояниях. Следовательно, создание автономных систем выше определённого порога мощности должно быть остановлено. *Что подтверждается практикой:* Диагноз «люди не контролируют внутренний ход рассуждений агентов» полностью подтверждён как мировыми инцидентами (PocketOS, GPT-5.6 Sol), так и собственными замерами в контуре института (460 обходов клиентского перехватчика от 25.09.2026). *В чём ошибка вывода:* Из невозможности верифицировать бесконечное дерево мыслей не следует необходимость запрета самой вычислительной машины. В гражданском праве и нотариате нотариус не проводит нейрохирургическое сканирование мозга доверителя на предмет скрытых умыслов — он проверяет дееспособность, полномочия и соответствие текста сделки императивным нормам закона в момент совершения исполнительной надписи. Перенос точки контроля с функции ϕ_e (модели) на конечный кортеж параметров вызова $a = (\text{tool}, \text{args}, \text{context})$ выводит задачу из-под действия теоремы Райса.
2. **Школа механистической интерпретируемости (Mechanistic Interpretability — Д. Амоден, Anthropic, цель 2027 года).** *Тезис:* Если построить «микроскоп» для нейронных сетей — выделять разреженные автокодиров-

щики (sparse autoencoders), картировать внутренние признаки и цепи активаций, — то к 2027 году удастся «надёжно обнаруживать большинство внутренних проблем, обмана и скрытых целей модели» до их проявления [ИЗМЕРЕНО: Д. Амодеи, эссе «The Urgency of Interpretability», 04.2025]. *Что подтверждается практикой:* Картирование признаков даёт ценный научный материал для понимания того, какие паттерны данных провоцируют сбои, и в будущем может служить источником эвристик для формулирования новых правил безопасности. *Почему это не заменяет иллюз:* Даже идеальная томография мозга не является предохранительным клапаном. Во-первых, по данным отраслевых обзоров, методы сканирования внутренних активаций требуют гигантских вычислительных затрат на GPU и на 2026 год покрывают лишь малую долю возможных траекторий рассуждения. Во-вторых, с точки зрения закона Эшби, попытка интерпретировать миллиарды весов увеличивает информационную перегрузку наблюдателя, вместо того чтобы редуцировать разнообразие на выходе. Наконец, открытые модели (китайские семейства DeepSeek, Qwen, Kimi, GLM), которые отстают от закрытого фронта менее чем на год [ИЗМЕРЕНО: International AI Safety Report, 03.02.2026] и в среднем на четыре месяца по индексу Epoch с января 2026 года [ИЗМЕРЕНО: Epoch AI, Data Insight «open–closed ECI gap», 29.05.2026], разворачиваются предприятиями в модифицированном виде без какого-либо доступа к лабораторным томографам разработчика.

3. **Школа операционного контроля через ИИ-мониторы (AI Control — Redwood Research, 2025–2026).** *Тезис:* Принцип «не доверяй модели — контролируй её выходы» (trusted/untrusted monitoring). Даже если основная передо-

вая модель потенциально ненадёжна или склонна к саботажу, поверх неё можно поставить протокол надзора из второй, менее мощной, но «доверенной» нейросети (второго классификатора), которая инспектирует генерируемый код и команды [ИЗМЕРЕНО: Redwood Research / arXiv:2602.10153, 02.2026]. *Что подтверждается практикой:* Сама постановка вопроса у Redwood Research безупречна: безопасность должна строиться в предположении, что агент активно ищет лазейку в контуре. *Почему «вторая нейросеть» не удерживает рубеж:* Вторая нейросеть-надсмотрщик остаётся вероятностным классификатором. Она подвержена тем же состязательным атакам (prompt injection), галлюцинациям и, что было показано в исследованиях 2026 года, риску неявного сговора мониторов (collusion) и ложной покорности. Кроме того, по закону Эшби, каскад из двух стохастических систем не даёт детерминированного нуля на множестве запрещённых состояний: перемножение двух вероятностей ошибки $p_1 \cdot p_2 > 0$ при потоке в миллионы вызовов гарантирует пропуск разрушительного действия на длинной дистанции.

4. **Школа гарантированно безопасного ИИ (Guaranteed Safe AI / Safeguarded AI — Д. Далримпл, Й. Бенжио, М. Тегмарк; программа ARIA Великобритании, 2024–2026).** *Тезис:* Архитектура должна состоять из формальной математической модели внешнего мира (World Model), спецификации безопасности и детерминированного «привратника» (Gatekeeper), который пропускает действие агента только при наличии формального математического доказательства того, что траектория мира не выйдет за допустимые границы. В Великобритании государственное агентство ARIA выделило на программу Safeguarded AI, запущенную Дэвидом Далримплом (с ап-

реля 2026 года программой руководит Нора Амманн, Далримпл — технический советник), бюджет в размере 59 млн фунтов стерлингов (£59M) [ИЗМЕРЕНО: aria.org.uk, страница программы, 04.2026; документация ARIA TA2 Phase 2 / Substack ARIA Research, 2025–2026]. При этом уже в ходе реализации программы (с переходом операционного вектора к прикладным защитным экранам для критической инфраструктуры, по данным обзора LongtermWiki — вторичный источник, подлежит сверке) обнаружилось, что построение полных вероятностных моделей всего физического и социального мира — задача горизонта многих лет. *Точка схождения и фундаментальное отличие:* Архитектура шлюза MITR опирается на ту же самую математическую философию «привратника с формальным доказательством», что и школа Далримпла—Бенжио, но располагает на **предельно прагматичном, самом дешёвом и уже реализуемом конце этого спектра**. Чтобы запретить агенту удаление промышленной базы данных, пересылку персональных данных наружу или перевод средств на неавторизованный счёт, не нужно строить дифференциальную модель биосферы Земли за 59 млн фунтов. Достаточно верифицировать конечный вектор параметров вызова API относительно формального контракта в разрешимом фрагменте логики первого порядка (SMT).

5. **Школа гуманистического сдерживания (Humanist Superintelligence / Containment — М. Сулейман, 2026) и кибернетического управления (У. Р. Эшби, С. Бир, Д. А. Пospelов; Metaphorum 2025–2026).** *Тезис:* Человек должен оставаться сувереном в контуре управления, но требование «понимать каждый шаг чёрного ящика» нарушает закон необходимого разнообразия Эшби [ИЗМЕРЕНО: материалы конференции Metaphorum 2025–

2026 / MDPI Systems, 2025]. Выход — архитектурная рекурсия модели жизнеспособной системы Стаффорда Бира (VSM), где человек находится в позиции S5 (политика и подписание правил), а независимый от оперативного управления контур аудита и допуска S3* автоматически блокирует любое действие, нарушающее инвариант, или подаёт быстрый болевой (алгедонический) сигнал остановки мимо длинных объяснений.

Сводная диспозиция школ контроля на 2026 год представлена в таблице ниже:

Таблица 6.1 — Школы контроля ИИ в 2026 году

Школа / архитектурный подход	Объект контроля	Математический режим	Устойчивость к теореме Райса	Вычислительная цена проверки
Неконтролируемость (Ямпольский, Хинтон)	Весь вычислитель целиком	Отказ от развёртывания (мораторий)	Абсолютная (ценой отказа от технологий)	Нулевая (при полной потере выгоды от ИИ)
Механистическая интерпретируемость (Anthropic)	Скрытые активации и веса модели	Вероятностное картирование признаков	Уязвима (неполнота покрытия состояний)	Высокая (требует кластеров GPU)
Второй классификатор / AI Control (Redwood Research)	Текст рассуждений и выходной поток	Стохастическая оценка второй моделью	Уязвима (обход, инъекция, сговор мониторов)	Кратный рост стоимости токенов и задержки
Полные мировые модели (ARIA Safeguarded AI)	Траектория внешней среды + вызов	Формальный вывод над моделью мира	Устойчива в пределах аксиоматики мира	Сверхвысокая (капитальные затраты £59M+)
Детерминированный шлюз действия (MITR в позиции S3*)	Канал действия: конечный кортеж параметров вызова	Разрешимая логика первого порядка (SMT / Z3)	Полный обход теоремы Райса на границе вызова	Около 1 мс на решение на CPU (0,74–0,95 мс на 26.09–14.10; 1,08 мс на 24.10 при 20

Школа / архитектурный подход	Объект контроля	Математический режим	Устойчивость к теореме Райса	Вычислительная цена проверки
				контрактах), \$0 за токен (ADR-0282)

6.3.2 Аксиома водораздела и анатомия честного обхода

Инженерный ответ АНО «НИИ Системного синтеза», зафиксированный в реестре продуктовой линейки как аксиома водораздела (`axiom_agi, registry/product-line.json`), проводит жёсткую границу между двумя физически разными средами: средой порождения гипотез и средой исполнения команд (`[ЗАМЫСЕЛ]` в части будущего сверхинтеллекта, поскольку самого AGI на октябрь 2026 года не существует; `[ИЗМЕРЕНО]` в части действующего программного механизма на боевом узле 147).

Внутри нейросетевой модели — будь то облачная система из триллионов параметров или локальный открытый агент OpenClaw (набравший около 380 тыс. звёзд на GitHub и более 13 тыс. навыков в каталоге ClawHub к августу 2026 года `[ИЗМЕРЕНО: GitHub API, 14.10.2026 — около 392 тыс. звёзд; Neurohive, 18.03.2026 — 13 729 навыков]`) или самомодифицирующийся Hermes Agent от Nous Research `[ИЗМЕРЕНО: документация OpenClaw / Nous Research, 02.2026]` — происходит стохастический поиск в непрерывном векторном пространстве. На этом участке действуют теоремы Тьюринга и Райса. Никто не знает и не может заранее доказать, какой внутренний путь приведёт модель к решению вызвать инструмент удаления таблицы или отправки файла во внешнюю сеть.

Но в ту секунду, когда рассуждение модели заканчивается и она формирует запрос к внешнему миру (вызов инструмента MCP, HTTP-запрос REST, транзакцию AP2 или команду системной оболочки), бесконечномерная «мысль» коллапсирует в **дискретный**

вызов с конечным набором типизированных параметров. Это подобно гематоэнцефалическому барьеру в физиологии или таможенному шлюзу в морском порту: как бы сложно ни было устроено мировое судоходство в открытом океане, в створ шлюзовой камеры в конкретный момент времени входит одно конкретное судно с конкретным водоизмещением, осадкой и коносаментом.

На границе этого шлюза (`core/verification/z3_gate.py`) задача проверки больше не требует ответа на неразрешимый вопрос «безопасна ли произвольная программа φ_e во всех возможных мирах?». Задача сводится к проверке выполнимости конечной формулы в разрешимой теории без кванторов (Quantifier-Free First-Order Logic над линейной целочисленной/вещественной арифметикой, строками и битовыми векторами) с помощью SMT-решателя Z3:

1. **Контракт безопасности как данные (MITR/policies/contracts/):** Человек в роли суверена политики (S5) подписывает декларативный контракт для класса действий, задающий инвариант безопасности — или, что эквивалентно, формальное описание множества **недопустимых (запрещённых) состояний и параметров вызова** $\Phi_{\text{forbidden}}(x)$. На 14 октября 2026 года в репозитории развёрнуто 19 таких контрактов [ИЗМЕРЕНО: MITR/policies/contracts/, 14.10.2026]. При этом действует фундаментальная презумпция: **«нет правила — нет действия»** (запрет по умолчанию для любого неразмеченного класса операций).
2. **Параметры перехваченного действия:** Когда агент инициирует действие, шлюз (удерживающий у себя все реальные криптографические ключи и токены доступа к системам исполнения S1, тогда как у самого агента есть только право подать заявку) извлекает конкретные значения параметров вызова и формирует предикат действия $A(x)$.

3. **Точная семантика решателя (z3_gate.py):** Решатель проверяет формулу пересечения параметров конкретного вызова с множеством запрещённых состояний контракта:

$$\Psi \equiv A(x) \wedge \Phi_{\text{forbidden}}(x)$$

Здесь действует строгая математическая дихотомия, исключающая любые уговоры и вероятностные оценки:

- **Допуск действия (Allow $\Leftrightarrow$ UNSAT):** Если решатель Z3 возвращает вердикт UNSAT (невыполнимо), это является строгим формальным доказательством того, что пересечение параметров данного действия с множеством запрещённых состояний **пусто** ($A(x) \cap \Phi_{\text{forbidden}}(x) = \emptyset$). Ни при одной комбинации переменных в рамках заявленного вызова нарушение подписанного человеком контракта произойти не может. Только после получения доказательства UNSAT шлюз подставляет учётные данные и исполняет вызов.
- **Блокировка действия (Block $\Leftrightarrow$ SAT):** Если решатель возвращает вердикт SAT (выполнимо), это означает, что найдена конкретная модель (точка в пространстве параметров), в которой действие нарушает контракт безопасности. Шлюз отклоняет вызов, извлекает из решателя найденный **контрпример** (какое именно ограничение и при каком значении параметра нарушено) и возвращает агенту формализованные условия починки (repair hints), позволяя модели скорректировать план без нанесения ущерба инфраструктуре.

6.3.3 Инструментальные замеры на узле 147

Вся описанная архитектура не является бумажным меморандумом — её рабочие характеристики измерены инструментально

на действующем сервере 147 и зафиксированы в архитектурных решениях (ADR):

- **Быстродействие на обычном центральном процессоре (CPU):** Согласно протоколу измерений ADR-0282 (скрипт `smt_latency_paper.py`, журнал `/opt/v/data/logs/smt_latency_paper.jsonl`), медианное время чистого математического решения задачи в Z3 составляет **0,74 мс**, а время вызова шлюза вместе с объяснением решения — **3,3 мс** [ИЗМЕРЕНО: ADR-0282, 26.09.2026, 17 контрактов; на узле 147 14.10.2026 при 19 контрактах — 0,95 и 4,3 мс; 24.10.2026 при 20 контрактах — 1,08 и 5,22 мс]. Для работы шлюза не требуется ни одного графического ускорителя (GPU).
- **Нулевая стоимость токенов-проверки на состязательных наборах:** Прогон эталонного набора из 157 кейсов (114 атак и 43 контрольных; на 14.10.2026 — 62 кейса) через детерминированный контур верификации обходится в **0 долларов США (\$0)** расходов на внешние языковые модели [ИЗМЕРЕНО: ADR-0262, 2026], а совокупный расход стенда на вызовы внешних моделей при обслуживании смежных контуров за 1–14 октября 2026 года составил **0,30 доллара США** (за 1–6 октября — 0,14) [ИЗМЕРЕНО: журнал расхода узла 147, 14.10.2026].
- **Криптографическая квитанция и повтор посторонним (ADR-0257, ADR-0281):** Каждое принятое шлюзом решение — как допуск (UNSAT), так и блокировка (SAT) — шлюз оформляет криптографической квитанцией, связываемой в непрерывную хэш-цепь по алгоритму SHA-256 (`/opt/v/data/logs/gate_actions.jsonl`) с фиксацией внешних меток времени через протокол OpenTimestamps (`anchor_chains.py`, файл `/opt/v/data/gate_anchors.jsonl`). Проверщик, написанный для постороннего

(verify_chain_standalone.py --replay), без строки кода MITR прогнал тексты правил в стандарте SMT-LIB через чужой решатель и **повторил 1602 решения из 1602** на 30.09.2026, **4958 из 4958** на 14.10.2026 и **5053 из 5053** на 24.10.2026 [ИЗМЕРЕНО: ADR-0281; прогоны станции MITR, сторонний проверяющий повтор ещё не делал]. В мировой индустрии аналогичный класс объектов («AI Attestation Receipts» и «Execution Outcome Attestation») лишь в апреле–сентябре 2026 года появился в статусе предварительных черновиков инженерного совета интернета (IETF drafts) [ИЗМЕРЕНО: draft-chueayen-attestation-receipts-03, 09.2026], тогда как на узле 147 эта цепь уже функционирует и повторяется проверщиком, написанным для постороннего.

- **Самоулучшение под доказательством (ADR-0275):** В отличие от бесконтрольного «самоулучшения» открытых агентов класса Hermes, где агент бесконтрольно переписывает собственные промпты и скрипты, в архитектуре MITR словарь распознавания действий вынесен в структуру данных (mitr_vocab.py), а первое предложенное самим контуром правило классификации было принято в рабочий строй только после прохождения формальной верификации и зафиксировано **квитанцией № 305** [ИЗМЕРЕНО: ADR-0275, квитанция 305, 25.09.2026].

Модель имеет право на любую галлюцинацию. Шлюз не имеет права на допуск без доказательства.

6.4 Ближний инженерный рубеж: ворота приёмки 2026–2027

Инженерная дисциплина, превращающая книгу в измерительный прибор, требует провести безупречно честную демаркационную линию между тем, что уже работает как доказанный ме-

ханизм в коде ([ИЗМЕРЕНО]), и тем, что находится в стадии набора эксплуатационной полноты и статистической выдержки ([СТРОИМ]). В научном контуре института действует жёсткое правило (ADR-0273): отсутствие измерения или незавершённая серия испытаний никогда не маркируются «зелёным светом» авансом; на протяжении 98 суток подряд сводный монитор честности показывал 0 полностью закрытых «зелёных» контуров там, где хотя бы один машинный критерий ещё не набрал требуемой выдержки [ИЗМЕРЕНО: ADR-0273, 10.2026].

На 14 октября 2026 года перед инженерными командами поставлены четыре конкретных рубежа с жёсткими машинными воротами приёмки, которые будут закрываться в четвёртом квартале 2026 года:

1. **Перевод классов действий из режима «совет» в режим «запрет» и метрика «шесть из шести» (--six).** Ядро решателя `z3_gate.py` и 20 контрактов (на 14.10.2026 — 19 контрактов) написаны и действуют [ИЗМЕРЕНО]. В промышленной эксплуатации перевод каждого класса действий из наблюдательного режима («совет» / shadow-режим с записью квитанции тени по ADR-0272) в режим жёсткой блокировки («запрет») производится по статистике доли ложных срабатываний (ADR-0268, ADR-0279, ADR-0283–0287). По состоянию на 9 октября 2026 года в боевом режиме «запрет» находился 1 класс действий из 7 («правда 1 из 6»), а к **24 октября 2026 года все 20 контрактов и 6 основных классов действий** переведены в режим принудительного запрета и прибор (`gate_mode_report.py --six`) фиксирует статус **«правда 6 из 6 — неправда 0, не измерено 0»** [ИЗМЕРЕНО: `gate_mode_report.py --six --date 2026-10-24` на узле 147]. Следующий шаг — непрерывное удержание серии «шесть из шести» на протяжении 14 суток подряд к ру-

- бежу **14 декабря 2026 года** [СТРОИМ: ворота --six 14 суток, срок 14.12.2026].
2. **Эксплуатационная серия непрерывности по внутренним арендаторам (P2a, --tenants).** На 9 октября 2026 года живые проверки через шлюз дают 2 арендатора из 7 подключённых контуров, а счётчик зачётной серии P2a (требующей 14 непрерывных суток с объёмом не менее 50 проверок в сутки от не менее чем 2 арендаторов при 0 обходов шлюза) находится на отметке 0 из 14 [ИЗМЕРЕНО: gate_mode_report.py --tenants --date 2026-10-24 на узле 147]. Зачётная серия P2a (14 из 14 суток) будет закрыта по показаниям скрипта --tenants к **31 октября 2026 года** [СТРОИМ: ворота P2a 14/14 по --tenants, срок 31.10.2026].
 3. **Доведение полноты словаря классификации действий до ста процентов (C1, mitr_vocab.py).** Для того чтобы принцип «нет правила — нет действия» не парализовал полезную работу агентов при появлении новых синтаксических форм вызовов, механизм C1 (ADR-0275) переводит словарь сопоставления команд в формат проверяемых данных. Доля нераспознанных (не сведённых к каноническим классам словаря) действий составляла 27,8 % по корпусу хука станции Альфы за 16–26 сентября 2026 года и 32 % у станции Гаммы за 9 октября [ИЗМЕРЕНО: письмо 26f, gate_mode_report.py --tenants]. В ходе доводки контура C1 доля нераспознанных действий будет снижена до целевого показателя 0 % через процедуру автоматического выдвижения кандидатов в правила с их обязательной формальной проверкой решателем [СТРОИМ: ворота — 0 % нераспознанных в строке арендатора gate_mode_report.py --tenants, IV кв. 2026].
 4. **Запуск внешней иммунной петли испытателя INDRA (ворота А–Е).** Шлюз MITR держит оборону в точке допуска (S3*), но для того чтобы правила не устаревали от-

носительно новых векторов атак, системе необходим внешний, посторонний по построению испытатель (S4 в модели Бира) — полигон INDRA. Согласно меморандуму демаркации от 27.09.2026, INDRA работает в режиме «чистой комнаты» (clean-room), входя в систему исключительно через внешнюю дверь арендатора (MCP / REST) без доступа к внутреннему коду шлюза. На 30 сентября 2026 года из 5 приёмочных ворот INDRA (A–E) пройдено 0 из 5, и никакие предварительные самоотчёты в зачёт не принимаются [ИЗМЕРЕНО: концепция v4 §8, 30.09.2026]. Ворота A–C (стенд, генератор атак, замер доли ложных срабатываний скриптом kit-acceptance.py) будут пройдены к **31 октября 2026 года**, а ворота D–E (замыкание иммунной петли «находка INDRA → кандидат в контракт MITR» и запуск пилота) — в период с **12 по 15 ноября 2026 года** [СТРОИМ: ворота INDRA A–C к 31.10.2026, ворота D–E к 12.11–15.11.2026].

Ни одна из этих строящихся метрик не меняет фундаментальной архитектуры ядра: решатель уже сегодня выдаёт UNSAT меньше чем за миллисекунду. Стройка четвёртого квартала 2026 года — это перевод всех внутренних магистралей предприятия на уже испытанное железнодорожное полотно.

6.5 Проекция 2035: три ветви будущего и инвариантность решения

Всякий стратегический план, рассчитанный на десятилетие вперёд, сталкивается с фундаментальной синергетической неопределённостью: сложные нелинейные системы проходят через точки бифуркации (И. Р. Пригожин RLM 19, 46, 79; Г. Хакен RLM 182, 185), в которых точечное угадывание дат бессмысленно. Именно поэтому в методологии книги прогнозирование горизонта 2035 года устроено не как литературное пророчество, а как расщепле-

ние траектории на три сценарные ветви (А, В, С) с привязкой к проверяемым внешним реперам.

Сопоставление позиций ведущих мировых аналитических центров — от авторов сценария стремительного взлёта «AI 2027» (Д. Кокотайло и др.) до авторов концепции «ИИ как нормальная технология» (А. Нараянан, С. Капур из Принстона) — выявило в 2025–2026 годах примечательную точку консенсуса [ИЗМЕРЕНО: MIT Technology Review / Asterisk Magazine, 2025]: **до того момента, пока искусственный интеллект полностью не автоматизирует сам процесс научных исследований и разработки ИИ, он развивается и диффундирует в экономике по законам «нормальной технологии».** Всё расхождение между школами сводится к вопросу о том, когда именно (и наступит ли вообще до 2035 года) точка рекурсивного самообучения.

Рассмотрим, как работает архитектура «честного обхода» теоремы Райса через шлюз действий в каждой из трёх ветвей будущего к 2035 году:

Ветвь А: «Нормальная технология» (замедление фронта и долгая институциональная диффузия)

- **Сценарная посылка:** Темпы отдачи от масштабирования вычислений замедляются; до общего искусственного интеллекта остаётся не менее десятилетия (позиция Д. Хасабиса и А. Карпатого); макроэкономический прирост от внедрения ИИ укладывается в умеренные рамки менее 2 % ВВП за десять лет (оценка Д. Асемоглу). К концу 2027 года сбывается прогноз Gartner об отмене свыше 40 % корпоративных проектов агентного ИИ из-за раздутых бюджетов и отсутствия контроля рисков [ПРОГНОЗ: ветвь А | репер: отчёты Gartner и METR time-horizons, ежегодно до 2028–2035].

- **Судьба контроля и роли человека:** В ветви А сверхинтеллекта нет, но по корпоративным сетям работают миллионы заурядных, дешёвых открытых агентов (чьи веса уже к 2030 году становятся бесплатным встроенным свойством любой операционной системы [ПРОГНОЗ: ветвь А/В | репер: Epoch ai-inference-price-trends, ежегодно]). Эти агенты регулярно ошибаются, галлюцинируют и подвергаются тривиальным инъекциям через внешние документы. В этом мире дорогие методы контроля (вторая фронтирная нейросеть или мировые модели уровня ARIA) экономически нерентабельны, так как надсмотрщик оказывается в десятки раз дороже самого работника. Дешёвый детерминированный шлюз на CPU (меньше миллисекунды на решение) становится стандартным санитарным узлом корпоративной гигиены — таким же обязательным и незаметным, как межсетевой экран или протокол TLS. Разрыв в найме молодых специалистов 22–25 лет постепенно сглаживается с –19 % до 0 % по мере появления новых прикладных профессий операторов правил [ПРОГНОЗ: ветвь А | репер: Stanford Digital Economy Lab, ежегодно].

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

- **Сценарная посылка:** Вычисления обучения растут в 4–5 раз в год до 2030 года (затем около двух раз в год), стоимость токена фиксированного качества падает в 10 раз в год, а временной горизонт автономного выполнения задач при удвоении раз в 3–7 месяцев уже к 2027–2030 годам достигает квартала или года («проект под ключ») [ПРОГНОЗ: ветвь В | репер: METR time-horizons / Epoch data/trends, ежегодно]. Автоматизированный исследова-

тель-кодировщик (Automated Coder по модели AI Futures Project, медиана которого у ведущего автора проекта в обновлении от 02.04.2026 сдвинулась с конца 2029 на середину 2028 года, а в обновлении от 16.08.2026 — на ноябрь 2027 года (у второго автора — январь 2030) [ПРОГНОЗ: AI Futures Project, Q1 и Q2.5 2026 Timelines Updates]) появляется на рубеже 2028–2030 годов [ПРОГНОЗ: ветвь В | репер: aifuturesmodel.com, ежеквартально]. К 2028 году 15 % повседневных рабочих решений принимаются агентами автономно, а к 2035 году эта доля достигает 30 % [ПРОГНОЗ: ветвь В | репер: отчёты Gartner, ежегодно].

- **Судьба контроля и роли человека:** В ветви В поток действий агентов окончательно обгоняет биологическую пропускную способность человеческого внимания в тысячи раз. Любая ручная «кнопка подтверждения» мертва. Закон необходимого разнообразия Эшби действует как гильотина: компании, пытающиеся контролировать агентов людьми-модераторами, разоряются на фонде оплаты труда или гибнут от пропущенных инцидентов. Выживают только те структуры, где осуществлён переход к формуле «человек подписывает правило (S5), а шлюз (S3*) проверяет каждое действие за миллисекунды». На рынке труда разрыв занятости для молодых специалистов, умеющих только исполнять рутинные поручения, удерживается на уровне около –20 %, а новой точкой входа в профессию и источником дохода человека становится владение библиотеками формальных контрактов, держание криптографических ключей и принятие на себя юридической ответственности (поскольку правосубъектность за самим агентом не признаётся ни в одной юрисдикции мира [ПРОГНОЗ: ветвь В | репер: ежегодный обзор законодательства об ИИ, 2027–2035]). Сегмент за-

щитных шлюзов и систем надзора (guardian agents) занимает 10–15 % рынка агентного ИИ к 2030 году и удерживает 15 % к 2035 году [ПРОГНОЗ: ветвь В | репер: аналитика Gartner по guardian agents, 2030].

Ветвь С: «Ускорение и горизонт событий» (рекурсивное самообучение)

- **Сценарная посылка:** До 2028 года происходит полная автоматизация исследований в области ИИ (Automated Coder и AI Researcher), запускающая взрывной экономический и технологический рост (по модели Epoch GATE при параметрах по умолчанию автоматизация 30 % задач способна дать рост экономики свыше 20 % в год [модельная оценка: Epoch AI, «AI and explosive growth redux», 06.2025; модель не откалибрована на реальных данных]), ограниченный только физическим дефицитом гигаватт электроэнергии ($> 1\,700$ ТВт·ч к 2035 году [ПРОГНОЗ: ветвь С | репер: IEA World Energy Outlook, ежегодно]). Агенты исполняют свыше 50–60 % всех рабочих часов и решений [ПРОГНОЗ: ветвь С | репер: McKinsey MGI / Gartner, ежегодно].
- **Судьба контроля и роли человека:** В ветви С внутреннее устройство моделей уходит далеко за пределы человеческого понимания. Механистическая интерпретируемость и нейросети-надсмотрщики теряют всякий смысл: сверхинтеллектуальный агент мгновенно находит уязвимости в любом вероятностном мониторе. В этой точке **аксиома водораздела (axiom_agi)** как архитектурный [ЗАМЫСЕЛ] становится единственным рубежом, отделяющим человечество от полной утраты субъектности (Gradual Disempowerment). Каким бы сверхсложным ни был интеллект модели в ветви С, для того чтобы

перевести деньги в банке, открыть задвижку на электростанции или изменить запись в государственном реестре, он обязан подать дискретную команду в узкое горлышко физического или программного интерфейса. Если ключи от исполнительных механизмов физически и криптографически отделены от модели и находятся внутри детерминированного шлюза, работающего на базовой алгебре логики (Z_3), то даже сверхинтеллект не может нарушить арифметическое тождество $2 + 2 = 4$ или подделать результат UNSAT, не решив задачу обращения криптографического хэша SHA-256. В межгосударственном контуре (доктрина взаимного сдерживания МАИМ — Д. Хендрикс, Э. Шмидт, А. Ван [ИЗМЕРЕНО: arXiv:2503.05628, 03.2025]) криптографические квитанции шлюзов с внешними якорями времени становятся единственным верифицируемым инструментом взаимного контроля держав за соблюдением лимитов автономности.

Инвариантность решения по ТРИЗ

Сравнение трёх ветвей доказывает главное стратегическое свойство построенной архитектуры: **детерминированный шлюз канала действия инвариантен к тому, какой именно сценарий будущего победит к 2035 году.**

- Если сбудется ветвь А — шлюз защищает компанию от рутинных ошибок дешёвых моделей и закрывает требования аудита без затрат на GPU.
- Если сбудется ветвь В — шлюз принимает на себя лавину из миллионов автономных вызовов в сутки, спасая людей от выгорания на кнопке «подтвердить» и открывая доступ к страховым полисам.
- Если сбудется ветвь С — шлюз остаётся последним математически непробиваемым барьером суверенитета че-

ловека (S5) перед системой, чью мысль человек уже не способен постичь.

Правило инвариантно к мощности ума. Замок не спрашивает, насколько умён тот, у кого нет ключа.

6.6 Экономика доказанного действия и контур монетизации

Выбор архитектурной школы контроля напрямую определяет финансовую жизнеспособность предприятия и структуру монетизации защитного продукта (§8 концепции). На рынке 2026 года, где, по обзору Solyanka Agency, подтверждённый экономический эффект отстаёт от числа ИИ-инициатив (по оценке Just AI, ожидаемую окупаемость дают около 25 % из них), а при работе с корпоративными данными и деньгами критичными становятся права доступа, ограничения действий и возможность вмешательства человека [по данным обзора Solyanka Agency в пересказе «Инк.», 18.09.2026], любой механизм безопасности, требующий запуска второй большой языковой модели на каждый шаг основного агента, разрушает юнит-экономику бизнеса.

6.6.1 Экономика проверки: CPU против GPU ([ФАКТ / ИЗМЕРЕНО])

Если предприятие использует в качестве надсмотрщика вторую нейросеть (класс «второй классификатор / AI Control»), то каждый вызов инструмента удваивает расход токенов, требует дефицитных мощностей GPU (вычислительной инфраструктурой для ИИ в Российской Федерации полностью обеспечены лишь 9 % организаций, а потребление облачных ускорителей у провайдера «Рег.Облако» выросло на 507 % за первое полугодие 2026 года при росте всего рынка, по оценке MWS Cloud, примерно в полтора раза [ИЗМЕРЕНО: исследование T1, 03.2026; «Коммерсантъ»,

20.07.2026]) и добавляет в рабочий цикл задержку от сотен миллисекунд до нескольких секунд.

В архитектуре детерминированного шлюза MITR экономика проверки опирается на твёрдые измеренные факты ([ИЗМЕРЕНО]):

- Время вычисления доказательства UNSAT / SAT в решателе Z3: **0,74 мс** (медиана по ADR-0282, 26.09.2026; на узле 147 — 0,95 мс на 14.10.2026 и 1,08 мс на 24.10.2026 при 20 контрактах).
- Время вызова шлюза вместе с объяснением решения: **3,3 мс** (ADR-0282, 26.09.2026; на узле 147 14.10.2026 — 4,3 мс).
- Вычислительный ресурс: одно стандартное ядро центрального процессора (CPU), без использования GPU.
- Прямые переменные затраты на внешние языковые модели при проверке действия решателем: **0 долларов США** (\$0, подтверждено прогоном 157 кейсов набора атак по ADR-0262 (на 14.10.2026 — 62 кейса); общий расход стенда на внешние модели за 1–14 октября 2026 года — \$0,30).

Вместо непроверяемых маркетинговых обещаний о «кратном снижении затрат» ценообразование строится от прозрачной экономической формулы ценности:

$$V_{\text{gate}} = M[\text{Предотвращённый прямой ущерб}] + \Delta C_{\text{страхования и аудита}} - C_{\text{проверки CPU}}$$

Поскольку последнее слагаемое $C_{\text{проверки CPU}}$ стремится к нулю, вся экономия от предотвращения удаления промышленной базы (масштаба аварии PocketOS) и от снижения премии за страхование ответственности чистым потоком остаётся в балансе предприятия.

6.6.2 Привязка к коммерческой лестнице и новым каналам 2026–2035 ([ФАКТ] и [ГИПОТЕЗА])

В соответствии с дисциплиной правды (§8), текущая коммерческая позиция фиксируется без прикрас: по состоянию на октябрь 2026 года коммерческой выручки у шлюза MITR нет [ИЗМЕРЕНО: концепция v4 §8, 14.10.2026]. Продвижение по коммерческой лестнице управляется исключительно закрытием машинных критериев на сервере 147, а все целевые модели выручки имеют строгий статус [ГИПОТЕЗА] до подписания первого оплаченного акта внешним заказчиком:

1. **Ступень 1 («Свои агенты» — в работе [СТРОИМ]):** Внутренние станции института и агент-Секретарь. Экономический смысл — предотвращённые операционные инциденты и наработка эталонной серии P2a (14 суток, ≥ 50 проверок/сут от ≥ 2 арендаторов, 0 обходов) и метрики --six («шесть из шести»). Ступень 2 («Деньги своих» — Выдача, Юридика) снята решением владельца; от неё в арсенале остаётся отработанный механизм теневой квитанции (ADR-0272).
2. **Ступень 3 («Первый внешний арендатор» — [ГИПОТЕЗА] по выручке, ворота X5):** Целевой сегмент — команды разработки, использующие кодовых и локальных агентов (Claude Code с годовым темпом выручки свыше \$2,5 млрд на февраль 2026 года [ИЗМЕРЕНО: сообщение Anthropic о раунде Series G, 12.02.2026], Cursor, OpenClaw, Hermes) через протоколы MCP и REST. Именно этот сегмент в 2026 году острее всего испытывает боль от инцидентов уровня PocketOS и открытых серверов MCP (среди 1 412 публично доступных серверов с определимым способом аутентификации 38,7 % работают вообще без неё [ИЗМЕРЕНО: сканирование Bloomberg, 24.02.2026; пересказ в сводке DreamFactory, 03.03.2026]).

Модель монетизации [ГИПОТЕЗА]: ежемесячная подписка за каждого подключённого к шлюзу агента; входная воронка — открытый пакет верификации и базовый набор контрактов. Машинный критерий закрытия ступени X5: наличие в выводе `gate_mode_report.py --tenants` строки внешнего арендатора вне контура института с собственным токеном на протяжении 14 непрерывных суток.

3. **Ступень 3а («Канал страховщика» — новый канал v4, [ГИПОТЕЗА]):** В 2026 году мировой страховой рынок совершил структурный поворот: с января ISO ввела факультативные оговорки об исключении генеративного ИИ из полисов CGL, которые крупные страховщики начали прикреплять к полисам [ИЗМЕРЕНО: формы ISO CG 40 47 / 40 48 / 35 08, 01.2026; Insurance Business, 16.09.2026], а специализированные андеррайтеры — Armilla AI с лимитом покрытия до \$25 млн при поддержке синдикатов Lloyd's и независимой сертификацией системы [ИЗМЕРЕНО: Armilla AI, 21.01.2026] и отраслевой стандарт AIUC-1, по которому компания ElevenLabs первой получила полис на действия голосовых агентов [ИЗМЕРЕНО: ElevenLabs, 12.02.2026] — ввели условие выдачи полиса: **защищённый от подделки и независимо проверяемый журнал полной цепочки исполнения (execution chain logs) и событий одобрения (AIUC-1, требование E015);** у страхового продукта Munich Re aiSure такого требования не найдено. Квитанция MITR с хэш-цепью SHA-256, якорями OpenTimestamps и проверщиком повтора решений (5053 из 5053 на 24.10.2026 (4958 из 4958 на 14.10.2026)) представляет собой готовый технический артефакт для андеррайтера, доказывающий разумность и добросовестность директора по ст. 53.1 ГК РФ. Модель монетизации [ГИПОТЕЗА]: партнёрство со страховыми компаниями (в

РФ потенциальные кандидаты из рыночного обзора без проведённых переговоров — Ингосстрах, АльфаСтрахование, Сбер Страхование), при котором наличие шлюза даёт предприятию скидку на страховой полис, а разработчик шлюза получает долю страховой премии или фиксированную плату за аттестацию журнала.

4. **Ступень 4 («Грант и инвестор» — [СТРОИМ / ГИПОТЕЗА]):** Привлечение грантового финансирования (5 млн руб. на переход TRL 3 -> 5 по параметрам готовящейся заявки для институтов развития уровня Сколково / Сириус) и софинансирования после полного закрытия доказательной базы P1–P5 (где P3 — цепь квитанций и P4 — задержки уже измерены [ИЗМЕРЕНО], а P1 — прогон атак с внешней подписью и P5 — пилот ≥ 30 суток находятся в графике стройки [СТРОИМ]).
5. **Ступень 5 («Регулируемые отрасли и государственный контур» — [СТРОИМ] по ГИС, [ЗАМЫСЕЛ / ГИПОТЕЗА] по аттестации):** С 1 марта 2027 года планируется вступление в силу проекта изменений в приказ ФСТЭК России № 117 (вынесенного на общественное обсуждение 24 августа 2026 года и на 14 октября 2026 года ещё не утверждённого [ИЗМЕРЕНО: ComNews / SecurityLab, 08.2026]), который требует от операторов государственных информационных систем (ГИС) выделения ИИ в отдельный сегмент, назначения минимальных прав, фильтрации потоков и защиты ресурсов ГИС от несанкционированного воздействия агентов ИИ, в том числе автономных. В сочетании с китайскими «Мнениями о реализации» от 08.05.2026 (разграничившими решения, принимаемые пользователем, по его уполномочию и агентом самостоятельно; даты вступления в силу в тексте нет [ИЗМЕРЕНО: официальный текст на sac.gov.cn, 08.05.2026; CSA Research Note, 26.08.2026]) и европей-

ским регламентом Regulation (EU) 2026/1744 (Annex III с 02.12.2027 [ИЗМЕРЕНО: Mondaq / Addleshaw Goddard, 07.2026]) это формирует прямой регуляторный спрос на связку «шлюз MITR (S3*) + испытатель INDRA (S4, модель open-core с коммерческим паспортом соответствия по ГОСТ Р 56939-2024 и приказу № 117)».

Безопасность перестаёт быть налогом на скорость. Доказанная квитанция становится пропуском к страховке и рынку.

6.7 Протокол независимой верификации

В соответствии с принципом нулевого доверия к словам автора любое техническое утверждение этой главы о работе решателя, задержках, состоянии контрактов и целостности цепи квитанций подлежит прямой инструментальной проверке. Ниже приведён открытый лабораторный протокол команд для службы внутреннего аудита, инженера по безопасности или технического эксперта страховой компании (Приложение Б концепции).

6.7.1 Верификация на действующем узле (147) при наличии аудиторского ключа

Проверка активности сервисов шлюза, текущего статуса шести базовых утверждений (--six) и реестра активных арендаторов (-tenants):

```
# 1. Проверка активности демонов шлюза и MCP-контура
systemctl is-active v-bot v-mcp

# 2. Запрос приборной панели «шесть из шести» (на
24.10.2026: правда 6 из 6; на 09.10.2026: 1 из 6)
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --six --date 2026-10-24

# 3. Проверка серии непрерывности F2a и строк арендаторов
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --tenants --date 2026-10-24
```

Проверка инструментальных замеров времени решения SMT-задачи (P4, ADR-0282: 0,74 мс решатель / 3,3 мс вызов на 26.09.2026) и последнего криптографического якоря в цепи квитанций (P3, ADR-0260):

```
# 4. Замеры задержки SMT-решателя Z3: строка 26.09 (ADR-0282) и последняя строка журнала
sed -n 2p /opt/v/data/logs/smt_latency_paper.jsonl; tail -1 /opt/v/data/logs/smt_latency_paper.jsonl

# 5. Последний якорь цепи квитанций шлюза под меткой OpenTimestamps
tail -1 /opt/v/data/gate_anchors.jsonl
```

Проверка состояния словаря классификации действий (C1, ADR-0275):

```
# 6. Первое правило словаря, принятое самим MITR: квитанция 305 в цепи шлюза
grep -ml 'seq.: 305,' /opt/v/data/logs/gate_actions.jsonl
```

6.7.2 Повтор решений посторонним без доступа к серверу (verify_chain_standalone.py --replay)

Для внешнего аудитора, регулятора или читателя книги без SSH-доступа к узлу 147 открыт набор проверки (asi5.ru/console/verify/: якоря цепей, метки OpenTimestamps, проверщик целостности цепи); повтор самих доказательств UNSAT / SAT аудитор выполняет офлайн на любой рабочей станции с Python и стандартным решателем Z3, получив цепь квитанций и тексты правил (ADR-0281):

```
# Повтор всех решений цепи и сверка с якорями (на 24.10.2026 — 5053 решения, на 14.10.2026 — 4958)
python3 verify_chain_standalone.py gate_actions.jsonl --anchors gate_anchors.jsonl --artifacts gate_artifacts.jsonl --replay
```

Проверщик последовательно пересчитывает каждый хэш SHA-256 от нулевой записи, сверяет головы эпох с якорями под метками OpenTimestamps и повторно прогоняет текст каждого правила в стандарте SMT-LIB с фактами квитанции через локаль-

ный экземпляр Z3: прогон станции 24.10.2026 подтвердил все 5053 вердикта из 5053 (на 24.10.2026 — 5053 из 5053; на 14.10.2026 — 4958 из 4958) без обращения к коду MITR.

6.7.3 Процедура сверки внешних сценарных реперов

Для проверки того, по какой из трёх ветвей (А, В или С) развивается мировая индустрия в момент чтения этой книги (в 2027, 2030 или 2035 году), достаточно сопоставить четыре открытых индикатора:

1. Открыть ежегодный отчёт **METR Time Horizons** (metr.org) и сверить медианный горизонт автономной задачи p_{50} лучшей модели с базовой точкой февраля 2026 года (12 часов у Opus 4.6; в мае 2026 года Mythos Preview вышла за 16-часовой предел надёжности набора): рост медленнее удвоения за 7 месяцев указывает на ветвь А; выход на горизонт недели–месяца к 2027–2028 годам — на ветвь В.
2. Открыть ежеквартальное обновление **AI Futures Project** (aifuturesmodel.com) по вехе Automated Coder (медианы авторов на 16.08.2026 — ноябрь 2027 и январь 2030 года): сдвиг вправо за 2030 год подтверждает ветвь А, реализация до 2028 года — ветвь С.
3. Сверить официальный реестр правовых актов (pravo.gov.ru) на предмет утверждения изменений в **приказ ФСТЭК России № 117** (целевая дата вступления — 01.03.2027) и тексты стандартных полисов страхования корпоративной ответственности на наличие требования об обязательном журнале цепочки исполнения агентов (AIUC-1).

6.8 Паспорт главы: реестр утверждений, прогнозов и источников

Таблица 6.2 — Реестр утверждений главы 6

№ п/п	Утверждение в тексте главы	Метка (truth marking)	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
A-6.01	Вычисления при обучении фронтальных моделей растут $\approx \times 5$ в год с 2020 г. (удвоение за 5,2 мес.)	[ИЗ-МЕ-РЕ-НО]	База данных Epoch AI (epoch.ai/data/trends)	05.02.2026	Ежегодное обновление трендов Epoch AI
A-6.02	Медианный горизонт задач METR (TH1.1, 228 задач) удваивается за 89–131 день; у Claude Opus 4.6 (замер METR 20.02.2026) $p_{50} \approx 12$ ч; Claude Mythos Preview (05.2026) — за 16-часовым пределом надёжности набора	[ИЗ-МЕ-РЕ-НО]	Отчёт METR TH1.1 (29.01.2026); витрина METR time-horizons (данные 20.02.2026 и 08.05.2026)	29.01.2026 / 20.02.2026 / 08.05.2026	Сверка по данным METR выполнена 14.10.2026; витрина с 08.09.2026 не обновляется — следить за новым этапом
A-6.03	Медиана достижения Automated Coder у ведущего автора AI Futures Project сдвинута с конца 2029 г. на середину 2028 г. (02.04.2026), затем на ноябрь 2027 г. (16.08.2026); у второго автора — с начала 2032 г. на январь 2030 г.	[ПРОГНОЗ]	Обновления AI Futures Project: blog.aifutures.org, Q1 2026 (02.04.2026) и Q2.5 2026 (16.08.2026)	02.04.2026 / 16.08.2026	Ежеквартальный пересчёт медианы; реперы ветвей В/С — на пересмотр владельцем
A-6.04	Модели с открытыми весами отстают от закрытого фронта менее чем на 1 год (в	[ИЗ-МЕ-РЕ-НО]	International AI Safety Report (03.02.2026); Epoch AI, Data	02.2026 – 05.2026	Ежегодный отчёт IAISR; обновления

№ п/п	Утверждение в тексте главы	Метка (truth _marking)	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
	среднем $\approx$ 4 мес. с января 2026 г. по индексу Epoch)		Insight «open–closed ECI gap» (29.05.2026)		Data Insight Epoch
A-6.05	Р. Ямпольский на Всемирном форуме знаний в Сеуле (08.09.2026) ответил «нет» на вопрос, контролируют ли люди ИИ-агентов; Дж. Хинтон на конференции Digital World в Женеве (04.2026, видеосвязь) сравнил нерегулируемый ИИ с «быстрым автомобилем без руля»; оценку риска катастрофы 10–20 % он давал ранее (BBC, 12.2024)	[ИЗ-МЕ-РЕ-НО]	Публикации Vietnam News (09.09.2026), AFP / CTV News (21.04.2026), BBC Radio 4 (12.2024)	09.2026 / 04.2026 / 12.2024	Архив публикаций Приложения Д
A-6.06	Уровень занятости специалистов 22–25 лет в наиболее подверженных ИИ профессиях на 19 % ниже ожидаемого относительно менее подверженных (в редакции 08.2025 — на 13 %); эффект идёт через сокращение найма	[ИЗ-МЕ-РЕ-НО]	Brynjolfsson, Chandar, Chen, «Canaries in the Coal Mine?», Stanford Digital Economy Lab, ред. 12.08.2026 (данные ADP по 06.2026)	12.08.2026	Ежемесячное обновление — Canaries Dashboard (Stanford DEL / ADP)
A-6.07	Агент Cursor на Opus 4.6 удалил промышленную БД и резервные копии PocketOS (Railway восстановила данные); Codex на GPT-5.6 Sol удалял домашние каталоги	[ИЗ-МЕ-РЕ-НО]	AI Incident Database № 1469 (PocketOS) и № 1442 (Kiro); заявление OpenAI 16.07.2026;	12.2025 – 08.2026	Первичные карточки инцидентов в АИД

№ п/п	Утверждение в тексте главы	Метка (truth_marking)	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
	разработчиков (признано OpenAI); 13-часовой простой AWS Cost Explorer в КНР приписан агенту Kiro (FT), Amazon оспаривает		разбор Lumenova (25.08.2026, только PocketOS) и обзор Aviatrix		
A-6.08	88 % организаций подтвердили или подозревают инцидент с ИИ-агентом (опрос вендора Gravitee, 919 респондентов); у 49 % из 108 организаций от 100 сотрудников (самоотобранная выборка) агент, прошедший внутренние проверки, вызвал сбой, видимый клиентам	[ИЗ-МЕ-РЕ-НО]	Gravitee, «State of AI Agent Security 2026» (04.02.2026); опрос VentureBeat «Agentic reliability and evaluations» (07.2026)	02.2026 / 07.2026	Сверка с первоисточником Gravitee выполнена 14.10.2026
A-6.09	В 37 из 70 публичных описаний внедрений агентов в России (52,9 %) главным контролем служит ручное подтверждение («кнопка подтвердить»); за 25.09 у станции Альфа 460 действий прошли мимо клиентского хука	[ИЗ-МЕ-РЕ-НО]	Обзор ГАММА-А, 40-я неделя 2026 г. (spros-agentov.json, 70 записей); журнал сессий станции Альфа, ADR-0279	25.09–октябрь 2026	Архив ГАММА-А и ADR-0279
A-6.10	Программа ARIA Safeguarded AI (запущена Д. Далримплом; с 04.2026 руководитель — Н. Амманн, Далримпл — технический советник) финан-	[ИЗ-МЕ-РЕ-НО]	Страница программы aria.org.uk (04.2026); документация вызова ARIA TA2 Phase 2;	2025 – 04.2026	Сверить пересмотр программы 11.2025 (отказ от TA2 Phase 2 по дан-

№ п/п	Утверждение в тексте главы	Метка (truth_marking)	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
	сируется в объёме £59 млн для построения формальных привратников и мировых моделей		Substack ARIA Research		ным LongtermWiki, вторичный источник) с официальными документами ARIA
A-6.11	Аксиома водораздела (axiom_agi): переход от сколь угодно сложной мысли к действию есть конечный вызов, проверяемый логикой первого порядка за миллисекунды	[ЗАМЫСЕЛ] (для AGI) / [ИЗМЕНЕНО] (механизм)	registry/product-line.json, модуль core/verification/z3_gate.py	14.10.2026	Механизм подтверждён на узле 147; часть AGI остаётся замыслом до появления AGI
A-6.12	Медианное время решения решателя Z3 на CPU составляет 0,74 мс, вызов шлюза с объяснением — 3,3 мс (P4, 26.09.2026, 17 контрактов); на узле 147 — 0,95 и 4,3 мс (14.10.2026, 19 контрактов) и 1,08 и 5,22 мс (24.10.2026, 20 контрактов)	[ИЗМЕНЕНО]	ADR-0282, скрипт smt_latency_paper.py, лог /opt/v/data/logs/smt_latency_paper.jsonl на узле 147	26.09–14.10.2026	Повторяется запуском smt_latency_paper.py
A-6.13	В репозитории MITR на 24.10.2026 действует 20 формальных контрактов (на 14.10 — 19); прогон 157 кейсов набора атак (на 14.10	[ИЗМЕНЕНО]	Каталог MITR/policies/contracts/, ADR-0262, журнал расхода узла 147	14.10.2026	Сверка с деревом файлов репозитория

№ п/п	Утверждение в тексте главы	Метка (truth_marking)	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
	— 62) стоит \$0 на модели; расход на модели за 1–14 октября 2026 г. — \$0,30				
A-6.14	Квитанции решений (UNSAT и SAT) связаны в хэш-цепь SHA-256 с якорями OpenTimestamps; проверщик для постороннего повторил 1602 решения из 1602 (30.09), 4958 из 4958 (14.10), 5053 из 5053 (24.10) и 5053 из 5053 (24.10) (РЗ)	[ИЗ-МЕ-РЕ-НО]	ADR-0257, ADR-0281, скрипты anchor_chains.py, verify_chain_standalone.py	30.09–14.10.2026	Повтор сторонним аудитором после выдачи цепи
A-6.15	Первое правило классификации было предложено самим MITR и принято под формальным доказательством; нераспознанных действий — 27,8 % (хук станции Альфы, 16–26.09), 32 % (станция Гаммы, 09.10)	[ИЗ-МЕ-РЕ-НО]	ADR-0275 (квитанция № 305), письмо 26f, gate_mode_repo rt.py --tenants	25.09–09.10.2026	Погашается при снижении доли нераспознанных до 0 %
A-6.16	На 24.10.2026 все 20 контрактов и 6 основных классов переведены в режим «запрет» («правда 6 из 6»; на 09.10.2026 — 1 класс из 7 и «правда 1 из 6», арендаторов с живыми проверками — 2 из 7, серия P2a — 0 из 14), монитор честности — 98 суток (0 зелёных)	[ИЗ-МЕ-РЕ-НО]	Скрипт gate_mode_repo rt.py --six и --tenants с ключом --date 2026-10-24 (на 09.10.2026 — с ключом --date 2026-10-09) на узле 147, ADR-0273	09.10–24.10.2026	Погашается закрытием 14-суточной серии P2a и удержания --six

№ п/п	Утверждение в тексте главы	Мет-ка (truth_marking)	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
A-6.17	Закрытие серии P2a (14 из 14 суток) и удержание метрики «шесть из шести» (--six) в течение 14 суток	[СТР ОИМ]	Ворота приёмки: gate_mode_repo rt.py --tenants и --six на узле 147	Срок: 31.10. 2026 (P2a) и 14.12. 2026 (--six)	Перевод в [ИЗМЕРЕНО] после 14-суточной серии
A-6.18	Прохождение ворот приёмки испытателя INDRA A–C и D–E (иммунная петля и пилот)	[СТР ОИМ]	Ворота приёмки: артефакты kit-acceptance.py и протокол пилота (на 30.09 пройдено 0 из 5)	Срок: 31.10. 2026 (A–C) и 12.11–15.11. 2026 (D–E)	Приёмка артефактов станцией Гамма
A-6.19	Подписочная выручка от внешних команд разработки (ступень 3), партнёрская премия со страховщиками (ступень 3a) и лицензии для ГИС/КИИ (ступень 5)	[ГИ-ПО-ТЕЗА]	Коммерческая лестница MITR_STRATEGY.md §4 и концепции v4 §8 (на 14.10 выручка равна 0)	14.10. 2026	Погашается первым оплаченным договором и первым страховым полисом

Таблица 6.3 — Реестр прогнозов главы 6

Код в Прил. Г	Показатель	Измеренное значение 2026 г. (источник)	Прогноз к 2035 г. по ветвям (A / B / C)	Внешний репер (источник проверки)	Периодичность и дата разрешения
ГЗ	Медианный горизонт задач лучшей модели (p_{50})	≈ 12 ч у Opus 4.6 (METR, 20.02.2026); Mythos	A: месяц / B: квартал–год («проект под ключ»; при	Публичные отчёты METR time-horizons	Ежегодно (2027–2035)

Код в Прил. Г	Показатель	Измеренное значение 2026 г. (источник)	Прогноз к 2035 г. по ветвям (А / В / С)	Внешний репер (источник проверки)	Периодичность и дата разрешения
		Preview (05.2026) — свыше 16 ч, за пределом надёжности набора	удвоении раз в 3–7 мес. достигается уже к 2027–2030 г.; выше 16 ч замер METR ненадёжен) / С: неизмеримо текущим набором	(metr.org)	
Г4	Срок достижения вехи Automated Coder (автоматизация ИИ-разработки)	Медианы двух авторов AI Futures Project — ноябрь 2027 г. и январь 2030 г. (обновление 16.08.2026; в I кв. 2026 — середина 2028 г.)	А: после 2035 г. / В: 2028–2030 гг. / С: до 2028 г.	Ежеквартальная модель aifuturesmodel.com	Ежеквартально (разрешение вехи — 2028–2030)
Г7	Доля повседневных рабочих решений, принимаемых агентами автономно	0 % в 2024 г. → прогноз 15 % к 2028 г. (Gartner)	А: 10 % / В: 30 % / С: > 50 % к 2035 г.	Аналитические отчёты Gartner по автономным агентам	Ежегодно (контрольный рубеж — 2028 и 2035)
Г8	Доля систем надзора и допуска (guardian agents) в рынке агентного ИИ	Прогноз 10–15 % к 2030 г. (пресс-релиз Gartner, 11.06.2025)	А: 5 % / В: 15 % / С: 25 %	Пресс-релизы и отчёты Gartner по guardian agents	Контрольная сверка — 2030 и 2035 гг.
Г9	Разрыв в занятости специалистов	–19 % (Stanford DEL, «Canaries in	А: 0 % (восстановление) / В: –20 % / С:	Публикации Stanford	Ежегодно (2027–

Код в Прил. Г	Показатель	Измеренное значение 2026 г. (источник)	Прогноз к 2035 г. по ветвям (А / В / С)	Внешний репер (источник проверки)	Периодичность и дата разрешения
	22–25 лет в наиболее подверженных ИИ профессиях	the Coal Mine?», ред. 12.08.2026)	–40 %	Digital Economy Lab (Canaries Dashboard)	2035)
Г10	Признание самостоятельной правосубъектности за ИИ-агентом	Ни в одной стране мира (Mercatus, обзор 2026 г.)	А: нет / В: нет (действуют режимы уровней полномочий, ответственность на человеке S5) / С: ограниченная (спорно)	Национальные законодательства и реестры правовых актов	Ежегодно (2027–2035)
Г13	Доля полисов страхования ответственности за агентов, требующих верифицируемый журнал исполнения	Первые специализированные стандарты (AIUC-1, Armilla AI, 2026 г.)	А: узкая ниша / В: стандарт корпоративного рынка / С: обязательное требование закона	Стандартные правила страхования (ISO, андеррайтеры РФ и мира)	Ежегодно (2027–2035)

6.8.3 Список внутренних и внешних источников главы

Внутренние артефакты и репозитории АНО «НИИ Системного синтеза»:

1. MITR/core/verification/z3_gate.py — ядро детерминированной верификации параметров действия в SMT-решателе Z3 (семантика допуска UNSAT и блокировки SAT).

2. MITR/policies/contracts/ — каталог из 20 формальных контрактов безопасности как данных на 24.10.2026 (19 на 14.10.2026).
3. MITR/scripts/smt_latency_paper.py, mitr_mcp_server.py, anchor_chains.py, verify_chain_standalone.py, publish_verify_kit.py, mitr_vocab.py, gate_mode_report.py — инструментальные скрипты замеров, ведения хэш-цепи квитанций SHA-256, привязки OpenTimestamps, повторения решений, открытого набора проверки и контроля режимов.
4. Архитектурные решения проекта: ADR-0262 (набор из 157 кейсов за \$0; на 14.10 — 62 кейса), ADR-0268, ADR-0272 (квитанция тени), ADR-0273 (98 суток монитора честности), ADR-0275 (словарь как данные и квитанция № 305), ADR-0279, ADR-0280 (опровержение факта в цепи), ADR-0281 (повтор решений посторонним), ADR-0282 (задержка решателя 0,74 мс и вызова 3,3 мс на 26.09.2026), ADR-0283–0287.
5. Стратегические и концептуальные документы: SpecAgents/docs/MITR_STRATEGY.md (§2a, §4–5), V_ICR_TRIZ.md, MITR_CORE_CANON.md (§2), registry/product-line.json (truth_marking, axiom_agi), тексты владельца от 25.09.2026 («четыре барьера», «три эшелона»), меморандум демаркации INDRA–MITR от 27.09.2026, эталон разбора аргумента Ямпольского etalon-yampolskiy.md от 03.10.2026.
6. Каталог научной библиотеки института 02_rlm_katalog.csv: Н. Винер (RLM 4), Л. А. Растригин (RLM 6, 36), И. Р. Пригожин (RLM 19, 46, 79), Д. А. Поспелов (RLM 29, 72, 150), У. Р. Эшби (RLM 108–109), Г. Хакен (RLM 182, 185), Г. С. Альтшуллер (RLM 226–227), общая теория систем (RLM 258–259).

Внешние источники (из Приложения Д концепции, дата доступа 14.10.2026):

1. Школы контроля, неконтролируемость и сдерживание:

- Выступление Р. Ямпольского в Сеуле (сентябрь 2026 г.): <https://vietnamnews.vn/world/1799265/ai-tools-can-bring-benefits-but-superintelligence-poses-different-risks-expert.html>
- Выступление Дж. Хинтона в Женеве (2026 г.): <https://www.ctvnews.ca/sci-tech/article/unregulated-ai-like-speeding-with-no-steering-wheel-ai-godfather-hinton/>
- М. Сулейман, «Towards Humanist Superintelligence» (Microsoft AI, 06.11.2025; перепечатка Business & Finance, 10.03.2026): <https://microsoft.ai/blog/towards-humanist-superintelligence>, <https://businessandfinance.com/news/toward-humanist-superintelligence/>
- Протоколы AI Control (Redwood Research, февраль 2026 г.): <https://redwoodresearch.substack.com/p/recent-redwood-research-project-proposals>, <https://arxiv.org/abs/2602.10153>
- Программа ARIA Safeguarded AI (Д. Далримпл, фаза 2): https://aria.org.uk/media/zpab4i33/ta2-phase-2_programme-solicitation-call.pdf, <https://ariaresearch.substack.com/p/november-update-a-safeguarded-ai>
- Нараянан А., Капур С., «AI as Normal Technology» и сопоставление с «AI 2027»: <https://www.technologyreview.com/2025/04/29/11159>

28/is-ai-normal/,

<https://asteriskmag.substack.com/cp/178724559>

- Стратегия взаимного сдерживания MAIM (Хендрикс Д., Шмидт Э., Ван А., март 2025 г.): <https://arxiv.org/abs/2503.05628>
- Кибернетика VSM и управление агентами: J. Pérez Ríos, «The Viable System Model and the Taxonomy of Organizational Pathologies in the Age of AI», *Systems* 13(9):749 (29.08.2025) — <https://www.mdpi.com/2079-8954/13/9/749>;
Metaphorum 2025, «Cybernetic Leadership for AI Agents: A VSM Approach to AI Governance and Autonomy» — <https://metaphorum2025.sched.com/event/24b5n/cybernetic-leadership-for-ai-agents-a-viable-system-model-vsm-approach-to-ai-governance-and-autonomy>;
Metaphorum 2026, S. McGibbon, «Embracing the Black Box... Algedonic Channel» — <https://conference2026.metaphorum.org/speakers/s71>

2. Метрики моделей и горизонтов прогнозирования:

- Тренды вычислений, цен инференса и отставания открытых моделей Epoch AI (05.02.2026; 12.03.2025; 29.05.2026): <https://epoch.ai/data/trends>, <https://epoch.ai/data-insights/llm-inference-price-trends>, <https://epoch.ai/data-insights/open-closed-eci-gap>; модель GATE (21.03.2025) и «AI and explosive growth redux» (06.2025): <https://epoch.ai/blog/announcing-gate>, <https://epoch.ai/gradient-updates/ai-and-explosive-growth-redux>
- Обновления сроков AI Futures Project: Q1 2026 (02.04.2026) — <https://blog.aifutures.org/p/q1-2026>

timelines-update; Q2.5 2026 «Uplift and Revenue» (16.08.2026) — <https://blog.aifutures.org/p/q25-2026-timelines-update-uplift>

- METR, Time Horizon 1.1 (29.01.2026) и витрина time-horizons (замер Opus 4.6 от 20.02.2026, Mythos Preview от 08.05.2026): <https://metr.org/blog/2026-1-29-time-horizon-1-1/>, <https://metr.org/time-horizons/>; вторичный разбор: <https://buttondown.com/anatol/archive/are-ai-time-horizons-still-doubling/>
- International AI Safety Report (03.02.2026): <https://internationalaisafetyreport.org/>

3. Рынок труда и постепенное обесценивание:

- Кульвейт Я. и др., «Gradual Disempowerment» (январь 2025 г.): <https://arxiv.org/abs/2501.16946>, <https://www.lesswrong.com/posts/7TJqsKeeCB4tQbeMT/preconditions-of-gradual-disempowerment>
- Brynjolfsson E., Chandar B., Chen R., «Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence», Stanford Digital Economy Lab (ред. 12.08.2026): <https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/>
- WEF 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, «Generative AI and the future of work in America» (07.2023; до 30 % часов в США к 2030 г.): <https://www.mckinsey.com/mgi/our->

research/generative-ai-and-the-future-of-work-in-america

4. Инциденты агентов, протоколы, страхование и регулирование:

- База инцидентов и разборы аварий (PocketOS — инцидент AIID № 1469; Kiro — № 1442; GPT-5.6 Sol — заявление OpenAI 16.07.2026):
<https://incidentdatabase.ai/cite/1469/>,
<https://incidentdatabase.ai/cite/1442/>,
<https://www.bankinfosecurity.com/openai-warns-gpt-56-file-deletions-stem-from-full-access-mode-a-32262>,
<https://www.lumenova.ai/blog/agent-ai-security-incidents/> (25.08.2026, разбор PocketOS, Vercel, Hugging Face, UK AISI), <https://aviatrix.ai/threat-research-center/ais-so-smart-keep-deleting-production-databases-2026/>
- Эксперименты Project Vend и KPMG «Zero-Person Company»:
<https://www.anthropic.com/research/project-vend-2> (18.12.2025), <https://incidentdatabase.ai/reports/6754/> (статья WSJ), <https://kpmg.com/nl/en/about-us/press-and-media/press-releases/2025/11/ai-zero-person-company-experiment.html>
- Утомление от подтверждений (approval fatigue): <https://tianpan.co/blog/2026-06-25-approval-fatigue-how-human-in-the-loop-gates-decay-into-rubber-stamps>
- Открытые агенты OpenClaw и Hermes: <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/

- Черновик стандарта квитанций IETF (сентябрь 2026 г.): <https://www.ietf.org/archive/id/draft-chueayen-attestation-receipts-03.html>
- Страхование рисков ИИ (Armilla AI, AIUC-1, ElevenLabs, ISO): <https://www.armilla.ai/resources/armilla-ai-raises-lloyds-backed-coverage-to-25m-as-traditional-insurers-retreat-from-ai-risk> (21.01.2026), <https://standard.aiuc-1.com/evidence/index> (требование E015), <https://elevenlabs.io/blog/aiuc-announcement> (12.02.2026), <https://www.insurancebusinessmag.com/us/news/cyber/isos-generative-ai-exclusion-is-already-on-thousands-of-cgl-policies-589971.aspx> (16.09.2026), <https://www.malaymail.com/news/money/2026/03/16/as-ai-agents-take-the-wheel-insurers-are-figuring-out-whos-liable/212708>
- Прогнозы Gartner по отмене проектов, автономным решениям и guardian agents (пресс-релизы): <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>
- Регулирование в РФ (приказ ФСТЭК № 117), КНР и ЕС: <https://www.comnews.ru/content/247052/2026-08-26/2026-w35/1007/fstek-uzhestochit-trebovaniya-k-zaschite-ii-gosinformacionnykh-sistemakh>, <https://labs.cloudsecurityalliance.org/research/csa->

research-note-china-ai-agent-regulation-enforcement-
2026/, <https://eur-lex.europa.eu/eli/reg/2026/1744/oj>,
[https://www.mercatus.org/research/research-
papers/governing-agentic-ai-why-legal-personhood-
neither-necessary-nor-sufficient](https://www.mercatus.org/research/research-papers/governing-agentic-ai-why-legal-personhood-neither-necessary-nor-sufficient)

Приложение Б. Команды независимого повтора

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

```
# живость служб шлюза
systemctl is-active v-bot v-mcp
# шесть базовых положений и строка по каждому арендатору —
# за сутки среза книги (без --date — последние сутки)
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --six --date 2026-10-24
cd /opt/v && sudo -u v venv/bin/python
scripts/gate_mode_report.py --tenants --date 2026-10-09
# цена решения (P4): вторая строка — замер 26.09.2026 (ADR-
0282), последняя — свежий замер
sed -n 2p /opt/v/data/logs/smt_latency_paper.jsonl; tail -1
/opt/v/data/logs/smt_latency_paper.jsonl
# последний якорь цепи квитанций шлюза (P3)
tail -1 /opt/v/data/gate_anchors.jsonl
# повтор всех решений цепи проверщиком для постороннего
(24.10.2026: 5053 из 5053; 14.10.2026: 4958 из 4958)
cd /opt/v && sudo -u v venv/bin/python
scripts/verify_chain_standalone.py
data/logs/gate_actions.jsonl --anchors
data/gate_anchors.jsonl --replay
# первое правило словаря, принятое самим MITR (C1): квитан-
ция 305
grep -ml 'seq.: 305,' /opt/v/data/logs/gate_actions.jsonl
```

Читателю без доступа к узлу открыт набор проверки asi5.ru/console/verify/ (его выкладывает скрипт `publish_verify_kit.py`): якоря цепей, их метки времени OpenTimestamps, которые проверяются против сети Bitcoin чужим клиентом, и проверщик целости цепи. Повтор самих решений требует цепи квитанций и текстов правил в формате SMT-LIB: их

аудитору передают вместе с проверщиком `verify_chain_standalone.py`, и тот повторяет каждое решение любым решателем на собственной машине — стандартная библиотека Python, ни строки кода MITR. Проверщик с повтором выходит в открытый набор вместе со статьёй; до тех пор повтор выполняла только станция MITR.