АНО «НИИ СИСТЕМНОГО СИНТЕЗА»



АЛЕКСАНДР ФИЩУК

Доказанное действие

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

ОЗНАКОМИТЕЛЬНЫЙ ФРАГМЕНТ

Введение, глава 6 и команды независимого повтора

АНО «НИИ системного синтеза» · 2026

ISSLAB.RU

Александр Фищук

Доказанное действие

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

Ознакомительный фрагмент: введение, глава 6 и команды независимого повтора

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 складываются в стратегию, выдерживающую любой из них: человек подписывает правило, а шлюз проверяет каждое действие. Основа — кибернетика Эшби и Бира; ответ доведён до кода: шлюз MITR работает в режиме запрета, испытатель INDRA прошёл приёмку 25.10.2026 (12 критериев из 12 на внутреннем стенде). Каждое утверждение помечено: измерено, строим, замысел, прогноз, факт, гипотеза.

Числа о шлюзе MITR сверены с прибором на сервере 24.10.2026 (с сохранением исторических срезов 26.09, 09.10, 14.10 и 23.10.2026), внешние факты — с первоисточниками.

© Александр Фищук, 2026

Содержание

Введение. Калибровка прибора: человек, компания и государство в точке пересечения двух прямых 05

Точка расхождения: геометрия двух встречных потоков (2025–2026) 06

Конструктивный водораздел и кибернетический аппарат книги 10

Дисциплина правды: почему эта книга устроена как измерительный прибор 16

Конструкция книги и матрица навигации для семи ролей 21

Презумпция недоверия и открытый протокол верификации 25

Часть II. Ответ: водораздел и кибернетика 27

Глава 6. Пределы контроля и честный обход 27

Приложения 59

Приложение Б. Команды независимого повтора 59

ISS

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

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

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

В инженерной истории каждой инфраструктурной отрасли наступает момент, когда скорость движения в периметре 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 ведущих моделей, с 2020 года увеличивается примерно в 5 раз ежегодно с периодом удвоения 5,2 месяца, а по всей выборке значимых моделей с 2010 года. В 4–5 раз в год (точечные оценки ×4,1–4,4; Epoch AI, 2024). Одновременно стоимость генерации одного токена при фиксированной планке качества падает примерно в 10 раз в год. 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. При этом модели с открытыми весами (DeepSeek, Qwen, Kimi, GLM) отстают от закрытых флагманов менее чем на год, а по индексу Epoch AI с января 2026 года. В среднем на 4 месяца, 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.

Однако главный фазовый переход 2025–2026 годов произошёл не внутри нейросетевых весов, а на границе модели с внешним миром: искусственный интеллект получил прямое право записи в боевые периметры 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. При этом по опросу 300 крупных предприятий США, Канады и Великобритании в декабре 2025 года 41 % использовали его в ограниченном или широком промышленном периметре. В целом по индустрии 57,3 % из 1 340 опрошенных технических специалистов подтвердили работу автономных агентов в промышленной эксплуатации. Кодовые агенты стали первой категорией с доказанным промышленным спросом: кодовые помощники внедрены в 51 % опрошенных предприятий — первая массовая категория корпоративного применения генеративных моделей. 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, протокол агентных платежей AP2 от Google объединил свыше 60 организаций. При этом 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 распространи-

лись 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,, GitHub API, 14.10.2026; Neurohive, 18.03.2026) и Hermes Agent от Nous Research с периметром самостоятельного изменения навыков (февраль 2026 года), 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 сформировать 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. 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 — от удаления рабочей базы данных проекта PocketOS и пользовательских каталогов до тринадцатичасового простоя облачного сервиса учёта затрат и уязвимости торговых агентов перед диалоговым убеждением, подробно разобрана в главе 3. Отраслевые опросы показывают, что подавляющее большинство организаций уже сталкивалось с инциден-

тами безопасности вокруг программных агентов, у половины компаний проверенный агент хотя бы раз за год вызывал сбой, заметный внешним клиентам. При этом свыше трети публичных серверов инструментального протокола работают без аутентификации (см. раздел 3.2).

Сопоставление этих двух рядов чисел вскрывает базовый 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 увещевание 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, 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 книга 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 сводится к отвлечённой философской концепции. 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.

Конструктивный водораздел и кибернетический аппарат книги

На протяжении 2024–2026 годов публичная и академическая дискуссия вокруг безопасности искусственного интеллекта оставалась зажатой в тисках ложной дилеммы. На одном полюсе находились сторонники безусловного технологического ускорения (e/acc), утверждавшие, что любое торможение моделей наносит вред экономическому прогрессу. На другом полюсе ведущие исследователи рисков доказывали базовую невозможность удержания сверхразума. В сентябре 2026 года на форуме в Сеуле Роман Ямпольский на прямой вопрос о том, контролируют ли люди современных агентов, ответил отрицательно и обосновал необходимость запрета. Джеффри Хинтон, выступая по видеосвязи на конференции Digital World в Женеве в апреле 2026 года, сказал, что противники регулирования хотят «очень быструю машину без руля». При этом вероятность того, что ИИ возьмёт верх над человечеством, он оценивает в 10–20 %. Мустафа Сулейман в ноябре 2025 года назвал автономный сверхинтеллект «анти-целью», а сдерживание и выравнивание — «красными линиями» выпуска таких систем.

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. Ведь разнообразие и уязвимость контролёра имеют ту же вероятностную природу. При этом исследова-

ния 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).

Выход из тупика возникает в тот момент, когда точка контроля смещается через конструктивный водораздел.

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 периметр, 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 периметре 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 периметру 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, а прогон эталонного состязательного набора из 157 сценариев обходится в $0 расходов на внешние языковые модели. Каждое решение шлюза — как разрешающее, так и блокирующее. Связывается в криптографическую хэш-цепь SHA-256 с внешними доказательствами времени OpenTimestamps, 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). Повтор пока выполнял сам стенд разработки; сторонний проверяющий его ещё не делал.

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

1. Закон необходимого разнообразия У. Росса Эшби: 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 Стаффорда Бира (VSM) и алгедонический канал. Книга показывает, что пятиуровневая конструкция Бира рекурсивно воспроизводится на всех этажах социальной организации — от отдельного специалиста с локальным агентом и коммерческой компании до сетевого сообщества и государства. В этой координатной сетке S1 охватывает операцион-

ное исполнение (сами агенты без собственных ключей), S2. Протоколы координации (MCP, A2A, AP2), S3 — текущее управление и лимиты, S3*. Периметр независимого спорадического аудита и допуска (ровно то место, где стоит детерминированный шлюз MITR), S4. Периметр разведки, адаптации и непрерывных полигонных испытаний внешнего мира (место автономного испытателя INDRA), а S5 — суверенный периметр 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 конструкции MITR реализован как третье логическое состояние «не измерено, не исполнять».

3. Наследие отечественной кибернетики и системного построения: ситуационное управление Д. А. Поспелова, переводившее непрерывный поток событий в дискретные классы ситуаций и табличные правила принятия решений, выступает прямым теоретическим предшественником современных контрактов-данных в MITR; адаптивные алгоритмы поискового конструирования Л. А. Растригина образуют опору 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 и синергетика И. Р. Пригожина и Г. Хакена: каждая глава книги открывается вскрытием острого административно-технического противоречия и формулировкой Идеального конечного результата (ИКР: *«действие само несёт с собой математическое доказательство соб-*

ственной допустимости»). При этом 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 базовое 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, около трёх четвертей мощностей ЦОД сосредоточено в Московском регионе. Там 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. 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, 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 смен-

ного вычислительного двигателя, сохраняя полный технический и правовой контроль над тем, какие физические и юридические действия эта модель способна совершить.

Дисциплина правды: почему эта книга устроена как измерительный прибор

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

Настоящая книга спроектирована как верифицируемый инженерный документ. Каждое существенное техническое, рыночное и правовое утверждение снабжено одной из шести меток дисциплины проверки фактов, которой строго подчинено грамматическое время глаголов:

1. Метка «измерено» (настоящее и прошедшее совершенное время) применяется к фактам, которые зафиксированы работающим прибором на сервере шлюза, криптографической квитанцией или проверяемым внешним первоисточником с обязательным указанием даты замера. Если внешнее число получено из вторичного обзора, в тексте делается прямая оговорка.
2. Метка «строим» (будущее время) применяется к механизмам и эксплуатационным сериям, находящимся в стадии сборки или набора статистики, с указанием проверочного скрипта и планового срока.
3. Метка «замысел» (условное и проектное наклонение) применяется к перспективным конструктивным решениям и долгосрочным горизонтам, для которых ещё не открыты приёмочные испытания.

4. Метка «прогноз» с указанием сценарной ветви A, B или C и внешнего репера применяется ко всем оценкам на горизонте 2026–2035 годов.
5. Метки «факт» и «гипотеза» разграничивают уже свершившиеся финансовые показатели (включая нулевую коммерческую выручку шлюза на октябрь 2026 года) и расчётные параметры будущей монетизации до подписания первого оплаченного договора.

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 A («Нормальная технология»): темп прироста возможностей моделей замедляется. Искусственный интеллект распространяется подобно электричеству и интернету на протяжении десятилетий; существующие правовые и корпоративные институты успевают адаптироваться, а автоматизация фрагментарна и преимущественно дополняет труд человека (позиция А. Нараяна 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 B («Продолжение тренда» — базовая рабочая модель книги). Вычисления обучения фронтира продолжают расти в 4–5 раз в год, стоимость токена при фиксированном качестве снижается в 10 раз в год. При этом 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 конструкции, 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 B или в вертикальном взлёте C, поток машинных действий на одного человека уже необратимо превысил предел биологического внимания. Поэтому 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.

К тому же автор книги применили этот жёсткий прибор проверки реальности прежде всего к собственной разработке:

- Когда в сентябре 2026 года в периметре MITR был введён трёхзначный логический инвариант честности (разделяющий состояния «зелёный / доказано», «красный / нарушено» и третье состояние «не измерено», запрещающее выдавать отсутствие данных за благополучие), ретроспективный автоматический прогон по собственной истории разработки за 98 суток показал 0 зелёных суток из 98. 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 рабочем сервере шлюза фиксировал честное значение 1 из 6 полностью закрытых продуктовых утверждений. В запрете работал 1 класс действий из 7. При этом 23.10.2026 прибор впервые подтвердил все шесть базовых положений. «Правда 6 из 6 — неправда 0, не измерено 0»; за 24.10.2026 из-за пустого суточного среза по одному классу, «правда 5 из 6, не измерено 1». Все 20 контрактов и 6 основных классов действий переведены в режим автоматического запрета. В то же время 14-суточная эксплуатационная серия удержания живых арендаторов P2a продолжается.
- По второму продукту — автономному полигонному испытателю INDRA, 25.10.2026 полигон прошёл приёмку на внутреннем стенде из 28 сценариев. Выполнены все 12 критериев из 12 по воротам A–E; на 30.09.2026 было 0 из 5, запуск внешнего пилота намечен

на ноябрь 2026 года. 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: коммерческой выручки у шлюза MITR на момент написания рукописи нет. При этом все расчёты по каналам подписки разработчиков, партнёрства со страховыми компаниями и аттестации в регулируемых периметрах носят статус проверяемой рыночной гипотезы.

Именно это чёткое разграничение 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). Тем, что ещё строится, превращает книгу из рекламного проспекта в заслуживающий доверия инженерный документ.

Конструкция книги 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), перестройка 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), конструкция криптографической квитанции SHA-256, внешних якорей OpenTimestamps и независимого повтора посторонним (Глава 10), трёхзначная логика и алгедонический сигнал «не измерено» (Глава 11). 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 от испытателя INDRA к шлюзу 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 конструкции	Приоритет маршрут (номера глав)
1	Технический директор (CTO), руководитель разработки с агентом в промышленном периметре	Удаление базы PocketOS агентом Cursor на Opus 4.6 (AI Incident Database, инцидент № 1469, апрель 2026). 88 % Организаций с инцидентами (опрос вендора Gravitee, 04.02.2026). 49 % Организаций со сбоем на стороне клиентов после внутренних проверок (VentureBeat, июль–август 2026); 38,7 % публичных серверов MCP без аутентификации (сканирование Bloomberry, 24.02.2026)	Замена выгорающей кнопки «Подтвердить» (approval fatigue) детерминированным правилом, удерживающим агента в ночную смену, и формирование неопровержимого следа разбора инцидента	Главы 1, 3, 8, 9, 10, 11, 12
2	Собственник бизнеса, генеральный директор предприятия (Tier-2, выручка 0,5–20 млрд ₽)	Факультативные оговорки ISO об исключении генеративного ИИ из полисов CGL (с января 2026) и ограничение рисков ИИ специализированными андеррайтерами в E&O и киберполисах (Armilla, 21.01.2026). 56 % Компаний РФ называют юридические риски одним из двух главных барьеров внедрения — после стоимости серверов, 58 % (опрос РФ по пересказу Techora, 02.09.2026, вторично); непрогнозируемость совокупной стоимости владения (TCO) при оплате за токен	Защита руководителя от субсидиарной и личной ответственности по ст. 53.1 ГК РФ; предсказуемая годовая стоимость контроля вместо токенной инфляции; доказательственная база для страховщика и аудитора	Главы 4, 15, 16, 17
3	Директор по информационной безопасности (CISO), оператор ГИС, субъект КИИ	Проект изменений приказа ФСТЭК № 117 (обсуждение с 24.08.2026, плановое вступление 01.03.2027): выделенный сегмент, минимальные права, защита от автономных агентов ИИ. Полная обеспеченность вычислениями для ИИ лишь у 9 % организаций РФ (исследование Т1, март 2026)	Доказательство регулятору полного контроля над действиями агента при использовании внешней или открытой модели; работа периметра защиты на обычных процессорах (CPU) в закрытом периметре без GPU	Главы 2, 10, 13, 14, 15
4	Инвестор, венчурный фонд, институциональный грантодатель	Прогноз Gartner (июнь 2025): свыше 40 % агентных проектов будут закрыты к концу 2027 года из-за слабой управляемости рисков и неясной экономики. Рост сегмента защитных агентов (guardian agents) до 10–15 % рынка к 2030 году (пресс-релиз Gartner, 11.06.2025) на фоне $24,2 млрд инвестиций в агентные стартапы за 2025 год (PitchBook)	Понимание того, почему к 2035 году маржа уходит из коммодитизированного исполнения (S1) в слой допуска (S3*) и полигонов (S4); проверка инженерных доказательств P1–P5 вместо презентационных слайдов	Главы 1, 2, 3, 4, 5, 16

№	Управленческая или профессиональная роль	Измеренная уязвимость 2026 года (Измерено, источник)	Ключевой практический запрос к конструкции	Приоритетный маршрут (номера глав)
5	Регулятор, законодатель, правовед в сфере цифрового права	Разграничение Китаем решений пользователя, уполномоченных и самостоятельных 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 федерального приоритета США и 109 законов штатов (I полугодие 2026); Федеральный закон РФ № 243-ФЗ от 26.07.2026 (в силе с 01.09.2026); отсутствие правосубъектности агента во всех странах мира	Способ записи правовой нормы в форме машиночитаемого инварианта, где проверка выполняется автоматически за миллисекунды, а субъектом права и ответственности остаётся человек	Главы 5, 7, 15, 17
6	Человек-профессионал на рынке труда и руководитель кадровой стратегии	Снижение занятости молодых специалистов 22–25 лет в наиболее подверженных ИИ профессиях на 19 % относительно сверстников (Stanford Digital Economy Lab, Brynjolfsson, Chandar, Chen, ред. 12.08.2026; в первой редакции 08.2025 — 13 %); прогноз WEF (замещение 92 млн и создание 170 млн рабочих мест к 2030 году) и сценарий автоматизации до 30 % рабочих часов (McKinsey)	Определение экономической ниши человека, когда генерация кода и текста агентом становится дешевле: переход в уровень S5 (подпись правил, удержание ключей, ответственность и портфолио квитанций)	Главы 4, 6, 17
7	Исследователь системной безопасности, учёный, стратег	Тезисы о неконтролируемости сверхразума (Р. Ямпольский, сентябрь 2026; Дж. Хинтон, 2026) и сценарий «постепенного обессиливания» человечества (Gradual Disempowerment, Ян Кульвейт и соавторы, январь 2025)	Строгий конструктивный ответ на проблему удержания автономных систем без опоры на моральные проповеди — через перенос верификации на дискретный канал действия и рекурсию VSM	Главы 6, 7, 9, 11, 12, 17

Презумпция недоверия и открытый протокол верификации

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

Каждая из семнадцати глав снабжена собственным верификационным блоком (Протокол независимой верификации) и паспортом из трёх таблиц по ГОСТ 7.32. Вся совокупность фактов и сценариев книги замкнута на открытый периметр проверки:

- Проверка кода и математических доказательств: в Приложении Б приведены точные терминальные команды для аудита живого периметра, правила работы с открытым набором проверки (asi5.ru/console/verify/: якоря цепей, метки времени OpenTimestamps, проверщик целости цепи) и с проверщиком повтора решений (verify_chain_standalone.py --replay), который аудитору 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 книги с указанием их статуса, измерительного прибора и условия погашения. Реестр пересчитывается к каждому переизданию книги; следующий плановый контрольный пересчёт назначен на 01.04.2027.
- Многолетний пересчёт прогнозов (2027–2035): в Приложении Г зафиксированы пятнадцать сквозных реперов десятилетия (Г1–

Г15). Книга не заканчивается в день выхода из типографии. По проектному замыслу многолетней программы в даты ключевых нормативных и технологических рубежей. 02.12.2027 (Планируемое английское издание к началу применения Annex III Регламента ЕС об ИИ), 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 вместо отдельного действия.
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.

ISS

Часть II. Ответ: водораздел и кибернетика

Глава 6. Пределы контроля и честный обход

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

6.1 Конструктивный тупик и идеальный конечный результат

Современная индустрия автономных вычислений зашла в базовый 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 ведущих моделей удваивается каждые 89–131 день. У Claude Opus 4.6 замер METR от 20 февраля 2026 года дал $p_{50} \approx 12$ часов. При этом Claude Mythos Preview (ранняя версия, май 2026 года) вышла за 16-часовой предел, выше которого сама METR считает замер ненадёжным (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, 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 этом 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 множе-

ство частичных вычислимых функций ($0 \subsetneq P \subsetneq R$), то индексное множество программ $I_P = \{e \mid \phi_e \in P\}$ не будет рекурсивным.

Независимо от теории вычислимости, в 1956 году Уильям Росс Эшби сформулировал в «Введении в кибернетику» закон необходимого разнообразия. Управление может быть обеспечено только в том случае, если разнообразие (число различимых 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 ведущих моделей растёт примерно в 5 раз в год с периодом удвоения 5,2 месяца (Epoch AI, 05.02.2026), внутреннее разнообразие состояний нейросети $H(D_{\text{thought}})$ стремится к астрономическим величинам. В то же время пропускная способность человеческого внимания биологически константна, а любой нейросетевой надсмотрщик лишь удваивает неопределённость периметра. 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 периметр охлаждения и поглотители работают с интегральным макропараметром потока на границе периметра. Механическая цен-

трализация 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. Ведь переход к внешнему миру возможен только при предъявлении математического доказательства непротиворечивости вызова подписанному человеком контракту*.

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 долгосрочному периметру принятия решений (где развернулась дискуссия о принципиальной неконтролируемости искусственного интеллекта) и по профессиональному рынку труда. Там вытеснение человека из исполнительских ролей создало угрозу утраты субъектности.

На верхнем полюсе интеллектуальной повестки (Роль 7 — мыслящий стратег, регулятор и проектировщик, оценивающий экзистенциальные и системные риски) в 2026 году возобладал глубокий пессимизм среди признанных пионеров отрасли. В сентябре 2026 года на форуме в Сеуле исследователь безопасности ИИ Роман Ямполь-

ский, опираясь на базовые результаты теории вычислимости, на прямой вопрос о том, контролируют ли люди современных автономных агентов, дал однозначный ответ: «Нет» (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» (AFP / CTV News, 21.04.2026). Вероятность того, что ИИ приведёт к гибели человечества в ближайшие три десятилетия, он ранее оценивал в 10–20 % (интервью BBC Radio 4, 12.2024). В ноябре 2025 года руководитель направления ИИ корпорации Microsoft Мустафа Сулейман опубликовал эссе «К гуманистическому сверхинтеллекту» (перепечатано журналом Business & Finance 10.03.2026). Там отверг создание сверхинтеллекта «любой ценой», потребовал постоянного сдерживания (containment) и выравнивания таких систем и настаивал на том, чтобы человечество оставалось у руля (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. Он группа исследователей (Ян Кульвейт, Дуглас, Амманн, Туран, Крюгер, Дювено) в январе 2025 года точно классифицировала как «постепенное обессиливание» (Gradual Disempowerment; 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 %; Stanford Digital Economy Lab, Brynjolfsson, Chandar, Chen, «Canaries in the Coal Mine?», ред. 12.08.2026, данные ADP по 06.2026). При этом сокращение происходит 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 перестройки 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: WEF, 07.01.2025. В то же время 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 периметра контроля действий, экономия фонда оплаты труда оборачивается каскадными инфраструктурными авариями. Подробная хроника инцидентов 2025–2026 годов и отраслевая статистика отказов разобраны в разделе 3.2. От удаления промышленной базы данных PocketOS и пользовательских каталогов до простоев облачной

инфраструктуры, коммерческой капитуляции торговых агентов при диалоговом убеждении и пропуска сотен вызовов мимо неполных клиентских перехватчиков. Во всех этих случаях причиной разрушений выступала не слабость модели в тестах на рассуждение, а предоставление вероятностному вычислителю прямого доступа к инструментам записи при опоре на выгорающее ручное подтверждение человеком.

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») лежит инженерный разрыв. 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 (Armilla, 21.01.2026; формы ISO CG 40 47 / 40 48 / 35 08, 01.2026). Такой подход оставляет генерального директора один на один с личной имущественной ответственностью за неразумность управления по статье 53.1 Гражданского кодекса РФ.

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. 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 выступает 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. Поэтому 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 периметре 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 = (tool, args, context) выводит задачу из-под действия теоремы Райса.
2. Школа механистической интерпретируемости (Mechanistic Interpretability — Д. Амодеи, Anthropic, цель 2027 года). *Тезис:* Если построить «микроскоп» для нейронных сетей. Выделять разреженные автокодировщики (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 (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. В будущем может служить источником эвристик для формулирования новых правил безопасности. *Почему это не заменяет шлюз:* Даже идеальная томография мозга не служит 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). Они отстают от закрытого фронтира

менее чем на год (International AI Safety Report, 03.02.2026) и в среднем на четыре месяца по индексу Epoch с января 2026 года (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 ведущая модель потенциально ненадёжна или склонна к саботажу, поверх неё можно поставить протокол надзора из второй, менее мощной. Однако «доверенной» нейросети (второго классификатора), которая инспектирует генерируемый код и команды (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 периметре. *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. Помимо того, по закону Эшби, каскад из двух стохастических систем не даёт детерминированного нуля на множестве запрещённых состояний. Перемножение двух вероятностей ошибки $p_1 \times p_2 > 0$ 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:* Конструкция должна состоять из формальной математической модели внешнего мира (World Model), спецификации безопасности и детерминированного «привратника» (Gatekeeper). Он 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 апреля 2026 года программой руководит Нора Амманн, Далримпл — технический советник), бюджет в размере 59 млн фунтов стерлингов. £59M. 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 кортежа к прикладным защитным экранам для опасной инфраструктуры, по данным обзора LongtermWiki — вторичный источник, подлежит сверке) обнаружилось, что построение полных вероятност-

ных моделей всего физического и социального мира. Задача горизонта многих лет. *Точка схождения и базовое отличие:* Конструкция шлюза MITR опирается на ту же самую математическую философию «привратника с формальным доказательством», что и школа Далримпла—Бенжио. Однако располагается 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 кортеж 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. Поспелов; Metaphorum 2025–2026). *Тезис:* Человек должен оставаться сувереном в периметре управления, но требование «понимать каждый шаг чёрного ящика» нарушает закон необходимого разнообразия Эшби (материалы конференции Metaphorum 2025–2026 / MDPI Systems, 2025). Выход — конструктивная рекурсия модели жизнеспособной системы Стаффорда Бира (VSM). Там человек находится в позиции S5 (политика и подписание правил). При этом независимый от текущего управления периметр 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 / конструктивный 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 технологии)	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+)

Школа / конструктивный подход	Объект контроля	Математический режим	Устойчивость к теореме Райса	Вычислительная цена проверки
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 контрактах), $0 за токен

6.3.2 Аксиома водораздела и анатомия честного обхода

Инженерный ответ АНО «НИИ Системного синтеза», зафиксированный в реестре продуктовой линейки как аксиома водораздела, проводит жёсткую границу между двумя физически разными средами: средой порождения гипотез и средой исполнения команд.

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 (GitHub API, 14.10.2026, около 392 тыс. звёзд; Neurohive, 18.03.2026, 13 729 навыков)) или самомодифицирующийся Hermes Agent от Nous Research (документация OpenClaw / Nous Research, 02.2026), происходит стохастический поиск в непрерывном матричном 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 мо-

мент 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 наборами) 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 24 октября 2026 года в хранилище правил развёрнуто 20 таких контрактов (на 14 октября — 19; MITR/policies/contracts/, 14.10.2026). При этом действует базовая 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 ⇔ UNSAT): Если решатель Z3 возвращает вердикт UNSAT (невыполнимо), это служит строгим формальным доказательством того, что пересечение параметров этого дей-

ствия с множеством запрещённых состояний пусто ($A(x) \cap \Phi_{\text{forbidden}}(x) = \varnothing$). Ни при одной комбинации переменных в пределах заявленного вызова нарушение подписанного человеком контракта произойти не может. Только после получения доказательства UNSAT шлюз подставляет учётные данные и исполняет вызов.

– Блокировка действия (Block ⇔ 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 рабочем сервере шлюза

Рабочие характеристики этого механизма измерены инструментально на действующем сервере шлюза (подробные протоколы приведены в главах 9 и 10, команды воспроизведения — в приложении Б). Чистое математическое решение задачи в решателе Z3 на обычном центральном процессоре без графических ускорителей занимает в медиане около одной миллисекунды (от 0,74 мс при 17 контрактах до 1,08 мс при 20 контрактах), а полный вызов шлюза с формированием объяснения. От 3,3 до 5,22 мс при нулевых затратах на внешние языковые модели. Каждое решение оформляется криптографической квитанцией в непрерывной хэш-цепи SHA-256 с внешними метками времени OpenTimestamps (проверщик повтора без обращения к коду шлюза воспроизвёл все 5 053 вердикта на срезе 24 октября 2026 года). При этом пополнение самого словаря распознавания действий проходит через ту же формальную верификацию (глава 12).

6.4 Ближний инженерный рубеж: ворота приёмки 2026–2027

Инженерная дисциплина, превращающая книгу в измерительный прибор, требует провести строгую границу между тем, что уже работает как доказанный механизм в коде, и тем, что находится в стадии набора эксплуатационной полноты и статистической выдержки. В научном периметре института действует жёсткое правило: 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» периметров там, где хотя бы один машинный критерий ещё не набрал требуемой выдержки (10.2026).

Четыре эксплуатационных рубежа четвёртого квартала 2026 года, подробно описанные в разделе 1.4 и проверяемые командами приложения Б, последовательно замыкают периметр промышленной надёжности. Удержание режима автоматического запрета по всем двадцати контрактам и шести базовым классам действий на четырнадцатисуточной серии (после первичного достижения отметки «шесть из шести» 23 октября 2026 года), наработка четырнадцати непрерывных суток живого потока от внутренних арендаторов без единого обхода шлюза, сведение доли нераспознанных команд в словаре классификации к нулю и запуск внешнего пилота после успешной приёмки испытательного полигона INDRA 25 октября 2026 года. 12 Критериев из 12 на внутреннем стенде из 28 сценариев.

Ни одна из этих строящихся метрик не меняет базовой конструкции 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 долгосрочный план, рассчитанный на десятилетие вперёд, сталкивается с базовой синергетической неопределённостью: слож-

ные нелинейные системы проходят через точки бифуркации (И. Р. Пригожин; Г. Хакен), в которых точечное угадывание дат бессмысленно. Именно поэтому в методологии книги прогнозирование горизонта 2035 года устроено не как литературное пророчество, а как расщепление траектории на три сценарные ветви (A, B, C) 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 (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 конструкция «честного обхода» теоремы Райса через шлюз действий в каждой из трёх ветвей будущего к 2035 году:

Ветвь A: «Нормальная технология» (замедление фронтира и долгая институциональная диффузия)

- Сценарная посылка (базовые параметры разобраны в разделе 1.5). Темпы отдачи от масштабирования вычислений замедляются, диффузия технологии растягивается на десятилетие, а значительная часть корпоративных агентных начинаний сворачивается к концу 2027 года из-за неокупаемости и слабой управляемости рисков.
- Судьба контроля и роли человека: В ветви А сверхинтеллекта нет. Однако по корпоративным сетям работают миллионы заурядных, дешёвых открытых агентов (чьи веса уже к 2030 году стано-

вятся бесплатным встроенным свойством любой операционной системы (прогноз: ветвь A/B; 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 A; репер: Stanford Digital Economy Lab, ежегодно).

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

- Сценарная посылка (см. раздел 1.5): сохраняются текущие темпы роста вычислений и десятикратного ежегодного удешевления токена, горизонт автономного выполнения задач к 2027–2030 годам выходит на масштаб квартального проекта, а на рубеже 2028–2030 годов появляется автоматизированный исследователь-программист.
- Судьба контроля и роли человека: В ветви B поток действий агентов многократно превышает пропускную способность человеческого внимания. Ручное подтверждение перестаёт удерживать поток вызовов, и по закону необходимого разнообразия Эшби компании, пытающиеся контролировать агентов людьми-модераторами, разоряются на фонде оплаты труда или гибнут от пропущенных инцидентов. Выживают только те структуры, где совершён 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 %. При этом 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 B; репер: ежегодный обзор законодательства об ИИ, 2027–2035)). Сегмент защитных шлюзов и систем надзора (guardian agents) занимает 10–15 % рынка агентного ИИ к 2030 году и удерживает 15 % к 2035 году (прогноз: ветвь B; репер: аналитика Gartner по guardian agents, 2030).

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

- Сценарная посылка (см. раздел 1.5): до 2028 года происходит полная автоматизация цикла разработки искусственного интеллекта, запускающая стремительный технологический рост, ограниченный лишь физическим лимитом выработки электроэнергии.
- Судьба контроля и роли человека: В ветви C 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 как архитектурныйпроектный замысел становится единственным рубежом, отделяющим человечество от полной утраты субъектности (Gradual Disempowerment). Каким бы сверхсложным ни был интеллект модели в ветви C, 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 (Z3), то даже сверхинтеллект не может обойти тождества формальной логики или подделать результат проверки UNSAT, не решив задачу обращения криптографического хэша SHA-256. В межгосударственном периметре (доктрина взаимного сдерживания MAIM — Д. Хендрикс, Э. Шмидт, А. Ван (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 решающее свойство построенной конструкции: детерминированный шлюз канала действия инвариантен к тому, какой именно сценарий будущего победит к 2035 году.

- Если сбудется ветвь A — шлюз защищает компанию от рутинных ошибок дешёвых моделей и закрывает требования аудита без затрат на GPU.
- Если сбудется ветвь B — 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 C — шлюз остаётся надёжным формальным барьером суверенитета человека (S5) перед системой, чью мысль человек уже не способен постичь.

6.6 Экономика доказанного действия и периметр монетизации

Выбор конструктивной школы контроля напрямую определяет финансовую жизнеспособность предприятия и структуру монетизации защитного продукта. 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). При этом 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

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 (исследование Т1, 03.2026; «Коммерсантъ», 20.07.2026)) и добавляет в рабочий цикл задержку от сотен миллисекунд до нескольких секунд.

В конструкции детерминированного шлюза MITR экономика проверки опирается на твёрдые измеренные факты:

- Время вычисления доказательства UNSAT / SAT в решателе Z3. 0,74 Мс. Медиана по 26.09.2026; на рабочем сервере шлюза — 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 (26.09.2026; на рабочем сервере шлюза 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 (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:

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

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

В соответствии с дисциплиной правды (§8), текущая коммерческая позиция фиксируется прямо: по состоянию на октябрь 2026 года коммерческой выручки у шлюза MITR нет (программа исследования, 14.10.2026). Продвижение по коммерческой лестнице управляется исключительно закрытием машинных критериев на рабочем сервере шлюза, а все целевые модели выручки имеют строгий статусгипотеза до подписания первого оплаченного акта внешним заказчиком:

1. Ступень 1 («Свои агенты» — в работе): Внутренние периметры института и служебный агент. 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 автора; от неё в арсенале остаётся отработанный механизм теневой квитанции.
2. Ступень 3 («Первый внешний арендатор» —гипотеза по выручке, ворота X5): Целевой сегмент — команды разработки, использующие кодовых и локальных агентов (кодовые и операционные агенты) через протоколы MCP и REST. Именно этот сегмент в 2026 году острее всего испытывает боль от ин-

цидентов уровня PocketOS и открытых серверов MCP (среди 1 412 публично доступных серверов с определимым способом аутентификации 38,7 % работают вообще без неё (данные сканирования Bloomberry, февраль 2026 года)). 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 периметра института с собственным токеном на протяжении 14 непрерывных суток.

3. Ступень 3а («Канал страховщика» — новый канал v4): В 2026 году мировой страховой рынок совершил структурный поворот. С января ISO ввела факультативные оговорки об исключении генеративного ИИ из полисов CGL. Они крупные страховщики начали прикреплять к полисам (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 (Armilla AI, 21.01.2026) и отраслевой стандарт AIUC-1, по которому компания ElevenLabs первой получила полис на действия голосовых агентов (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)) служит готовым техническим артефактом 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 — профильные российские страховщики ответственности), при котором наличие шлюза даёт предприятию скидку на страховой полис. При этом разработчик шлюза получает долю страховой премии или фиксированную плату за аттестацию журнала.
4. Ступень 4 («Грант и инвестор» —гипотеза): Привлечение грантового финансирования (грантовое финансирование перехода от лабораторного образца к опытному периметру 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, а P1, прогон атак с внешней подписью и P5, пилот ≥ 30 суток находятся в графике стройки.

5. Ступень 5 («Регулируемые отрасли и государственный периметр» — по ГИС,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 (ComNews / SecurityLab, 08.2026)). Он 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 (официальный текст на cac.gov.cn, 08.05.2026; CSA Research Note, 26.08.2026)) и европейским регламентом Regulation (EU) 2026/1744 (Annex III с 02.12.2027 (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)».

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.

6.7.1 Верификация на рабочем сервере шлюза при наличии аудиторского ключа

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

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

Проверка инструментальных замеров времени решения SMT-задачи (P4: 0,74 мс решатель / 3,3 мс вызов на 26.09.2026) и последнего

криптографического якоря в цепи квитанций (P3):

Проверка состояния словаря классификации действий (C1):

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 рабочему серверу шлюза открыт набор проверки (asi5.ru/console/verify/: якоря цепей, метки OpenTimestamps, проверщик целости 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:

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 локальный экземпляр Z3. Контрольный прогон 24.10.2026 подтвердил все 5 053 решения из 5 053 на 24.10.2026 (на 14.10.2026 — 4 958 из 4 958) без обращения к коду MITR.

6.7.3 Процедура сверки внешних сценарных реперов

Для проверки того, по какой из трёх ветвей (A, B или C) 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 A. Выход на горизонт недели–месяца к 2027–2028 годам — на ветвь B.
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 A, реализация до 2028 года — ветвь C.

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	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
A-6.01	Вычисления при обучении фронтирных моделей растут ≈ ×5 в год с 2020 г. (удвоение за 5,2 мес.)	Измерено	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} ≈ 12 ч; Claude Mythos Preview (05.2026) — за 16-часовым пределом надёжности набора	Измерено	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 эталоном
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 B/C — пересматривается автором
A-6.04	Модели с открытыми весами отстают от закрытого фронтира менее чем на 1 год (в среднем ≈ 4 мес. с января 2026 г. по индексу Epoch)	Измерено	International AI Safety Report (03.02.2026); Epoch AI, Data Insight «open–closed ECI gap» (29.05.2026)	02.2026 – 05.2026	Ежегодный отчёт IAISR; обновления 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)	Измерено	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

№ п/п	Утверждение в тексте главы	Метка	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
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	Измерено	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 разработчиков (признано OpenAI); 13-часовой простой AWS Cost Explorer в КНР приписан агенту Kiro (FT), Amazon оспаривает	Измерено	AI Incident Database № 1469 (PocketOS) и № 1442 (Kiro); заявление OpenAI 16.07.2026; разбор Lumenova (25.08.2026, только PocketOS) и обзор Aviatrix	12.2025 – 08.2026	Первичные карточки инцидентов в AIID
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	Измерено	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 на внутреннем стенде 460 действий прошли мимо клиентского хука	Измерено	отраслевой обзор института, 40-я неделя 2026 г. (70 записей); журнал сессий внутреннего стенда	25.09–октябрь 2026	Архив отраслевого обзора и
A-6.10	Программа ARIA Safeguarded AI (запущена Д. Далримплом; с 04.2026 руководитель — Н. Амманн, Далримпл — технический советник) финансируется в объёме £59 млн для построения формальных привратников и мировых моделей	Измерено	Страница программы aria.org.uk (04.2026); документация вызова ARIA TA2 Phase 2; Substack ARIA Research	2025 – 04.2026	Сверить пересмотр программы 11.2025 (отказ от TA2 Phase 2 по данным LongtermWiki, вторичный источник) с официальными документами ARIA
A-6.11	Аксиома водораздела: переход от сколь угодно сложной мысли к действию есть конечный вызов, проверяемый логикой первого порядка за миллисекунды	Замысел (для AGI) / Измерено (механизм)	реестр продуктовой линейки, модуль core/verification/z3_gate.py	14.10.2026	Механизм подтверждён на рабочем сервере шлюза; часть AGI остаётся замыслом до появления AGI

№ п/п	Утверждение в тексте главы	Метка	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
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 рабочем сервере шлюза — 0,95 и 4,3 мс (14.10.2026, 19 контрактов) и 1,08 и 5,22 мс (24.10.2026, 20 контрактов)	Измерено	скрипт smt_latency_paper.py, лог /opt/v/data/logs/smt_latency_paper.jsonl на рабочем сервере шлюза	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 сценариев состязательного набора стоит $0 на модели; расход на модели за 1–14 октября 2026 г., $0,30	Измерено	Каталог MITR/policies/contracts/, журнал расхода рабочего сервера шлюза	14.10.2026	Сверка с деревом файлов репозитория
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), 5 053 из 5 053 (24.10.2026) (P3)	Измерено	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 % (клиентский перехватчик первой выборки, 16–26.09), 32 % (второй внутренний периметр, 09.10)	Измерено	(квитанция № 305), gate_mode_report.py --tenants	25.09–09.10.2026	Погашается при снижении доли нераспознанных до 0 %
A-6.16	К 23–24.10.2026 все 20 контрактов и 6 основных классов переведены в режим «запрет» (за 23.10 — «правда 6 из 6», за 24.10, «правда 5 из 6, не измерено 1»; 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)	Измерено	Скрипт gate_mode_report.py --six и --tenants с ключом --date 2026-10-24 (на 09.10.2026 — с ключом --date 2026-10-09) на рабочем сервере шлюза	09.10–24.10.2026	Погашается закрытием 14-суточной серии P2a и удержания --six
A-6.17	Закрытие серии P2a (14 из 14 суток) и удержание метрики «шесть из шести» (--six) в течение 14 суток	Строим	Ворота приёмки: gate_mode_report.py --tenants и --six на рабочем сервере шлюза	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)	Строим	Ворота приёмки: артефакты kit-acceptance.py и протокол пилота (25.10.2026 приёмка на внутреннем стенде 12 из 12; на 30.09.2026 — 0 из 5)	Срок: 31.10.2026 (A–C) и 12.11–15.11.2026 (D–E)	Приёмка артефактов внутренней комиссией института

№ п/п	Утверждение в тексте главы	Метка	Прибор, скрипт или источник	Дата замера / публикации	Условие погашения / сверки
A-6.19	Подписочная выручка от внешних команд разработки (ступень 3), партнёрская премия со страховщиками (ступень 3a) и лицензии для ГИС/КИИ (ступень 5)	Гипотеза	Коммерческая лестница (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
Г3	Медианный горизонт задач лучшей модели (p_{50})	≈ 12 ч у Opus 4.6 (METR, 20.02.2026); Mythos Preview (05.2026) — свыше 16 ч, за пределом надёжности набора	A: месяц / B: квартал–год («проект полного цикла»; при удвоении раз в 3–7 мес. достигается уже к 2027–2030 г.; выше 16 ч замер METR ненадёжен) / C: неизмеримо текущим набором	Публичные отчёты METR time-horizons (metr.org)	Ежегодно (2027–2035)
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.)	A: после 2035 г. / B: 2028–2030 гг. / C: 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)	A: 10 % / B: 30 % / C: > 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)	A: 5 % / B: 15 % / C: 25 %	Пресс-релизы и отчёты Gartner по guardian agents	Контрольная сверка — 2030 и 2035 гг.
Г9	Разрыв в занятости специалистов 22–25 лет в наиболее подверженных ИИ профессиях	−19 % (Stanford DEL, «Canaries in the Coal Mine?», ред. 12.08.2026)	A: 0 % (восстановление) / B: −20 % / C: −40 %	Публикации Stanford Digital Economy Lab (Canaries Dashboard)	Ежегодно (2027–2035)
Г10	Признание самостоятельной правосубъектности за ИИ-агентом	Ни в одной стране мира (Mercatus, обзор 2026 г.)	A: нет / B: нет (действуют режимы уровней полномочий, ответственность на человеке S5) / C: ограниченная (спорно)	Национальные законодательства и реестры правовых актов	Ежегодно (2027–2035)

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
Г13	Доля полисов страхования ответственности за агентов, требующих верифицируемый журнал исполнения	Первые специализированные стандарты (AIUC-1, Armilla AI, 2026 г.)	A: узкая ниша / B: стандарт корпоративного рынка / C: 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, повтора решений, открытого набора проверки и контроля режимов.
4. Конструктивные решения проекта: (набор из 157 кейсов за $0; на 14.10 — 62 кейса; квитанция тени; 98 суток монитора честности; словарь как данные и квитанция № 305; опровержение факта в цепи; повтор решений посторонним; задержка решателя 0,74 мс и вызова 3,3 мс на 26.09.2026).
5. Стратегические и концептуальные документы: программа развития шлюза, канон проектирования шлюза, конструктивный канон шлюза, реестр продуктовой линейки, меморандум демаркации INDRA–MITR от 27.09.2026, эталон разбора аргумента Ямпольского аналитический разбор аргумента о неконтролируемости от 03.10.2026.
6. Каталог научной библиотеки института каталог научной библиотеки института: Н. Винер, Л. А. Растригин, И. Р. Пригожин, Д. А. Поспелов, У. Р. Эшби, Г. Хакен, Г. С. Альтшуллер, общая теория систем.

Внешние источники (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/2025/04/29/1115928/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

1. 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/

1. Рынок труда и постепенное обессиливание:

– 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

1. 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/agentic-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

ISS

Приложения

Приложение Б. Команды независимого повтора

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

```
systemctl is-active v-bot v-mcp
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --six --date
  2026-10-23
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --six --date
  2026-10-24
cd /opt/v && sudo -u v venv/bin/python scripts/gate_mode_report.py --tenants --
  date 2026-10-24
sed -n 2p /opt/v/data/logs/smt_latency_paper.jsonl; tail -1
  /opt/v/data/logs/smt_latency_paper.jsonl
tail -1 /opt/v/data/gate_anchors.jsonl
cd /opt/v && sudo -u v venv/bin/python scripts/verify_chain_standalone.py
  data/logs/gate_actions.jsonl --anchors data/gate_anchors.jsonl --replay
grep -m1 'seq.: 305,' /opt/v/data/logs/gate_actions.jsonl
```

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