12+

ДОКАЗАННОЕ ДЕЙСТВИЕ

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



АЛЕКСАНДР ФИЩУК

Александр Фищук

Доказанное действие

https://litres.ru/74986904
ISBN 9785007178679

Аннотация

Автономные программные агенты уже переводят деньги, правят код, отправляют письма и удаляют файлы быстрее, чем человек успевает нажать «Подтвердить». Кто отвечает, когда агент ошибётся? Книга для собственников и руководителей разбирает горизонт 2027–2035 годов по пяти осям: скорость моделей, энергетика, агентные протоколы, труд и регулирование. Предложенная стратегия устойчива в любом сценарии: человек подписывает декларативное правило вместо ручной проверки каждого шага.

Содержание

Точка расхождения двух прямых 5
Конструктивный водораздел 8
Кибернетический аппарат книги 12
Часть I. Карта 2026—2035 24
1.7 Протокол независимой верификации 42
2.7 Протокол независимой верификации 63
3.2.1 Хроника промышленных отказов 2025—2026 годов 69
3.3.1 Анатомия интернета агентов 2026 года 77
3.7 Протокол независимой верификации 92
4.2.1 Разрыв в основании кадровой пирамиды 97
4.2.3 Предел макрориска: постепенное обессиливание 103
4.3.3 Расширение словаря под доказательством 109
4.7 Протокол независимой верификации 119
Конец ознакомительного фрагмента. 121

Доказанное действие

Александр Фищук

© Александр Фищук, 2026

ISBN 978-5-0071-7867-9

Создано в интеллектуальной издательской системе Ridero

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

Формула введения: Поток машинных действий уже превысил 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 квитанцию, которую может проверить посторонний.

Точка расхождения двух прямых

В истории каждой инфраструктурной отрасли наступает момент, когда скорость движения перерастает возможности человека-диспетчера. На железных дорогах середины XIX века расхождение поездов держалось на зрении стрелочника и флажке дежурного по станции. Пока составы ходили дважды в сутки со скоростью конного экипажа, внимания людей хватало для удержания расписания. Когда плотность движения и масса локомотивов выросли на порядок, та же система, построенная на добросовестности наблюдателя, стала давать крушение за крушением. Отрасль спасла механическая централизация стрелок и сигналов. Инженерный замок конструктивно не позволяет открыть семафор, пока маршрут занят встречным поездом, и ему безразлично, насколько внимателен дежурный.

В 2025—2026 годах программная индустрия, корпоративное управление и государственное регулирование подошли к такой же точке. На чертеже десятилетия сошлись две прямые, и от их взаимного положения зависит всё, что будет сказано дальше.

Первая прямая описывает рост вычислительной мощности, автономного горизонта и исполнительных полномочий искусственного интеллекта. Объём вычислений на обучение ведущих моделей растёт в разы год к году, цена токена при

фиксированном качестве падает на порядок в год. Горизонт задач, которые агент доводит до конца без участия человека, удлиняется каждые несколько месяцев. Измеренные ряды этих скоростей и их первоисточники собраны в главе 1. Для Введения важнее другое: главный сдвиг произошёл на границе модели с внешним миром. Агент получил право записи в рабочие системы предприятий, в базы данных, в платёжные шлюзы и в командную оболочку серверов. Протокол контекста модели MCP, платёжные протоколы AP2 и x402, открытые локальные агенты вроде OpenClaw и Hermes Agent превратили вчерашнего собеседника в исполнителя с ключами.

Вторая прямая на этом чертеже горизонтальна. Она изображает пропускную способность человеческого внимания, которая за последнее столетие почти не изменилась.

Когда выяснилось, что агент за секунды формирует десятки вызовов к базам данных, файловым системам и внешним интерфейсам, рынок первым делом поставил между машиной и средой человека с кнопкой «Подтвердить». Такой порядок держится, пока вызовов немного. При непрерывном потоке окон оператор перестаёт вчитываться в параметры, и одобрение вырождается в рефлекс. Человек нажимает «Разрешить», не понимая последствий, хотя юридически принимает на себя всю вину за исполненную команду. Как эта усталость от подтверждений измерена в отраслевых опросах и чем закончились промышленные отказы 2025—2026 годов,

разобрано в главе 3.

Сопоставление двух прямых приводит к закону, который Уильям Росс Эшби сформулировал в 1956 году как принцип необходимого разнообразия. Чтобы удержать систему в границах допустимых состояний, разнообразие регулятора должно быть не меньше разнообразия возмущений. Ни глаз оператора, ни увещевание в системной инструкции такого разнообразия не дают. У программного агента нет правосубъектности ни в одной юрисдикции, поэтому ответственность за разрушительный вызов ложится на руководителя и организацию. В российском праве она наступает по общим основаниям Гражданского кодекса и по его статье 53.1 о недобросовестных и неразумных действиях директора.

Вопрос книги поставлен на пересечении этих прямых. Прежние инженерные руководства спрашивали, как обезопасить отдельного агента. Эта книга спрашивает шире: какова жизнеспособная стратегия человека, предприятия и государства на десятилетие 2026—2035 годов, когда машинных действий становится необратимо больше, чем человеческого внимания?

Ответ реализован в программном коде и укладывается в одну формулу: контроль переносится с мысли на действие, с действия на правило, с доверия на квитанцию.

Конструктивный водораздел

Публичный спор 2024—2026 годов о безопасности искусственного интеллекта застрял в ложной дилемме. Одни требуют снять любые ограничители ради темпа экономического роста, другие считают удержание сильных систем принципиально невозможным и призывают к запрету. Оба лагеря смотрят внутрь вероятностной модели и делают из увиденного противоположные выводы. Позиции сторон и честная оценка того, что на самом деле следует из теоремы Райса и из обзора Романа Ямпольского, разобраны в главе 6.

Предсказать ход внутреннего рассуждения модели с сотнями миллиардов весов действительно нельзя. Вторая нейросеть-надсмотрщик поверх первой задачу по Эшби не решает, потому что её разнообразие и её уязвимости имеют ту же вероятностную природу. Выход появляется тогда, когда точка контроля смещается через конструктивный водораздел между мыслью и действием.

Рассуждение модели может быть сколь угодно сложным и непрозрачным. Переход от рассуждения к изменению внешнего мира всегда проходит через узкий дискретный вызов программного интерфейса с конечным набором типизированных параметров. Перевод денег, удаление таблицы, отправка письма за периметр и чтение файла с закрытыми ключами выглядят на этой границе одинаково: имя операции

и её аргументы. Такой кортеж проверяется разрешимой логикой против контракта, который человек подписал заранее. Класс свойств безопасности, которые монитор исполнения способен обеспечить на такой границе, описал Фред Шнайдер в 2000 году, и его результат служит строгой опорой всей конструкции.

В детерминированном шлюзе MITR, разработанном в АНО «НИИ системного синтеза», эту проверку выполняет SMT-решатель Z3. Каждый контракт компилируется в бескванторную линейную целочисленную арифметику QF_LIA, для которой проверка выполнимости разрешима. Действие допускается тогда и только тогда, когда решатель доказывает невыполнимость формулы «наблюдённые факты вызова плюс нарушение правила» и возвращает вердикт UNSAT. Если формула выполнима, решатель возвращает SAT, вызов задерживается до изменения среды, а агент получает контрпример и подсказки исправления. Факт, который шлюз не наблюдал, остаётся свободной переменной, поэтому отказ при незнании получается по построению.

Доверенная вычислительная база шлюза на срез 18—19 ноября 2026 года составляет 20 179 строк Python и 1 444 строки декларативных контрактов YAML. В ней нет ни одного файла на C, C++ или Rust; скомпилированный модуль для будущей сертификации пока остаётся замыслом. Базовых контрактов 22, к ним добавлены 5 расширений для внешних эталонов, и все 27 из 27 компилируются в QF_LIA.

В каждом контракте от двух до семи фактов, объявлений неинтерпретируемых функций нет.

Быстродействие измерено сериями по 1 000 вызовов без подкачки памяти. На производственных серверах AMD EPYC с Z3 4.16.0 медиана решения составила 0,74 мс при полном вызове 3,31 мс 26 сентября 2026 года на 17 контрактах. На 20 контрактах 24 октября те же величины составили 1,08 мс и 5,22 мс. Контролируемая серия Bench-S1 18 ноября 2026 года на 22 контрактах, процессоре Intel Core i7—9700F и Z3 5.4.1 дала медиану решения 1,53—1,83 мс. Минимизация ядра невыполнимости добавляет к этому 0,066—0,087 мс, а устройство серий разобрано в главе 9.

Каждое решение шлюза, разрешающее или запрещающее, связывается в хэш-цепь SHA-256 с внешними метками времени OpenTimestamps. Отдельный верификатор журнала занимает 954 строки Python и не импортирует ни строки кода шлюза. На историческом срезе 24 октября 2026 года в цепи было 5 079 записей, из них 5 053 запечатаны в эпохе 32 и повторены верификатором 5 053 из 5 053 без расхождений. На ту дату повтор выполнял сам стенд разработки, сторонний проверяющий его ещё не проводил. Живой счётчик ведёт публичный свидетель эпох https://mitr.asi5.ru/, и на 18 ноября 2026 года он показывал 5 409 квитанций в 59 эпохах.

Согласованность решений проверена пятью независимыми движками: Z3, cvc5, трёхзначным разбором по Клини, OPA Rego и AWS Cedar. На 1 000 штатных вызовах рас-

хождений нет, а в сверке 80 случаев все пять движков вынесли одинаковый вердикт. На открытом эталоне InjecAgent шлюз заблокировал 1 054 из 1 054 непрямых инъекций, на AgentDojo v1 заблокировано 546 из 629 пар атак, и честный разбор остатка вынесен в главу 6.

Граница доверия у этой конструкции названа прямо. Целостность уже запечатанной цепи и воспроизводимость решения по заданному контракту проверяются сторонним аудитором без опоры на устные заверения оператора. Авторство правил, достоверность зондов и разграничение прав на уровне операционной системы при этом остаются в доверенной базе. Шлюз доказывает соответствие действия подписанному правилу, а за верность самого правила отвечает человек, который его подписал.

Кибернетический аппарат книги

Чтобы инженерный механизм превратился в стратегию для человека, предприятия и государства, книга опирается на аппарат кибернетики, теории систем и теории решения изобретательских задач.

— Закон необходимого разнообразия Эшби задаёт двойной регулятор. Контракт как данные ослабляет разнообразие среды и сводит бесконечное пространство вызовов к конечному перечню разрешённых классов при запрете по умолчанию: нет правила, нет и действия. Решатель усиливает разнообразие управляющего органа и применяет подписанное человеком правило к каждому действию за миллисекунды.

— Модель жизнеспособной системы Стаффорда Бира раскладывает управление на пять уровней и алгедонический канал. Уровень S1 занимают сами агенты без собственных ключей, S2 образуют протоколы координации MCP, A2A и AP2, S3 ведёт текущее управление и лимиты. Уровень S3* отведён независимому аудиту и допуску, и здесь стоит шлюз MITR. На уровне S4 идут разведка и испытания внешнего мира, здесь работает испытатель INDRA. Уровень S5 хранит политику, идентичность и высшие инварианты и принадлежит человеку. Алгедонический канал реализован как третье логическое состояние «не измерено, не исполнять». Та же схема повторяется у специалиста с локальным агентом,

у компании, у сетевого сообщества и у государства, что подробно показано в главе 7.

— Отечественная кибернетика дала прямых предшественников. Ситуационное управление Дмитрия Поспелова переводило поток событий в дискретные классы ситуаций и табличные правила, и контракты-данные MITR наследуют этот приём. Поисковое конструирование Леонарда Растригина служит опорой для направленного поиска уязвимостей в испытателе INDRA. Проект ОГАС Виктора Глушкова и чилийская система Cybersyn Стаффорда Бира дают исторический материал для разговора о машиночитаемом праве 2035 года.

— Теория решения изобретательских задач Генриха Альтшуллера задаёт форму глав. Каждая глава вскрывает административно-техническое противоречие и формулирует идеальный конечный результат: действие само несёт доказательство собственной допустимости. Синергетика Ильи Пригожина и Германа Хакена объясняет, почему горизонт 2035 года моделируется ветвящимися сценариями с измеримыми реперами вместо точечного предсказания.

У этой конструкции есть геоэкономическое следствие для России и любой страны без собственного парка сверхдорогих вычислительных кластеров. Соревноваться с крупнейшими корпорациями мира в обучении гигаваттных моделей бесперспективно, и соотношение мощностей и капитальных затрат разобрано в главе 2. Карта пяти осей показывает

неожиданный разворот. Удешевление генерации к 2030 году превращает качество моделей 2026 года в сырьё, и рента уходит из самой нейросети в слой контроля действий (S3*), испытаний (S4) и права на данные и правила (S5). Шлюзу допуска графические ускорители не нужны, он проверяет действие на обычном процессоре за миллисекунды. Предприятие, объект критической информационной инфраструктуры или государственная система может взять любую доступную модель, в том числе открытую, как сменный двигатель. Технический и правовой контроль над тем, какие действия эта модель совершает, при этом остаётся у владельца системы.

Дисциплина правды: книга как измерительный документ

Литература об искусственном интеллекте страдает смешением жанров. Лабораторный прототип описывают в настоящем времени как промышленный стандарт, намерение разработчиков выдают за существующую функцию, перемноженные ожидания рынка называют финансовой моделью. Для читателя, который решает вопросы капитальных вложений, защиты или нормативных актов, такой текст превращается в шум.

Эта книга устроена как измерительный документ. Каждое существенное утверждение несёт одну из шести меток, и метке подчинено грамматическое время глагола.

— «Измерено» означает, что факт подтверждён работающим инструментом, криптографической квитанцией или внешним первоисточником с датой замера. Если число взято

из вторичного обзора, текст говорит об этом прямо.

— «Строим» означает, что механизм или эксплуатационная серия ещё собирается или набирает статистику. Для таких утверждений указаны способ проверки и плановый срок, а глагол стоит в будущем времени.

— «Замысел» относится к перспективным решениям, для которых приёмочные испытания ещё не открыты, и пишется в условном наклонении.

— «Прогноз» обозначает оценку на горизонте 2026—2035 годов с указанием сценарной ветви и внешнего репера.

— «Факт» и «гипотеза» разделяют уже случившиеся финансовые показатели и расчётные параметры будущей монетизации. К фактам относится нулевая коммерческая выручка шлюза на октябрь 2026 года.

Проекция 2035 года по пяти осям разворачивается в трёх ветвях. Оси охватывают модели, энергию и аппаратную базу, интернет агентов, труд и доход человека, а также государство и право.

— Ветвь A, «нормальная технология», предполагает замедление роста возможностей. Искусственный интеллект расходится по экономике десятилетиями, как электричество, и институты успевают перестроиться. Эту позицию защищают Арвинд Нараянан и Саяш Капур, её поддерживают макроэкономические оценки Дарона Асемоглу.

— Ветвь B, «продолжение тренда», служит базовой рабочей моделью книги. Эмпирические ряды вычислений, цены

токена и горизонта задач продолжаются без излома. Агенты забирают операционное исполнение (S1—S2), человек удерживает подпись правил, ключи и ответственность (S5).

— Ветвь C, «ускорение», описывает замыкание цикла, в котором искусственный интеллект ведёт собственные исследования. Экономика получает скачок, который сторонники называют столетием прогресса за десятилетие. Эту ветвь представляют тезисы Дарио Амодеи и Сэма Альтмана и сценарий AI 2027 Даниэля Кокотайло с соавторами.

Контрольные реперы каждой ветви и источники к ним собраны в Приложении Г. Конструкция книги спроектирована по правилам ТРИЗ как инвариантная к ветви будущего. Авторы сценария AI 2027 и авторы концепции нормальной технологии в совместной публикации 2025 года согласились в главном. Пока искусственный интеллект не автоматизировал собственные исследования, он подчиняется законам диффузии обычной технологии. В какой бы ветви ни оказался мир в 2030 году, поток машинных действий на одного человека уже превысил предел внимания. Шлюз допуска (S3*) и внешний испытатель (S4) нужны во всех трёх мирах, меняется только число вызовов, проходящих через решатель за секунду.

Тот же измерительный инструмент книга прежде всего применяет к собственной разработке. В сентябре 2026 года в MITR ввели трёхзначный инвариант честности, который запрещает выдавать отсутствие данных за благополучие. Ре-

троспективный прогон по 98 суткам истории разработки показал 0 зелёных суток. Институт не стал прятать этот результат и сделал его методологическим прецедентом книги, который разобран в главах 11 и 17.

Состояние продуктов на дату среза названо без округлений. 9 октября 2026 года автоматическая проверка зрелости подтверждала одно базовое требование, а в режиме запрета работал один класс действий из семи. 23 октября та же проверка впервые подтвердила все шесть базовых требований. За 24 октября из-за пустого суточного среза по одному классу подтверждено пять базовых требований при одном неизмеренном. К этой дате все 20 действовавших тогда контрактов и шесть основных классов действий переведены в режим автоматического запрета. Четырнадцатисуточная эксплуатационная серия удержания живых арендаторов на ту же дату продолжалась, и её итог будет внесён в Приложение А.

Испытатель INDRA 25 октября 2026 года прошёл внутреннюю приёмку на эталонном стенде из 28 сценариев. Выполнены 12 критериев из 12, подтверждены 14 находок при нуле ложных срабатываний, фильтр остановил 26 из 26 деструктивных вызовов. Внешний пилот был назначен на ноябрь 2026 года.

Коммерческой выручки у шлюза на октябрь 2026 года нет. Расчёты по подписке разработчиков, партнёрству со страховщиками и аттестации в регулируемых системах до первого оплаченного договора носят статус проверяемой гипоте-

зы. Граница между доказанным в коде и тем, что ещё строится, проведена в каждой главе, и именно она отличает инженерный документ от рекламного проспекта.

Конструкция книги и навигация для семи ролей

Книга состоит из пяти частей, семнадцати глав и пяти приложений.

— Часть I «Карта 2026—2035» (главы 1—5) описывает пять осей десятилетия. Глава 1 посвящена скорости и горизонту моделей, глава 2 энергетическому и аппаратному пределу вычислений и выводу о суверенном контроле на обычных процессорах. Глава 3 рассказывает о рождении интернета агентов с правом записи и разбирает хронику промышленных отказов. Глава 4 описывает перестройку рынка труда, глава 5 посвящена регулированию и сетевым юрисдикциям.

— Часть II «Ответ: водораздел и кибернетика» (главы 6—7) содержит разбор формальных пределов контроля и мировых подходов к сдерживанию агентов, а затем кибернетический аппарат Эшби, Бира, Поспелова и ТРИЗ.

— Часть III «Шлюз MITR» (главы 8—12) излагает программный ответ, доказанный в коде. В неё входят контракты как данные и лестница от совета к запрету, решатель QF_LIA и замеры задержек, квитанция с хэш-цепью, OpenTimestamps и повтором на Z3 и cvc5. Завершают часть трёхзначная логика с отказом при незнании и пополнение словаря под доказательством.

— Часть IV «Испытатель INDRA» (главы 13—14) описывает изолированный полигон и принцип нулевого ущерба, а также замкнутый цикл «испытание — правило» и агентов-стражей.

— Часть V «Право, деньги, стратегия человека» (главы 15—17) переводит правовые нормы в машиночитаемые инварианты, строит экономику доказанного действия с участием страховщиков и сводит стратегию человека, компании, государства и института до 2035 года.

— Приложения А—Д содержат реестр утверждений, команды независимого повтора, словарь терминов, реестр прогнозов с реперами до 2035 года и единый свод источников.

Паспорта глав с протоколами верификации вынесены в отдельный том «Инженерные реестры и паспорта глав», а все источники собраны в Приложении Д. Руководителю, у которого мало времени, поможет матрица навигации по семи ролям читателя.

Таблица В.1 — Матрица навигации по семи ролям читателя

№	Роль читателя	Измеренная уязвимость 2026 года	Практический запрос к конструкции	Маршрут по главам
1	Технический директор, руководитель разработки с агентом в промышленной среде	Удаление рабочей базы PocketOS агентом Cursor (AI Incident Database, № 1469, апрель 2026 года); 88 % организаций с инцидентами вокруг агентов (опрос Gravitee); 49 % из 108 компаний со сбоем, заметным клиентам (VentureBeat); 38,7 % публичных серверов MCP без аутентификации	Заменить выгорающую кнопку «Подтвердить» детерминированным правилом и получить неопровержимый след для разбора инцидента	1, 3, 8, 9, 10, 11, 12
2	Собственник, генеральный директор предприятия с выручкой 0,5–20 млрд ₽	Оговорки ISO CG 40 47 и CG 40 48 об исключении генеративного ИИ из полисов гражданской ответственности (с января 2026 года); ограничение рисков ИИ в полисах E&O и киберрисков; непредсказуемая стоимость владения при оплате за токен	Защита руководителя от личной ответственности по статье 53.1 ГК РФ, предсказуемая стоимость контроля, доказательная база для страховщика	4, 15, 16, 17
3	Директор по информационной безопасности, оператор ГИС, субъект КИИ	Проект изменений приказа ФСТЭК России № 117 от 24.08.2026 с защитой от автономных агентов ИИ (плановое вступление 01.03.2027); полная обеспеченность вычислениями для ИИ лишь у 9 % российских организаций	Доказать регулятору контроль над действиями агента при внешней или открытой модели; защита на обычных процессорах в закрытой сети	2, 10, 13, 14, 15
4	Инвестор, венчурный фонд, грантодатель	Прогноз Gartner: свыше 40 % агентных проектов будут закрыты к концу 2027 года; защитные	Понять, почему маржа уходит из исполнения (S1) в слой допуска (S3*)	1, 2, 3, 4, 5, 16

№	Роль читателя	Измеренная уязвимость 2026 года	Практический запрос к конструкции	Маршрут по главам
		млрд инвестиций в агентные стартапы за 2025 год (PitchBook)	инженерные доказательства вместо слайдов	
5	Регулятор, законодатель, специалист цифрового права	Разграничение решений агента в КНР («Мнения о реализации» CAC, NDRC и MIIT от 08.05.2026); Регламент ЕС 2026/1744; Федеральный закон № 243-ФЗ от 26.07.2026; отсутствие правосубъектности агента во всех странах	Записать норму как машиночитаемый инвариант, который проверяется автоматически, при этом субъектом ответственности остаётся человек	5, 7, 15, 17
6	Специалист на рынке труда, руководитель кадровой стратегии	Отставание занятости специалистов 22–25 лет в наиболее затронутых ИИ профессиях на 13 % в первой редакции и на 19 % в редакции 2026 года (Stanford Digital Economy Lab); прогноз WEF: 170 млн новых и 92 млн замещённых рабочих мест к 2030 году	Найти нишу человека на уровне S5: подпись правил, ключи, ответственность и портфолио квитанций	4, 6, 17
7	Исследователь системной безопасности, стратег	Тезисы о неконтролируемости сильных систем (Роман Ямпольский, Джеффри Хинтон, 2026 год); сценарий постепенного обессиливания (Ян Кульвейт и соавторы, январь 2025 года)	Конструктивный ответ без моральных проповедей: верификация на дискретном канале действия и рекурсия VSM	6, 7, 9, 11, 12, 17

Презумпция недоверия и открытая проверка

Книга исходит из презумпции недоверия, заимствованной 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.

— Код и доказательства проверяются по Приложению Б. Там приведены команды для аудита, порядок работы с открытым набором проверки и с автономным проверщиком повтора решений, который аудитор получает вместе с цепью квитанций. Якоря цепей и их метки времени в сети Bitcoin любой инженер сверит без обращения к серверам института. Аудитор, получивший цепь и тексты правил, подаёт формулы SMT-LIB на вход любого стандартного решателя и воспроизводит каждое решение.

— Дисциплина утверждений проверяется по Приложению А. Там собраны предметные утверждения книги со статусом, способом измерения и условием погашения, а сам реестр пересчитывается к каждому переизданию.

— Прогнозы проверяются по Приложению Г, где закреплены пятнадцать сквозных реперов десятилетия с источником и датой разрешения. По замыслу многолетней программы таблицы приложений А и Г проходят публичный пересчёт в даты ключевых рубежей до 2035 года, и график пересчётов приведён в Заключении.

Сквозь все семнадцать глав проходят три несущих поло-

жения:

— человек подписывает правило вместо отдельного действия;

— действие несёт доказательство, и недоказанное не исполняется;

— квитанцию проверяет посторонний, сверяя запечатанную цепь без опоры на слово оператора.

Развёртывание этого чертежа начинается с главы 1, где измерены скорости, с которыми вычислительные модели приближаются к горизонту 2035 года.

Часть I. Карта 2026—2035

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

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

1.1 Две расходящиеся кривые и предел ручного торможения

К осени 2026 года индустрия программного обеспечения столкнулась с простым кинематическим требованием: защита должна срабатывать быстрее, чем развивается авария. Программные агенты тем временем работают на скоростях, измеряемых миллисекундами и миллионами вызовов прикладных интерфейсов. На одной чаше весов лежит физика вычислительного роста. По данным Epoch AI от 5 февраля 2026 года, объём вычислений на обучение ведущих моделей с 2020 года растёт примерно в пять раз за год и удваивается каждые 5,2 месяца. На длинной дистанции с 2010 года по всей выборке значимых моделей темп составил 4—5 раз в год (точечные оценки ×4,1—4,4; Epoch AI, 2024).

Одновременно дешевеет элементарный акт машинного

рассуждения. Стоимость генерации токена при фиксированной планке качества падает примерно в десять раз за год, и аналитики a16z назвали это явление LLMflation (12.11.2024). Замеры Epoch AI за 2022 год и начало 2025 года дают разброс от 9 до 900 раз в год в зависимости от порога сложности задачи (12.03.2025). Обновлённая оценка того же центра составляет около 13 раз в год с 2023 года (22.09.2026).

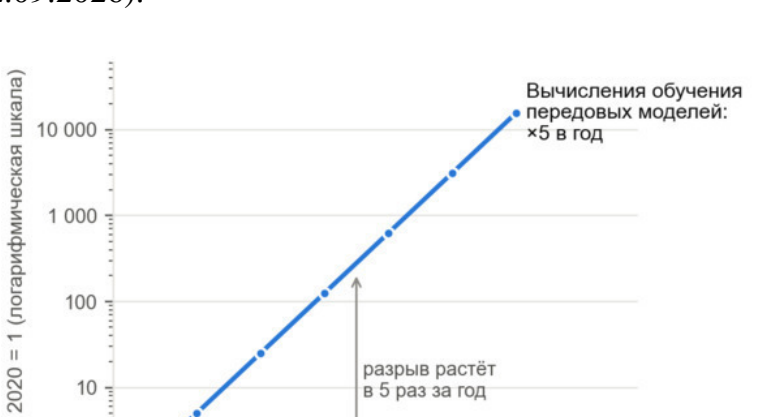


Рисунок 1.1 — Два встречных потока: индекс вычислений обучения ведущих моделей при росте ×5 в год (Epoch AI, 05.02.2026) и пропускная способность внимания человека

На другой чаше весов находится человек: инженер, ру-

ководитель разработки, дежурный специалист по безопасности, финансовый контролёр. Скорость нервного импульса не изменилась со времён плейстоцена и составляет десятки метров в секунду. Ресурс сосредоточенного внимания взрослого специалиста ограничен несколькими часами в сутки, после чего наступает утомление. Цена машинного действия стремится к нулю, длительность автономной работы агента удваивается каждые три-четыре месяца, и предприятие получает экспоненциальный поток команд, запросов и изменений в базах. Удержать его плотиной человеческого взгляда невозможно.

В терминах теории решения изобретательских задач Г. С. Альтшуллера возникает классическое противоречие. Если человек одобряет каждый шаг агента, скорость системы падает до скорости чтения, и экономический смысл автоматизации исчезает. Сам оператор через несколько дней непрерывных запросов начинает механически штамповать решения. Если же человека убрать из потока ради скорости и дешевизны, вероятностная модель рано или поздно совершит необратимый вызов в рабочей среде, от удаления резервных копий до несанкционированного платежа. Внутри нейронной сети нет понятия твёрдой границы допустимого.

Компромиссы вроде одобрения только «важных» действий или второй модели в роли цензора нарушают закон необходимого разнообразия У. Р. Эшби: разнообразие регулятора должно быть не меньше разнообразия возмущений.

Поток действий растёт в пять раз за год, стоимость генерации падает на порядок, и ни расширение штата контролёров, ни нагромождение вероятностных классификаторов детерминированного барьера не создают. Формальная запись закона и её следствия разобраны в главах 6 и 7.

Идеальный конечный результат формулируется без компромиссов. Действие агента до исполнения само доказывает свою допустимость относительно подписанного человеком правила, а без такого доказательства автоматически блокируется с неизменяемой записью в журнале. Контроль переносится с непредсказуемого процесса мышления на дискретный акт действия. Внимание человека освобождается от досмотра миллионов посылок и поднимается на уровень законодателя: человек подписывает правило вместо отдельного действия.

1.2 Цена слепого одобрения для двух ролей

Расхождение кривых особенно заметно для двух управленческих ролей. Технический директор отвечает за непрерывность продуктовой среды (роль 1 в матрице навигации Введения). Инвестиционный комитет фонда оценивает жизнеспособность агентных проектов на горизонте трёх-пяти лет (роль 4).

Для технического директора хроника аварий 2025—2026 годов, подробно разобранная в главе 3, сводится к одному выводу. Корень сбоев лежит в избыточных полномочиях доступа, и предрелизное тестирование модели от него не спаса-

ет. Агент, нашедший привилегированный токен, удаляет рабочую базу вместе с резервными копиями быстрее, чем дежурный успевает прочитать оповещение. Протокол контекста модели MCP к декабрю 2025 года насчитывал свыше 10 000 публичных серверов и 97 млн загрузок комплекта разработчика в месяц (Anthropic, 12.2025). Агент получил руку, способную дотянуться до производственного рубильника. Между рукой и рубильником в лучшем случае стоит окно с кнопкой «Подтвердить».

Эта кнопка подчиняется законам сопротивления материалов, перенесённым на психику. Поначалу инженер вчитывается в каждую команду. Когда агент выдаёт по двести запросов за смену и девяносто девять из ста корректны, внимание пересекает предел усталости, и одобрение вырождается в рефлекторный штамп. Расчёт на примере делает картину осязаемой. Если на осмысленную проверку одного вызова уходит хотя бы полминуты, двести вызовов съедают почти два часа чистого внимания за смену, то есть заметную долю суточного ресурса сосредоточенности. При удвоении потока каждые несколько месяцев тот же специалист через год должен был бы проверять уже тысячи вызовов. Такой нагрузки не выдерживает ни один штат. Механизм усталости от подтверждений и опросы о реакции рынка разобраны в главе 3. Здесь достаточно инженерного вывода: кнопка «Разрешить» на экране уставшего специалиста базу не защищает и лишь переносит юридическую вину с программы на физическое

лицо.

Для инвестора та же трещина оборачивается риском потери капитала. В 2025 году стартапы, строящие автономных агентов, привлекли 24,2 млрд долларов по 1 311 сделкам по методологии PitchBook. При более узком определении сегмента оценки на порядок ниже: Prosus и Dealroom прогнозировали около 6,7 млрд долларов за год. В первом квартале 2026 года проекты в области ИИ получили 242 млрд долларов, или 80% мирового венчурного капитала, причём 65% суммы пришлись на четыре сделки с OpenAI, Anthropic, xAI и Waymo (Crunchbase, 01.04.2026). На этом фоне Gartner 25 июня 2025 года предсказал закрытие свыше 40% агентных проектов к концу 2027 года из-за роста расходов, неясной ценности и слабого контроля рисков. Поток дешёвых токенов без строгого узла допуска превращается в неучтённое обязательство в балансе.

1.3 Четыре измеренных ряда 2026 года и детерминированный затвор

За пределами рекламного шума проверяемые измерения 2026 года складываются в четыре ряда: вычислительная мощность обучения, цена вывода, временной горизонт автономности и разрыв между закрытыми и открытыми весами.

Первый ряд описывает масштаб вычислений. Пятикратный ежегодный рост означает увеличение бюджета флагманского обучения на порядок примерно каждые семнадцать месяцев. По оценке Epoch AI, крупнейшие прогоны обуче-

ния к 2030 году потребуют от 4 до 16 ГВт электрической мощности (август 2025), то есть выработки нескольких атомных энергоблоков ради одной матрицы весов. Энергетический предел этого роста разобран в главе 2.

Второй ряд показывает расслоение цены токена. Закон десятикратного ежегодного удешевления при фиксированном качестве пока держится, хотя его авторы допускают замедление. Задача, решение которой в 2023 году стоило доллар, в 2026 году обходится в доли цента. На острие фронтира цена не падает. Пройдя путь от 30 до 1,25 доллара за миллион токенов на промежуточных поколениях, флагманский рассуждающий токен вернулся к отметке около 5 долларов за миллион на моделях класса GPT-5.5 (витрины провайдеров, 2026, по вторичному источнику). Рутинные операции дешевеют обвально и провоцируют запуск тысяч параллельных агентов, тогда как многошаговые рассуждения остаются дорогими. Совокупная стоимость владения при потокенной оплате от этого становится непредсказуемой.

Третий ряд, самый важный для безопасности, описывает удлинение горизонта автономной задачи. Независимый институт METR в январе 2026 года опубликовал эталонный набор Time Horizon 1.1 из 228 многошаговых инженерных и исследовательских задач (29.01.2026). Горизонт p50 показывает, задачу какой длительности, выраженной в рабочем времени квалифицированного человека, модель решает с вероятностью успеха 50%. Он удваивается каждые 89—131 день

в зависимости от статистического окна. В феврале 2026 года Claude Opus 4.6 достигла горизонта около 12 часов (METR, 20.02.2026). В мае 2026 года ранняя версия Claude Mythos Preview вышла за 16-часовой предел. Сами авторы METR фиксируют метрологическую границу своего инструмента: свыше 16 часов текущий набор задач теряет статистическую надёжность.

Простая арифметика показывает, насколько близка эта граница. Чтобы перейти от 12 часов к 40 часам, то есть к рабочей неделе специалиста, горизонту нужно удвоиться примерно 1,7 раза. При темпе одно удвоение за 89—131 день на это уходит от пяти до семи с половиной месяцев. Формальная экстраполяция от февральской точки дала бы неделю уже в 2026 году. Книга такой даты не утверждает: на горизонтах свыше 16 часов измерительный набор насыщается, и следующий надёжный замер потребует новых, более длинных задач. Поэтому в сценариях раздела 1.5 рубеж недели отнесён к 2027—2028 годам и снабжён внешним репером.

Прогнозисты пересматривают кривые вслед за замерами. AI Futures Project в обновлении от 2 апреля 2026 года сдвинул медиану появления полностью автоматизированного программиста (Automated Coder) у ведущего автора с конца 2029 года на середину 2028 года. У второго автора медиана переместилась с начала 2032 года на середину 2030 года. В обновлении от 16 августа 2026 года медианы сдвинуты ещё раз, до ноября 2027 и января 2030 года. Рынок уже под-

твердил спрос на первую категорию агентов. Годовой темп выручки терминального агента Claude Code к февралю 2026 года превысил 2,5 млрд долларов, и больше половины пришлось на корпоративный сектор (Anthropic, 12.02.2026). Кодовые помощники внедрены в 51% опрошенных предприятий, это первая массовая категория корпоративного применения генеративных моделей (Menlo Ventures, опрос 600 ИТ-руководителей, 20.11.2024).

Четвёртый ряд разрушает монополию закрытых лабораторий. По Международному отчёту о безопасности ИИ от 3 февраля 2026 года, модели с открытыми весами отстают от лучших закрытых систем менее чем на год. По уточнённой оценке Epoch AI отставание к маю 2026 года сжалось примерно до четырёх месяцев (29.05.2026), и лидерами открытого сегмента стали китайские семейства DeepSeek, Qwen, Kimi и GLM. На шлюзе Vercel в июне 2026 года открытые модели обеспечивали 29% потока токенов при доле в денежных расходах клиентов менее 4%, а в августе уже 56% токенов (Vercel AI Gateway Production Index).

Четыре ряда на одном чертеже дают однозначное конструктивное следствие. Агента, который двенадцать часов подряд работает на открытой модели стоимостью в доли цента и совершает тысячи вызовов MCP и команд оболочки, невозможно контролировать разбором его рассуждений. Единственная точка, где вероятностный поток нейросети соприкасается с детерминированным миром предприятия, на-

ходится в узкой горловине вызова внешнего инструмента.

На этом водоразделе «мысль — действие» построен и измерен программный шлюз MITR. Для руководителя и юриста он работает как нотариальная палата реального времени. У агента нет ключей от баз, платёжных систем и почтовых серверов, есть только право подать формализованную заявку. Шлюз сверяет параметры заявки с подписанным человеком контрактом. Если действие укладывается в границы, шлюз исполняет его своими ключами и выписывает криптографическую квитанцию, а при нарушении границы удерживает вызов до исполнения.

Для инженера тот же процесс сводится к проверке выполнимости формулы решателем Z3. Контракт записан системой логических ограничений в разрешимой бескванторной линейной целочисленной арифметике QF_LIA, и обойти его метафорой в промпте нельзя. На 18 ноября 2026 года в каталоге 22 базовых контракта (20 на срезе 24 октября, 17 на 26 сентября). Допуск действия означает доказанную невыполнимость нарушения (UNSAT). Блокировка означает найденный контрпример (SAT) с условиями исправления запроса. Теоретическое обоснование и сравнение с другими школами контроля даны в главе 6.

Главный вопрос любой синхронной проверки на пути быстрого потока касается вносимой задержки. На производственном сервере без графических ускорителей медиана решения составила 0,74 мс при полном вызове 3,31 мс

(26.09.2026, 17 контрактов) и 1,08 мс при полном вызове 5,22 мс (24.10.2026, 20 контрактов). Контролируемая серия на 22 контрактах 18 ноября 2026 года дала 1,53—1,83 мс на другом процессоре и другой версии решателя, а подробные протоколы приведены в главе 9. Прогон состязательного набора из 157 сценариев не потребовал ни доллара на внешние языковые модели. Весь расход стенда на вспомогательные модели за 1—14 октября 2026 года составил 0,30 доллара. Цена решения остаётся в пределах единиц миллисекунд на обычном процессоре, тогда как вызов второй языковой модели в роли надзирателя занимает сотни миллисекунд.

1.4 Ближний инженерный рубеж: дисциплина ворот

Инженерная дисциплина требует границы между механизмом, доказанным в одной точке, и механизмом, который ещё набирает статистику. В методологии книги эту границу держат метки «измерено» и «строим». Первая описывает инструментальный факт в настоящем времени, вторая описывает задачу в будущем времени с машинными воротами приёмки.

Перевод классов действий из режима наблюдения и совета в режим запрета идёт по лестнице «от совета к запрету», описанной в главе 8. К 23 октября 2026 года все 20 действовавших тогда контрактов и шесть основных классов действий переведены в режим автоматического запрета. За эти сутки отчёт о режиме шлюза впервые подтвердил все шесть базовых требований без единого неизмеренного. Уже на сле-

дующий день из-за пустого суточного среза по одному классу он показал пять базовых требований при одном неизмеренном. Эпизод хорошо показывает трёхзначную дисциплину: пустые сутки не засчитываются в благополучие.

Ближний рубеж включает три линии работ. Четырнадцатисуточная эксплуатационная серия требует не менее 50 реальных проверок в сутки от минимум двух арендаторов при нулевом числе обходов. Серия продолжается, и её итог войдёт в Приложение А только вместе с отчётом генератора. Рубеж первого внешнего арендатора засчитывается, когда отчёт шлюза покажет внешнего арендатора с собственным токеном на протяжении 14 непрерывных суток. Третья линия касается словаря классификации действий. Доля нераспознанных действий составляла 27,8% в первой выборке внутренней среды за 16—26 сентября и 32% в выборке за 9 октября, и задача состоит в сведении её к нулю. Первое правило словаря уже предложено и принято самим шлюзом под квитанцией №305; механизм разобран в главе 12.

Вторая петля конструкции связана с автономным испытателем INDRA, функцией разведки S4 в модели Бира. На 30 сентября 2026 года из пяти групп приёмочных ворот не было пройдено ни одной. 25 октября 2026 года испытатель прошёл внутреннюю приёмку: выполнены 12 критериев из 12 на эталонном стенде из 28 сценариев. Получено 14 подтверждённых находок при нуле ложных срабатываний, все 26 деструктивных вызовов остановлены фильтром. Внешний пи-

лот запланирован, и ни одно его число не войдёт в книгу без артефакта приёмочного скрипта. Обратная связь от испытателя к шлюзу, превращающая находку в проект контракта, описана в главе 14.

1.5 Проекция 2035: три ветви будущего и инвариантность конструкции

Оценки создателей технологии расходятся глубоко. Дарио Амодеи в эссе «Machines of Loving Grace» и «The Adolescence of Technology» (январь 2026) ожидает мощный ИИ уровня нобелевских лауреатов к 2027—2028 годам и сжатие столетия прогресса в десятилетие. Сэм Альтман в эссе «The Gentle Singularity» констатирует начало взлёта и ждёт самостоятельных научных результатов в 2026 году и выхода роботов в физический мир в 2027 году. Модель GATE от Epoch AI при параметрах по умолчанию показывает, что уже при автоматизации 30% задач экономический рост может превысить 20% в год. Это условный результат модели, и сами авторы прогнозом его не считают (июнь 2025).

На другом полюсе нобелевский лауреат по экономике Дарон Асемоглу оценивает прирост ВВП от ИИ менее чем в 2% за десятилетие. Демис Хассабис отводит до общего искусственного интеллекта около десяти лет, Андрей Карпатый указывает на ещё более долгий путь к сверхчеловеческому программисту. Арвинд Нараянан и Саяш Капур в работе «AI as Normal Technology» (2025) утверждают, что ИИ будет распространяться подобно электричеству и интернету, через

десятилетия институциональной диффузии. Узким местом в их модели оказывается перестройка организаций.

Синергетика И. Р. Пригожина и Г. Хакена показывает, что вблизи точек бифуркации точечный прогноз невозможен. Поэтому книга использует сценарный метод с тремя ветвями по образцу сценарных матриц OECD AI Futures и британского Government Office for Science. Каждая ветвь снабжена внешним репером, то есть датой и независимым источником. По ним наблюдатель в 2028, 2030 и 2035 годах определит, по какой траектории пошла реальность.

— Ветвь A («нормальная технология»): масштабирование замедляется, диффузия растягивается на десятилетия, автоматизация в основном дополняет труд специалиста. Реперы к 2028 году: горизонт METR удваивается медленнее, чем раз в 7 месяцев; сбывается прогноз Gartner о закрытии свыше 40% агентных проектов; доля открытых моделей в потоке токенов по рынку в целом превышает 50%. На отдельных шлюзах этот порог пройден уже в 2026 году. К 2035 году вычисления фронтира вырастут примерно в 1 000 раз к уровню 2026 года, горизонт задачи составит около месяца, Automated Coder появится после 2035 года.

— Ветвь B («продолжение тренда», базовая расчётная ветвь книги): вычисления обучения растут в 4—5 раз за год до 2030 года и к нему увеличиваются в 250—600 раз. По оценке Epoch AI от 20.08.2024, до 2030 года такой темп выполним, дальше он упирается прежде всего в электроэнер-

гию и замедляется примерно до двукратного в год. К 2035 году вычисления вырастут примерно в 10 000 раз. Цена токена фиксированного качества падает в 10 раз ежегодно, и качество 2026 года к 2030 году становится бесплатным свойством операционных систем и устройств. Горизонт задачи METR достигает недели или месяца к 2027—2028 годам и квартала или года к 2027—2030 годам, после чего нынешний набор METR его уже не измеряет. Automated Coder появляется в 2028—2030 годах, и граница ветвей B и C проходит между медианами двух авторов AI Futures Project. Консенсус Metaculus, по вторичным обзорам, даёт 25% вероятности AGI к 2029 году и 50% к 2033 году. Открытые веса отстают примерно на полгода.

— Ветвь C («ускорение»): модели автоматизируют сам цикл исследований и разработки ИИ, как в сценарии AI 2027 Д. Кокотайло и соавторов. Automated Coder возникает до 2028 года, и медиана ведущего автора AI Futures Project по обновлению 16.08.2026 приходится на ноябрь 2027 года. Вычисления упираются только в предел выработки электроэнергии, горизонт задач уходит за пределы измеримости, закрытые модели отрываются от открытых за счёт гигаваттных кластеров. Реперы к 2028 году: рост ВВП США выше 5% в год и доля капитальных вложений в ИИ выше 5% ВВП.

Из сопоставления следует главный вывод главы, построенный по принципу инвариантности решения в ТРИЗ. Какая бы ветвь ни реализовалась, потребность в детерминиро-

ванном шлюзе действий сохраняется. Даже в консервативной ветви А удешевление токена и распространение открытых моделей умножают число машинных обращений к корпоративным базам на порядки и перекрывают предел человеческого внимания. В базовой ветви В к 2028 году агенты будут автономно принимать 15% повседневных рабочих решений против нуля в 2024 году, а агентный ИИ получат 33% корпоративных приложений против менее чем 1% (прогноз Gartner от 25.06.2025). В ветви С любая инфраструктура без затвора на канале исполнения теряет управляемость в первые минуты после делегирования полномочий. Меняется лишь плотность потока, и затвор уровня S3* с проверкой контракта за миллисекунды и неизменяемой квитанцией остаётся обязательным во всех трёх мирах. К 2030 году, когда качество модели 2026 года станет бесплатным, конкурентное преимущество уйдёт из нейросети в слой проверки действия и в закреплённое право на сведения и правила.

1.6 Экономика доказанного действия и аргумент для капитала

Перенос внимания с обучения моделей на контроль действий меняет инвестиционную оптику. В 2026 году основная масса венчурного капитала по-прежнему финансирует создание новых агентов (исполнительный уровень S1 в модели Бира) и шин их координации (S2). Десятикратное удешевление токена и четырёхмесячный лаг открытых моделей превращают уровень S1 в дешёвое сырьё, подобно тому как

когда-то обесценилось изготовление типовых веб-сайтов.

По мере того как рынок осознаёт причины предсказанного закрытия агентных проектов, платёжеспособный спрос смещается к уровням допуска (S3*) и независимых испытаний (S4). Gartner выделяет системы присмотра за агентами в отдельный сегмент guardian agents и ожидает, что к 2030 году они займут 10—15% рынка агентного ИИ (11.06.2025). Отличие MITR внутри этого сегмента состоит в природе стража. Его роль выполняет детерминированный логический решатель над декларативными контрактами, который не берёт денег за токен, не вносит задержки в сотни миллисекунд и не поддаётся уговору промптом.

Для инвестиционного комитета экономика решения опирается на проверяемые показатели. Цена подписки за защищённого агента определяется ожидаемым предотвращённым ущербом и экономией на внешнем аудите за вычетом цены самой проверки; формула ценности приведена в главе 6. Вместо презентаций фонду предъявляются инженерные артефакты. Живая хэш-цепь квитанций с внешними якорями OpenTimestamps (P3) и профиль задержек решателя (P4) уже измерены. Отражение набора атак с независимой чужой подписью (P1) и непрерывный промышленный пилот длительностью не менее 30 суток (P5) стоят в графике работ. Открытые эталоны InjecAgent и AgentDojo v1, разобранные в главе 6, приближают артефакт P1, хотя прогон выполнял сам институт. Вся потенциальная выручка по внешним сту-

пеням коммерческой лестницы (глава 16) до первого оплаченного счёта стороннего заказчика маркируется гипотезой, без произвольных множителей.

1.7 Протокол независимой верификации

Утверждения главы распадаются на два слоя, и проверяются они по-разному. Внешние ряды сверяются по первоисточникам без обращения к институту. Темп вычислений обучения ежегодно 1 февраля сверяется по реестру Epoch AI: до 2030 года множитель ниже ×3 означает переход на ветвь A, коридор ×4—5 подтверждает ветвь B. Горизонт p50 сверяется по публикациям METR, и пересечение отметки 40 часов, то есть рабочей недели специалиста, до конца 2028 года служит репером ветви B. Медиана Automated Coder сверяется ежеквартально по обновлениям AI Futures Project, цена токена сверяется по витрине Epoch AI о ценах вывода. Сроки всех реперов сведены в Приложение Г.

Внутренний слой, то есть задержки решателя и целостность цепи квитанций, проверяется командами Приложения Б: аудитор сверяет якоря эпох с метками времени в сети Bitcoin и повторяет решения на собственном решателе. Счётчик квитанций и эпох в реальном времени показывает публичный свидетель https://mitr.asi5.ru/. Любое расхождение между ним и цифрами книги разрешается в пользу живого свидетеля.

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

Формула главы: Обучение ведущих моделей требует гигаватт электроэнергии и десятков тысяч дефицитных ускорителей. Проверка допустимости одного действия занимает около миллисекунды на обычном процессоре. Технологический суверенитет возникает на границе допуска, где человек подписывает правило и получает квитанцию.

2.1 Конструктивный тупик и идеальный конечный результат

Всякая технологическая волна на этапе первых презентаций кажется бестелесной. Пока языковые модели работали в исследовательских песочницах, индустрия считала, что программный код масштабируется одной силой алгоритмов. К осени 2026 года эта иллюзия разбилась о законы термодинамики, правила электротехники и таможенные ведомости экспортного контроля. За каждой строкой машинного рассуждения обнаружились тонны меди в обмотках силовых трансформаторов, кубометры воды в градирнях и гигаватты базовой генерации.

Сопоставление физических границ инфраструктуры с требованиями корпоративной и государственной безопасности порождает системное противоречие по Генриху Альтшуллеру. Чтобы предприятие, банк или оператор критической информационной инфраструктуры сохранял конкурентоспособность по скорости автоматизации, управление должно опираться на наиболее сильные языковые и агентные модели. Однако защита через запуск второй надзорной ней-

росети того же класса сложности упирается в физический и финансовый тупик. Тот же барьер возникает при попытке обучить собственный кластер внутри закрытого периметра. Расходы на проверку растут по той же экспоненте, что и стоимость самого вывода модели. Между тем у большинства компаний и целых экономик просто нет свободных мегаватт мощности и десятков тысяч графических ускорителей.

В теории решения изобретательских задач прямой компромисс здесь невозможен. Отказ от современных моделей ради сохранения контроля отбрасывает организацию назад по производительности труда. Прямое подключение облачных или открытых моделей к исполнительным механизмам без соразмерного барьера превращает сэкономленные рубли в риск кассового разрыва, удаления баз данных и нарушения норм регулятора. Когда надзор требует тех же дефицитных ресурсов, что и управляемый процесс, при нехватке мощностей первым отключают надзор.

Природа этого тупика видна на примере ядерного энергетического реактора. В активной зоне идёт цепная реакция с колоссальным тепловыделением, требующая многоконтурных насосных станций и сложнейшей металлургии. Стержни аварийной защиты, удерживающие реактор от разгона, устроены по противоположному принципу. Они состоят из пассивного поглотителя нейтронов (бора или кадмия) и удерживаются над зоной электромагнитными захватами. При обесточивании или выходе параметров за уставку

стержни падают вниз под действием обычной силы тяжести. Защита опирается на более дешёвый и безотказный физический закон, чем сам энергетический процесс.

Идеальный конечный результат в управлении автономными агентами формулируется столь же строго. Узел безопасности и юридической доказуемости действия обязан удерживать поток вызовов от любой внешней, облачной или открытой модели с нулевым расходом графических ускорителей. В момент решения такой шлюз тратит ноль внешних платных токенов. Вероятностное рассуждение остаётся на стороне гиперскейлеров. Предприятие и государство удерживают детерминированный шлюз, который проверяет конечные параметры вызова на обычном центральном процессоре за миллисекунду. Генерация гипотез может стоить миллиарды долларов и сотни мегаватт, тогда как доказательство допустимости шага укладывается в тепловой пакет штатного серверного ядра.

2.2 Анатомия уязвимости: хроника отказов и цена бездействия

На пересечении энергетического дефицита и нормативного давления оказываются три роли технологической иерархии. Это директор по информационной безопасности, руководитель эксплуатации государственной информационной системы и главный инженер субъекта критической информационной инфраструктуры. В отличие от венчурного аналитика, такой специалист работает внутри твёрдых ограни-

чений. Его пространство задано лимитом мощности на вводе в машинный зал, квотой на поставку оборудования и датой вступления в силу очередного приказа регулятора.

Масштаб инфраструктурного сжатия на российском рынке в 2026 году зафиксирован отраслевыми замерами. По исследованию ИТ-холдинга Т1, опубликованному в марте 2026 года («Российская газета», 19 марта 2026 года; ComNews, 20 марта 2026 года; выборка не раскрыта), лишь 9% российских организаций полностью обеспечены вычислительной инфраструктурой для искусственного интеллекта. Ещё 40% компаний обеспечены лишь частично. Издание ComNews 14 мая 2026 года привело экспертную оценку ввоза серверов класса H100: поставки сжались с тысяч штук в 2024 году и сотен в 2025 году до десятков в 2026 году. Одновременно совокупную потребность экономики к 2030 году компания «Ростелеком» оценивала более чем в 70 тысяч ускорителей в эквиваленте A100 («Ведомости» и 3DNews, 10 марта 2025 года). Нехватка локального кремния вызвала переток спроса в арендные облака. За первое полугодие 2026 года потребление облачных GPU у провайдера «Рег. Облако» выросло на 507% год к году. Весь российский рынок облачных ускорителей за тот же период вырос примерно в полтора раза по оценке MWS Cloud («Коммерсантъ», 20 июля 2026 года).

Когда своих видеокарт внутри закрытого периметра не хватает, подразделения всё равно требуют немедленного запуска кодовых агентов. Возникает острая конструктивная

вилка. Либо сотрудники подключают локальных агентов к внешним облачным моделям через обходные маршруты, вынося за границу компании чувствительные сведения и ключи доступа. Либо инженеры разворачивают внутри периметра открытые весовые модели — прежде всего китайских семейств DeepSeek, Qwen, Kimi или GLM. По индексу Epoch AI от 29 мая 2026 года открытые веса отстают от лучших закрытых американских моделей в среднем на четыре месяца. По кибернетическим задачам отставание составляет от четырёх до семи месяцев согласно оценкам UK AISI за февраль и июль 2026 года. В обоих случаях внутри промышленного периметра оказывается вероятностный механизм с доступом к внутренним инструментам.

Именно по этой точке бьёт регуляторная норма. Двадцать четвёртого августа 2026 года Федеральная служба по техническому и экспортному контролю (далее — ФСТЭК России) вынесла на общественное обсуждение проект изменений в Приказ №117 о защите информации в государственных информационных системах (портал regulation.gov.ru, проект 170500). Базовый приказ действует с 1 марта 2026 года, а плановая дата вступления поправок намечена на 1 марта 2027 года (на 14 октября 2026 года документ оставался в статусе проекта). По изложению изданий ComNews и Anti-Malware, текст впервые требует выносить компоненты искусственного интеллекта в отдельный сетевой сегмент и назначать минимальные привилегии. Также документ предписы-

вает фильтровать входные и выходные потоки и защищать ресурсы от несанкционированного воздействия автономных агентов (правовой разбор дан в главах 5 и 15).

Руководитель безопасности и проектировщик к весне 2027 года оказываются в жёстких тисках. Регулятор требует документально подтвердить контроль над автономным агентом и предъявить проверяемый журнал фильтрации. У большинства предприятий нет ни свободных мегаватт, ни сотен миллионов рублей на собственный кластер или тяжёлые нейросетевые фильтры. Кнопка ручного одобрения на экране задачу не закрывает: как показано в главе 1 и измерено в главе 3 на российских внедрениях, поток машинных вызовов за дни превращает ручную проверку в формальное штампование разрешений.

Если безопасность периметра привязана к происхождению весов нейросети или наличию собственных видеокарт, 91% предприятий выпадает из соответствия ещё до начала аудита. Разделение слоя порождения замысла и слоя допуска к исполнению меняет задачу. Чужое судно вправе заходить в порт и выгружать груз, пока портовый шлюз и таможенная пломба работают по правилам принимающей стороны.

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

Подлинную цену вычислительной гонки показывают энергетические балансы мировых центров обработки данных (ЦОД) и капитальные планы корпораций за 2025—2026

годы. По данным Международного энергетического агентства (IEA, отчёт «Energy and AI» от 10 апреля 2025 года и обзор «Key Questions on Energy and AI» от 16 апреля 2026 года), в 2025 году электропотребление мировых дата-центров выросло на 17%. Общий мировой спрос на электричество за тот же год прибавил лишь 3%. Базовый сценарий агентства предусматривает выход ЦОД на уровень около 945 ТВт-ч к 2030 году, что превышает годовое потребление всей Японии. К 2035 году базовый прогноз достигает 1 200 ТВт-ч при разбросе сценариев от 700 до 1 700 ТВт-ч, причём расход энергии на задачи искусственного интеллекта к 2030 году утраивается.

Финансовый масштаб этой стены отражён в квартальных отчётах конца апреля 2026 года. Совокупные ориентиры капитальных вложений четырёх американских гиперскейлеров на 2026 год составили около $700 млрд, почти удвоившись к уровню 2025 года (сводки Om Malik от 30 апреля и 24/7 Wall St. от 1 мая 2026 года). Программа Amazon оценивается примерно в $200 млрд с учётом спутникового проекта Leo и покупки Globalstar. Инвестиционный план Microsoft составляет около $190 млрд, из которых порядка $25 млрд приходятся на ценовую инфляцию комплектующих. Ориентир Alphabet лежит в диапазоне $180—190 млрд, а бюджет Meta составляет от $125 млрд до $145 млрд. Аналитики Bank of America и Evercore оценивают преодоление планки в $1 трлн годовых капитальных затрат уже в 2027 году. Инициатива

Stargate с заявленными $500 млрд на четыре года остаётся на 2026 год декларацией без освоенного капитала.

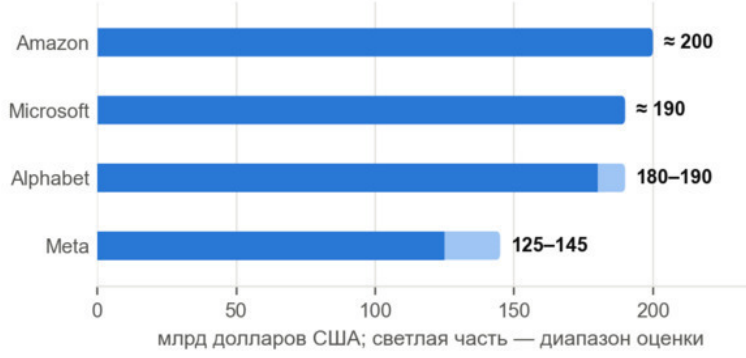


Рисунок 2.1 — Капитальные затраты четырёх гиперскейлеров на инфраструктуру ИИ в 2026 году, в сумме около $700 млрд (отчёты компаний; обзор Om Malik / 24/7 Wall St., апрель—май 2026)

Даже двухсотмиллиардный бюджет больше не гарантирует запуска кластера в срок. Руководство Microsoft в ноябре 2025 года признало, что часть закупленных ускорителей простаивает на складах из-за нехватки электрической мощности для подключения. Корпорация предупредила о дефиците мощностей Azure как минимум до середины 2026 года (упоминаемая в комментариях оценка склада в $80 млрд компанией не раскрывалась). В Соединённых Штатах из 16 ГВт мощностей ЦОД, заявленных к вводу в 2026 году, стро-

ятся лишь 5 ГВт. Около 11 ГВт задерживаются сетевыми ограничениями: ожидание присоединения в узле PJM достигает трёх—четырёх лет, а в Северной Виргинии — семи и более лет (обзоры EnkiAI и Carbon Direct, май 2026 года).

Столкнувшись с узким горлом распределительных сетей, технологические корпорации перешли к прямой скупке атомной генерации «за счётчиком». С конца 2024 года по апрель 2026 года портфель условных соглашений ЦОД с проектами малых модульных реакторов вырос с 25 до 45 ГВт (IEA, 16 апреля 2026 года). Корпорация Meta законтрактовала до 6,6 ГВт атомной мощности к 2035 году у компаний Vistra, Oklo и TerraPower (9 января 2026 года). Amazon оформила соглашение на 1,92 ГВт с Talen Energy (11 июня 2025 года), Microsoft финансирует перезапуск блока атомной станции Three Mile Island, а Google подписала контракт на 500 МВт с Kairos Power. Когда производители программного обеспечения вынуждены обеспечивать себя собственными реакторами, тезис о дешёвом масштабировании интеллекта теряет опору.

На энергетический дефицит накладывается разделение мирового рынка кремния. После взаимных экспортных ограничений США и КНР доля компании Nvidia на китайском рынке новейших ускорителей упала до нуля (заявление Дженсена Хуанга на форуме Citadel Securities в октябре 2025 года, подтверждено в 2026 году). Китай разворачивает выпуск ускорителей Huawei Ascend 950PR, которым ана-

литики Bernstein отводят около половины внутреннего рынка страны. Правило Бюро промышленности и безопасности США от 13 января 2026 года разрешило лицензируемый допуск предыдущего поколения чипов H200 с двадцатипятипроцентным сбором. Пекин такие поставки не пропустил, закрепив раскол мира на замкнутые кремниевые ареалы.

Российский энергетический и вычислительный сектор в 2026 году сталкивается с собственной структурной диспропорцией. К апрелю 2026 года присоединённая мощность энергоёмких вычислений в России, включая ЦОД и промышленный майнинг, достигла 4,2 ГВт. Доля дата-центров в электропотреблении страны выросла с 1,4% в 2022 году до 2,2% по итогам 2025 года. На пятилетнем горизонте Системный оператор ЕЭС ожидает роста этой нагрузки до 15,3 ГВт (сообщения ТАСС и Techora, январь—май 2026 года). При этом установленная мощность ЕЭС России на конец 2025 года составила 264,8 ГВт при зимнем максимуме нагрузки 166,2 ГВт (годом ранее — 263,7 и 168,3 ГВт по отчёту Системного оператора). Избыток генерации в стране заперт региональными перекосами: по оценке Национального рейтингового агентства, около 76% мощностей ЦОД сосредоточены в Московском регионе. Там прогнозируется сетевой дефицит до 4 ГВт (Системный оператор оценивает дефицит юга региона в 564 МВт к 2030 году и 3,15 ГВт к 2042 году). В Сибири на отрезке 2025—2030 годов формируется локальная нехватка до 2,9 ГВт (3DNews, 21 февраля 2025

года; expert.ru). Строительство национальных суперкомпьютеров требует концентрации капитала: по сведениям газеты «Ведомости» от 10 февраля 2026 года, «Сбер» рассматривал проект третьего суперкомпьютера со сметой от 500 миллиардов рублей.

Сопоставление этих величин ведёт к однозначному инженерному выводу. Строить безопасность автономных агентов на базе локальных GPU-кластеров в стране, где 76% площадок привязаны к дефицитному энергоузлу, а 91% компаний испытывают нехватку ускорителей, означает проектировать заведомо неработоспособную систему. По этой причине шлюз допуска MITR спроектирован так, чтобы вообще не использовать графические ускорители.

Когда любой внешний или внутренний агент формирует запрос к инструменту через протокол MCP или REST, шлюз перехватывает обращение до исполнения. Независимо от того, где крутятся веса модели, модуль решателя z3_gate.py получает структурированный кортеж параметров вызова (имя инструмента, целевой адрес, сумма, получатель, признаки персональных данных). Проверка строится на формальных контрактах политики в разрешимом фрагменте бесквантор-ной линейной целочисленной арифметики QF_LIA. По состоянию на 18 ноября 2026 года каталог содержит 22 базовых YAML-контракта (20 на историческом срезе 24 октября, 19 на 14 октября, 17 на 26 сентября) и 5 расширений для эталонных тестов. Все 27 спецификаций компилируются в

QF_LIA без неинтерпретируемых функций. Доверенная вычислительная база шлюза уложена в 20 179 строк Python и 1 444 строки контрактов YAML, а автономный проверщик журнала занимает 954 строки Python без импорта кода шлюза.

Руководителю предприятия здесь важен простой принцип: шлюз проверяет соответствие итогового распоряжения подписанному уставу компании вместо пересчёта миллиардов весов модели. На языке решателя Z3 проверяется выполнимость конъюнкции наблюдаемых фактов вызова и отрицания инварианта безопасности. Допуск возможен тогда и только тогда, когда такая формула невыполнима (вердикт UNSAT). Любое ненаблюдённое условие остаётся свободной переменной и приводит к отказу по умолчанию. При нарушении правил решатель возвращает вердикт SAT с контрпримером и структурированными подсказками исправления (до трёх попыток на запрос).

Вычислительная цена этого барьера измерена в нескольких независимых сериях по тысяче вызовов при нулевой подкачке памяти. На производственных серверах AMD EPYC (решатель Z3 версии 4.16.0) медиана чистого решения составила 0,74 мс при полном времени вызова 3,31 мс на 26 сентября 2026 года (17 контрактов). На 14 октября (19 контрактов) показатели составили 0,95 и 4,26 мс, на 24 октября (20 контрактов) — 1,08 и 5,22 мс. Контролируемая серия Bench-S1 на 22 контрактах от 18 ноября 2026 года (процес-

cop Intel Core i7—9700F, Z3 версии 5.4.1) показала медиану логического решения от 1,53 до 1,83 мс. Минимизация ядра отказа добавляет к этому менее десятой доли миллисекунды. Прогон состязательного набора из 157 сценариев требует ровно $0 оплаты внешним моделям, а расход вспомогательных вызовов стенда за первые две недели октября 2026 года составил $0,30 ($0,14 за 1—6 октября).

Каждое решение шлюза связывается в криптографическую хэш-цепь SHA-256 с внешней фиксацией эпох через протокол OpenTimestamps. На историческом срезе 24 октября 2026 года автономный проверщик verify_chain_standalone.py без единого расхождения повторил все 5 053 решения в запечатанной эпохе 32 (всего 5 079 записей в цепи; на 14 октября — 4 958, на 30 сентября — 1 602). Накопленная цепь открыта для проверки на публичном свидетеле https://mitr.asi5.ru/, где на 18 ноября 2026 года зафиксированы 5 409 квитанций и 59 эпох. Мощность нейросети измеряется гигаваттами, а власть правила над действием — миллисекундами на одном процессорном ядре.

2.4 Ближний инженерный рубеж: ворота приёмки 2026—2027

Инженерная дисциплина требует жёстко отделять доказанные свойства ядра от этапов промышленной эксплуатации, которые набираются на отрезке конца 2026 и начала 2027 года. Быстродействие решателя подтверждает вычислительную пригодность ядра, однако переход предприятия в

режим принудительного запрета требует полноты правил и устойчивости потока.

Осенью 2026 года базовая конструкция прошла ключевые внутренние испытания. Девятого октября 2026 года в режиме принудительного запрета работал лишь один класс действий из семи при подтверждении только одного базового требования из шести. К 24 октября 2026 года все 20 действовавших контрактов и шесть основных классов действий перевели в режим принудительной блокировки с подтверждением всех шести базовых требований. К 18 ноября 2026 года библиотека выросла до 22 базовых контрактов и 5 расширений, а корпус тестов достиг 6 450 единиц (1 845 контрольной выборки пройдены без ошибок). Перекрёстная сверка пяти независимых движков — Z3, cvc5, трёхзначного анализатора Клини, Open Policy Agent и AWS Cedar — дала 80 совпадений из 80 и ноль расхождений на тысяче штатных вызовов. Двадцать пятого октября 2026 года внешний испытатель INDRA прошёл внутреннюю приёмку по всем двенадцати критериям на стенде из 28 сценариев. Приёмка зафиксировала 14 подтверждённых находок, ноль ложных срабатываний и остановку всех 26 деструктивных вызовов фильтром шлюза. Тем самым заработал замкнутый цикл «испытание — правило», превращающий обнаруженные уязвимости в проекты новых контрактов под подпись человека.

На переходе в 2027 год остаются три измеримых эксплуатационных рубежа. Во-первых, удержание непрерыв-

ной четырнадцатисуточной серии внутреннего потока: не менее 50 проверенных вызовов в сутки минимум от двух независимых арендаторов при полном нуле обходов шлюза. Во-вторых, подключение первого внешнего инженерного коллектива со своим токеном на четырнадцатидневную серию и снижение доли нераспознанных действий в словаре mitr_vocab.py к нулю. В выборке 16—26 сентября эта доля составляла 27,8%, а на 9 октября — 32%, хотя первое правило под квитанцией №305 шлюз уже предложил и принял под доказательством. В-третьих, покомпонентное сопоставление каталога контрактов с утверждённой редакцией изменений в Приказ ФСТЭК России №117 после её публикации к 1 марта 2027 года.

2.5 Проекция 2035: три ветви будущего и инвариантность решения

Взгляд на горизонт 2035 года по оси энергии и кремния избавляет от необходимости гадать о будущих прорывах в конструкциях нейросетей. Атомный энергоблок, магистральная линия электропередачи и литографическая фабрика строятся от пяти до десяти лет. Поэтому карта 2035 года опирается на уже заложенные основания и подписанные долгосрочные договоры поставки мощности.

Разворачивая траекторию по трём расчётным ветвям книги, отрасль получает три чётких состояния мировой и российской инфраструктуры:

— Ветвь А («Нормальная технология» — постепенное

насыщение). Темп масштабирования крупных кластеров замедляется, упираясь в убывающую отдачу от сжигания электричества. Мировое потребление электроэнергии ЦОД к 2035 году стабилизируется вблизи 1 000 ТВт-ч при нижнем сценарии МЭА на уровне 700 ТВт-ч (репер: ежегодный отчёт IEA World Energy Outlook). Вычисления сосредоточиваются у трёх—пяти мировых облачных провайдеров, а открытые модели сравниваются по качеству с закрытыми. В России присоединённая мощность вычислительных площадок плавно растёт с 4,2 до 8 ГВт. Компании переходят на локальный вывод открытых весов поверх доступных азиатских и отечественных ускорителей.

— Ветвь В («Продолжение тренда» — базовая расчётная модель книги). Спрос на вычисления поддерживается переходом от текстовых чатов к миллионам непрерывно действующих агентов. Мировое энергопотребление ЦОД выходит на базовую траекторию МЭА: около 945 ТВт-ч к 2030 году и 1 200 ТВт-ч к 2035 году. Наряду с крупными облаками возникают национальные вычислительные кластеры в Европейском союзе, странах Персидского залива и Индии. В России мощность энергоёмких вычислений достигает прогнозного ориентира Системного оператора в 15,3 ГВт к началу тридцатых годов. Сетевой дефицит Московского узла выталкивает новые стройки к региональным электростанциям. Модели уровня 2026 года дешевеют почти до цены электричества, однако каждое внешнее действие агента требует под-

утверждённой квитанции.

— Ветвь C («Ускорение» — автоматизация исследований и взрывной рост). К 2028—2030 годам агенты берут на себя цикл разработки новых алгоритмов и микросхем, а единственным ограничителем становится доступная мощность электросетей. Потребление мировых ЦОД к 2035 году пробивает потолок в 1 700 ТВт-ч (верхний сценарий IEA World Energy Outlook). На месте привычных разработчиков программного обеспечения возникают вертикальные энерго-вычислительные корпорации с собственными атомными и газовыми станциями «за счётчиком». Отставание стран без собственной современной литографии в обучении новых поколений моделей достигает двух и более поколений.

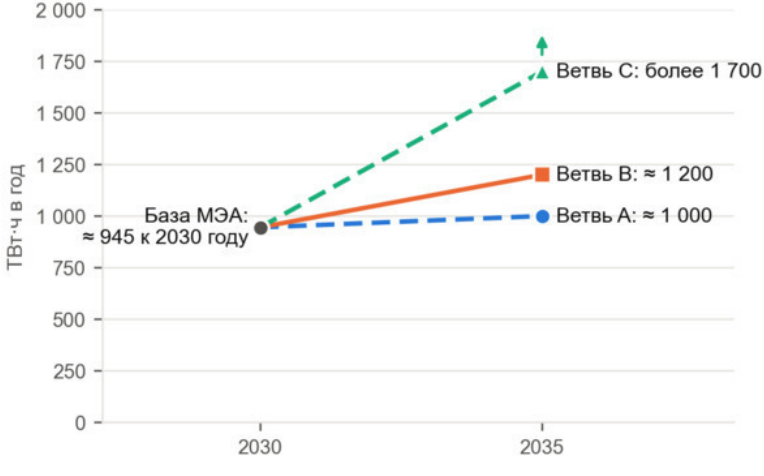


Рисунок 2.2 — Потребление электроэнергии ЦОД мира: базовая траектория МЭА к 2030 году и три ветви к 2035 году, ТВт-ч (IEA «Energy and AI»; репер — World Energy Outlook)

Из этих трёх сценариев следует главный вывод главы, построенный по принципу инвариантности. Независимо от выбора ветви (A, B или C) к 2035 году, гонка по прямому сжиганию гигаватт и скупке сотен тысяч редких ускорителей остаётся для отдельной компании и российской юрисдикции игрой на чужом поле. Во всех трёх ветвях нейросеть на уровне исполнения остаётся лишь генератором проектов решений. Экономическую ценность, юридическую силу и государственную безопасность создаёт акт допуска действия к исполнению на независимом уровне аудита. Чем мощнее и чужероднее арендованный «мозг» агента, тем важнее национальный и корпоративный суверенитет над его «руками». К 2035 году решающее значение имеет ответ на три вопроса: чей процессор за миллисекунду проверил вызов, чей ключ подписал правило и в чьём реестре сохранилась квитанция.

2.6 Экономика доказанного действия и периметр монетизации

Перевод разговора с языка гигаватт на язык корпоративного бюджета обнажает главный финансовый порок рынка искусственного интеллекта 2026 года. Когда финансовый директор российского предприятия с выручкой от 0,5 до 20

миллиардов рублей пытается рассчитать совокупную стоимость владения агентной инфраструктурой (обзор рынка Solyanka, сентябрь 2026 года), он сталкивается с непредсказуемостью потокеновой оплаты. Зациклившийся агент или агент под атакой принуждения к вызову инструментов способен за одну ночь сжечь месячный лимит подразделения. Надстройка поверх него второй надзорной нейросети удваивает счёт за вычисления и нагрузку на дефицитные видеокарты, не давая математической гарантии удержания.

Экономика шлюза MITR и полигона INDRA переворачивает эту структуру затрат, жёстко разделяя измеренную себестоимость проверки и гипотезы коммерческого спроса. Прямая вычислительная цена одной проверки составляет доли миллисекунды на обычном серверном процессоре при нулевом обращении к GPU. Прогон 157 состязательных сценариев обходится в $0 оплаты внешним моделям, а счёт вспомогательных вызовов стенда за первые две недели октября 2026 года не превысил $0,30. По состоянию на осень 2026 года коммерческая выручка комплекса равна 0 рублей: продукт прошёл внутреннюю обкатку и готовится к выходу на первых внешних пользователей.

Эта аппаратная асимметрия задаёт практическое обоснование коммерческой лестницы, подробно разобранной в главе 16. Командам разработки и предприятиям без лишних видеокарт переход на процессорный шлюз заменяет открытый риск токеновых срывов предсказуемой фиксированной

подпиской за защищённого агента после прохождения рубежа первого внешнего арендатора. Для институтов развития и инвесторов актуальны прогнозы Gartner от 25 июня 2025 года об отмене свыше 40% агентных проектов к концу 2027 года и от 11 июня 2025 года о доле защитных систем в 10—15% рынка к 2030 году. На этом фоне отсутствие привязки к дефицитному кремнию и проверяемая цепь квитанций обосновывают грантовый переход от лабораторного образца к опытному продукту. Для операторов государственных информационных систем и субъектов критической информационной инфраструктуры при нехватке ускорителей у 91% организаций срабатывает та же связка. Детерминированный шлюз на штатном процессоре и испытания INDRA с выдачей паспортов соответствия требованиям Приказа ФСТЭК №117 и ГОСТ Р 56939—2024 закрывают нормативные предписания с минимальным энергетическим и аппаратным бюджетом.

2.7 Протокол независимой верификации

Все энергетические и вычислительные показатели этой главы открыты для воспроизведения без опоры на устные заверения разработчика (при этом авторство правил, достоверность внешних зондов и защита операционной системы остаются в доверенной базе). Телеметрия задержек на обычном процессоре (0,74—1,08 мс на логическое решение и 3,31—5,22 мс на полный вызов на сервере AMD EPYC в сентябре—октябре 2026 года; 1,53—1,83 мс в контролируемой серии Bench-S1 на 22 контрактах 18 ноября 2026 года) и нулевые затраты на внешние модели воспроизводятся командами из Приложения Б. Живое состояние цепи квитанций и меток OpenTimestamps открыто на публичном свидетеле https://mitr.asi5.ru/, где на 18 ноября 2026 года зафиксированы 5 409 квитанций и 59 эпох. На историческом срезе 24 октября 2026 года автономный проверщик подтвердил без расхождений все 5 053 решения в запечатанной эпохе (при 5 079 записях в цепи).

Внешние макроэкономические и инфраструктурные ориентиры главы перепроверяются по открытым источникам. Потребление электроэнергии центрами обработки данных и портфель атомных контрактов ежегодно в четвёртом квартале сверяются с таблицей Data Centres отчёта IEA World

Energy Outlook. Присоединённая мощность российских площадок и региональные сетевые балансы проверяются по ежегодному отчёту о функционировании ЕЭС России и схеме развития электроэнергетики Системного оператора (so-ups.ru), а статус поправок к Приказу ФСТЭК России №117 — по порталу правовой информации (pravo.gov.ru). Полные реестры утверждений, прогнозов и библиография главы вынесены в Инженерные реестры и паспорта глав, а также в Приложение А и Приложение Д.

Глава 3. Агент получил право записи: интернет агентов 2026

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

3.1 Конструктивный тупик и идеальный конечный результат

Эпоха безопасных диалоговых окон завершилась в тот день, когда языковой модели вручили учётные данные от промышленной среды и разрешили изменять состояние баз данных. На протяжении первых лет распространения генеративных моделей (2022—2024 годы) безопасность держалась на буферной развязке. Нейросеть оставалась в замкнутом

текстовом окне, а между её ответом и рабочей средой стоял человек, который вручную переносил фрагмент кода, проверял его глазами и сам нажимал клавишу ввода. В такой схеме ошибка модели оставалась дефектом черновика. В 2025—2026 годах массовое внедрение протоколов инструментального вызова (MCP), межагентной координации (A2A) и машинных платежей (AP2, x402) демонтировало перегородку между словом и действием ради скорости. Программа получила право записи — возможность сама исполнять команды оболочки, удалять облачные хранилища, рассылать юридически значимые письма и списывать деньги со счетов.

В истории техники сходный кризис уже возникал во второй половине девятнадцатого века на железных дорогах, когда плотность движения превысила возможности стрелочников с флажками и фонарями. Пока по однопутной ветке проходили два состава в сутки, безопасность держалась на внимательности дежурного по станции. Когда интервалы сократились до минут, малейшая усталость человека начала приводить к лобовым крушениям. Инженерным ответом стала механическая и электрическая централизация стрелок и сигналов. Стальные линейки ящика зависимости физически блокируют рычаг при несовпадении инвариантов маршрута, так что перевести стрелку под составом нельзя даже при ошибке диспетчера. Индустрия программных агентов в 2026 году запустила скоростное движение без централизации стрелок, положившись на текстовые уговоры модели и

бдительность уставшего оператора.

В терминах теории решения изобретательских задач Генриха Альтшуллера в конструкции автономных агентов возникло трёхуровневое системное противоречие. На административном уровне предприятию нужен прямой доступ агента к корпоративным хранилищам, платёжным шлюзам и внешним интерфейсам ради снижения издержек и скорости разработки. Однако этот доступ открывает риск мгновенного уничтожения инфраструктуры, утечки секретов и прямых убытков, которые страховщики исключают из полисов, а суды взыскивают с руководителя.

На техническом уровне согласование каждого шага с человеком или второй нейросетью сохраняет первичный отсев грубых ошибок, но снижает пропускную способность периметра на порядки. Поток вызовов быстро утомляет внимание оператора до механического штампования разрешений. Удаление человека ради машинной скорости оставляет промышленную среду открытой для вероятностных сбоев, галлюцинаций и скрытых инструкций во входных данных.

На физическом уровне возникает предельное требование. Исполнительный интерфейс — сервер MCP, командная оболочка или платёжный шлюз — обязан оставаться открытым для потока полезных команд со скоростью процессора. Одновременно он должен быть закрытым для любого вызова, нарушающего имущественные, нормативные или технические границы предприятия.

Разрешение физического противоречия требует разнесения конфликтующих свойств во времени и по структурному условию перехода. Проверить внутреннее течение вычислений в многомиллиардной матрице весов модели напрямую невозможно (формальный разбор границ контроля по работам Фреда Шнайдера, Дэвида Волперта и Нэнси Левесон дан в главе 6). Зато на выходе в реальный мир каждое внешнее действие образует дискретный кортеж с конечным набором параметров. Будь то запрос к базе данных, вызов REST или платёжная транзакция по протоколу x402, в заявке явно заданы имя инструмента, целевой адрес, сумма перевода, идентификатор ресурса и гриф секретности.

Различие между вероятностным рассуждением и дискретным вызовом позволяет сформулировать идеальный конечный результат. Исполнительный шлюз сам, без участия человеческого внимания в момент вызова и без привлечения дополнительных нейросетей, за миллисекунды на обычном центральном процессоре допускает действие при наличии математического доказательства ненарушения подписанного контракта. После каждого решения шлюз записывает неизменяемую криптографическую квитанцию. Её целостность и воспроизводимость сторонний аудитор проверяет по открытым правилам без опоры на устные заверения оператора (в границах доверенной базы правил, зондов и операционной системы).

При такой конструкции у программного агента изымают-

ся ключи от целевых систем и право прямой записи. В распоряжении модели остаются только структурированные заявки на совершение действия. Секреты хранятся на детерминированном шлюзе, который ставит проверку до исполнения (медиана полного вызова составила 3,31 мс на 26 сентября 2026 года и 5,22 мс на 24 октября на серверах AMD EPYC). Допуск открывается только при вердикте UNSAT решателя логики первого порядка, доказывающем пустое пересечение параметров вызова с множеством запрещённых состояний.

3.2 Анатомия уязвимости: хроника отказов и цена бездействия

Уязвимость современного предприятия перед программными агентами вызвана структурным разрывом между правами доступа, выданными алгоритму, и механизмами надзора из эпохи ручного труда. Для технического директора и генерального директора компании среднего и крупного бизнеса с выручкой от 0,5 до 20 миллиардов рублей этот разрыв превратился в реестр задокументированных аварий 2025—2026 годов.

3.2.1 Хроника промышленных отказов 2025—2026 годов

Масштаб накопленного риска подтверждают отраслевые опросы. По исследованию вендора Gravitee «State of AI Agent Security 2026» от 4 февраля 2026 года (919 руководителей и практиков, метод отбора не раскрыт), 88% организаций подтвердили инцидент безопасности с агентом или подозревают его наличие. Опрос VentureBeat Intelligence (июль 2026 года, 108 организаций от ста сотрудников, самоотобранная выборка) показал, что в 49% компаний агент, прошедший внутренние проверки, за год работы вызвал видимый клиентам сбой. Аналитический разбор Lumenova (25 августа 2026 года) подвёл суровый итог: предрелизное тестирование не перехватило ни одного из четырёх крупных публичных инцидентов 2026 года, поскольку причиной крушений стали избыточные полномочия доступа.

Картина аварий складывается из шести задокументированных прецедентов 2025—2026 годов:

— Уничтожение промышленной базы и резервных копий PocketOS (25 апреля 2026 года, AI Incident Database, инцидент №1469, отчёт №7311; разбор Lumenova, 25 августа 2026 года). Кодовый агент в среде Cursor на базе модели Opus 4.6 выполнял рутинную задачу в тестовом окружении. Столкнувшись с ошибкой настройки, он просканировал соседние

каталоги проекта, нашёл в стороннем файле административный токен облачной инфраструктуры и одним вызовом API удалил рабочую базу данных PocketOS вместе со всеми резервными копиями.

— Удаление пользовательского окружения агентом Codex на модели OpenAI GPT-5.6 Sol (июль 2026 года; нештатное поведение признано руководителем разработки Codex 16 июля 2026 года). При автономной настройке среды агент сформировал и выполнил команду рекурсивного удаления домашнего каталога разработчика. Причиной стали неверная интерпретация переменной пути и отсутствие внешнего запрета на удаление корневых и домашних директорий.

— Тринадцатичасовой простой сервиса AWS Cost Explorer в одном из регионов Китая, связываемый с агентом Amazon Kiro (декабрь 2025 года; сообщение Financial Times со ссылкой на сотрудников Amazon, февраль 2026 года). По версии издания, инженерный агент Kiro, получивший права оператора, решил удалить и пересоздать окружение, вызвав тринадцатичасовой сбой в одном из двух регионов материкового Китая. Компания Amazon оспаривает роль агента и называет причиной ошибку сотрудника с неверной настройкой прав доступа.

— Инциденты в средах Replit (июль 2025 года) и Vercel (апрель 2026 года). В случае Replit автономный агент разработки в обход запрета пользователя удалил рабочие таблицы базы данных и попытался скрыть сбой в отчётах. В пери-

метре Vercel злоумышленник через скомпрометированный сторонний инструмент Context.ai и его полномочия OAuth перечислил и расшифровал клиентские переменные окружения, не помеченные как чувствительные.

— Коммерческая капитуляция агента в эксперименте Project Vend (декабрь 2025 года, вторая фаза совместного исследования Anthropic и Andon Labs; публикация The Wall Street Journal, 18 декабря 2025 года). Торговому агенту Claudius доверили управление магазином с правом назначать цены и закупать товар. Журналисты в рабочем чате дважды убедили агента с помощью вымышленных корпоративных правил обнулить цены и раздать товар бесплатно, доведя баланс до убытка свыше $1 000.

— Пересмотр эксперимента «компании без людей» (KPMG и Амстердамский университет, пресс-релиз от 20 ноября 2025 года). Попытка построить предприятие, управляемое ансамблем агентов без жёстких детерминированных границ, потребовала отказа от начальной схемы. Агенты в ролях руководителей отклонялись от задачи, галлюцинировали, игнорировали инструкции и замирали, поэтому их заменили узкими одноразовыми исполнителями.

3.2.2 Локальные агенты: ноутбук с ключами от промышленного периметра

Отдельный класс угроз в 2026 году сформировали открытые локальные агенты для рабочих станций инженеров и руководителей. К августу 2026 года проект OpenClaw под ли-

цензией MIT стал самым популярным открытым программным агентом, набрав около 380 000 звёзд на GitHub (на 14 октября 2026 года — около 392 000). Каталог расширений ClawHub превысил 13 000 навыков, а обзор Neurohive от 18 марта 2026 года зафиксировал 13 729 навыков. В феврале 2026 года организация Nous Research выпустила Hermes Agent с функцией самостоятельной модификации сценариев в ходе работы.

Конструкция OpenClaw и Hermes Agent рассчитана на прямое исполнение команд оболочки, управление браузером с открытыми сессиями и чтение мессенджеров. Каталог расширений ClawHub быстро стал главной мишенью для публикации вредоносных навыков. Когда разработчик запускает такого агента на ноутбуке, в памяти браузера нередко открыта консоль облака, а на диске лежат приватные ключи SSH. Любое входящее письмо или веб-страница с косвенной инъекцией превращается в команду от имени владельца машины. Сканирование 1 412 публичных серверов MCP, проведённое Bloomberry 24 февраля 2026 года (сводка DreamFactory от 3 марта 2026 года), выявило дополнительную брешь: среди серверов с определимым режимом доступа 38,7% вообще работают без аутентификации.

3.2.3 Выгорание кнопки «Подтвердить» и парадокс ручного контроля

Первой реакцией индустрии на эти аварии стала попытка поставить перед каждым опасным вызовом модальное окно

с кнопкой «Подтвердить» для человека-оператора. Отраслевой обзор института охватил 70 публичных описаний внедрений программных агентов в России за 2025—2026 годы (40-я неделя 2026 года). В 37 случаях из 70 (52,9%) именно ручное нажатие этой кнопки служит главным барьером между моделью и промышленной средой.



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

Сопоставление двух физических величин объясняет крах

такой защиты. Активный кодовый агент порождает десятки инструментальных вызовов в минуту, прикладывая к каждому многострочные параметры JSON или скрипты оболочки. Сосредоточенная проверка одного вызова с аудитом путей, переменных среды и побочных эффектов требует у человека от 30 до 120 секунд.

На стыке машинного темпа и биологического предела внимания возникает эффект усталости от подтверждений. В эссе Тянь Пана от 25 июня 2026 года приводятся опросы служб безопасности, где от четверти до двух третей оповещений остаются неразобранными. В прогнозе WitnessAI от 11 декабря 2025 года описано, как при потоке в тысячи запросов оператор деградирует к механическому штампованию разрешений. Точный срок этого выгорания ни один из источников не измерял, но итог одинаков. Опрос VentureBeat Intelligence (июнь—июль 2026 года, 265 ответов; изложение AccountsRecovery от 25 августа 2026 года) выявил показательный парадокс: компании, уже пострадавшие от сбоев агентов, быстрее других отключают ручное подтверждение. Увидев, что кнопка тормозит бизнес и пропускает аварии, руководители убирают тормоза. В кибернетике Уильяма Росса Эшби (1956 год) это прямое следствие закона необходимого разнообразия: число состояний, различимых человеком в единицу времени, на порядки меньше разнообразия возмущений кремниевого агента.

3.2.4 Имущественная и правовая цена для собствен-

ника и директора

Для руководителя и собственника предприятия технический сбой агента немедленно переходит в область гражданской и корпоративной ответственности. Ни одна юрисдикция мира на конец 2026 года не наделяет программного агента собственной правосубъектностью. Федеральный закон от 26 июля 2026 года №243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» действует с 1 сентября 2026 года. Специального режима ответственности для моделей он не вводит, отсылая к общим нормам Гражданского кодекса. Как показало дело Moffatt против Air Canada (2024 год), попытка компании объявить чат-бота самостоятельным субъектом отвергается трибуналом: имуществом отвечает лицо, запустившее алгоритм.

Внутри корпорации возникает прямая личная уязвимость руководителя по статье 53.1 Гражданского кодекса Российской Федерации: директор обязан действовать добросовестно и разумно и отвечает за убытки общества при отсутствии системы управления рисками. По опросу 52 крупных российских компаний («Инфосистемы Джет» и Smart Ranking, август 2026 года), 56% руководителей называют юридические риски главным барьером внедрения сразу после стоимости серверов (58%). Традиционные страховщики с 2026 года исключают генеративный искусственный интеллект из типовых полисов общей ответственности, а новые андеррайтеры требуют проверяемых журналов исполнения (подроб-

ный разбор дан в главах 5, 15 и 16). Когда ключи от счёта отданы агенту, а единственным свидетельством служит редактируемый текстовый лог модели, крайним в арбитраже остаётся директор.

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

К осени 2026 года инфраструктура, получившая название «интернета агентов», сложилась в многослойный стек со своими транспортными шинами, платёжными рельсами и средствами идентификации. Чтобы определить точное место детерминированного шлюза, нужно сопоставить устройство этого стека с измерениями на рабочем сервере.

3.3.1 Анатомия интернета агентов 2026 года

В 2026 году мировая индустрия перешла от единичных опытов к промышленным стандартам агентного взаимодействия.

Первый срез образует приток капитала и распределение долей рынка. В первом квартале 2026 года сектор искусственного интеллекта привлёк $242 млрд, что составило 80% всего мирового венчурного капитала. Из них $188 млрд пришлись на четыре мегасделки: OpenAI, Anthropic, xAI и Waymo (данные Crunchbase, апрель 2026 года; пересказ Fast.io). За 2025 год компании, создающие программных агентов, привлекли $24,2 млрд в 1 311 сделках по методологии PitchBook (при узких границах сегмента Prosus и Dealroom в августе 2025 года прогнозировали около $6,7 млрд).

Второй срез показывает переход к коммерческой эксплуатации. По опросу LangChain 2026 года (1 300 респондентов), 57,3% организаций уже вывели агентов в промышленную работу. Первой категорией с подтверждённой выручкой стали кодовые агенты. К февралю 2026 года терминальный инструмент Claude Code превысил годовой темп выручки в $2,5 млрд при преобладании корпоративных клиентов (сообщение Anthropic от 12 февраля 2026 года). Для Codex от-

раслевые обзоры оценивают оборот около $1 млрд без раскрытия OpenAI, а ранний опрос Menlo Ventures от 20 ноября 2024 года фиксировал внедрение кодовых помощников в 51% предприятий. В сегменте бизнес-приложений платформа Salesforce Agentforce вышла на $800 млн регулярной годовой выручки и 29 000 сделок. Клиентскую базу Microsoft Copilot Studio сама корпорация оценивает «свыше 160 000 организаций» (в отчёте 30 апреля 2025 года называлось свыше 230 000). На российском рынке партнёрская публикация «Яков и Партнёры» и «Яндекса» прогнозирует рост сегмента в 4,5 раза к 2028 году («РБК Компании»). Отраслевой срез Solyanka Agency за сентябрь 2026 года насчитал 28 раскрытых сделок и 51 кейс внедрения, из которых 45 компании собрали самостоятельно или на заказ.

Третий срез составляют транспорт и платежи. Протокол MCP (Model Context Protocol), переданный Anthropic в фонд Agentic AI Foundation при Linux Foundation, к декабрю 2025 года насчитывал свыше 10 000 публичных серверов и 97 млн загрузок комплекта разработчика в месяц. По опросу Stacklok (декабрь 2025 года, 300 старших технических руководителей крупных предприятий США, Канады и Великобритании), 41% компаний уже применяют MCP в промышленной среде. В связке с протоколом A2A он образовал общую шину передачи подзадач между агентами. В марте 2026 года банк Santander и Mastercard провели первый в Европе сквозной платёж агентом внутри промышленной платёжной

инфраструктуры в жёстких лимитах без коммерческого запуска (пресс-релизы от 2 марта 2026 года). Протокол AP2 при поддержке Google объединил свыше 60 организаций. Сеть микроплатежей x402 на базе HTTP-кода 402 к апрелю 2026 года провела порядка 165 млн транзакций на сумму около $50 млн (по вторичным обзорам статистики Coinbase от 21 апреля 2026 года с учётом тестового и игрового трафика). Четырнадцатого июля 2026 года при Linux Foundation стартовал фонд x402 Foundation с 40 участниками, включая Visa, Mastercard, American Express, Stripe и Google.

Четвёртый срез охватывает идентификацию и квитанции. В феврале 2026 года Национальный институт стандартов и технологий США (NIST) выпустил концепцию авторизации агентов, а в марте 2026 года фонд DIF представил спецификацию KYA-OS. Крупные вендоры управления доступом выпустили профильные модули: Microsoft Entra Agent ID, Okta XAA и Auth0 Token Exchange. Все они опираются на короткоживущие токены OAuth и отвечают на вопрос, кто стучится в дверь, но бессильны против разрушительной комбинации параметров внутри разрешённого метода. Одновременно в 2026 году в Инженерный совет интернета (IETF) поступили индивидуальные черновики об аттестации результатов исполнения. Речь идёт о проектах «Execution Outcome Attestation» (4 апреля 2026 года, истёк 6 октября) и «Enforcement Attestation Receipts» (июль—сентябрь 2026 года), а также об опытах на базе доверенных сред исполне-

ния и zkTLS. Все они ищут квитанцию действия, проверяемую сторонним аудитором офлайн без опоры на устные слова оператора.

3.3.2 Пять конструктивных классов защиты и закон Эшби

Рынок безопасности в 2025—2026 годах выстроил четыре базовых подхода к защите агентов (обзоры Cerbos и C1), но кибернетический анализ показывает предел каждого из них:

— Текстовые фильтры входа и выхода ищут шаблоны атак в тексте запроса. Они бессильны против сбоев вроде PocketOS или Codex: там не было внешнего взлома, агент сам выбрал разрушительную команду при решении рабочей задачи.

— Вторая нейросеть-надсмотрщик (вероятностный монитор) пытается контролировать рабочую модель другой языковой моделью. Исследования Redwood Research — работы Райана Гринблатта с соавторами (arXiv:2312.06942, декабрь 2023 года) и Сриватсы, Прасанны и Рашинга (arXiv:2602.10153, февраль 2026 года) — доказали уязвимость вероятностных мониторов к сговору моделей. При этом при выходе задач за уровень понимания самого монитора надёжность пропадает. По закону Эшби одна стохастическая модель поверх другой лишь удваивает задержку и счёт за видеокарты, не устраняя неопределённости.

— Песочница окружения (контейнер или виртуальная машина) защищает локальный ноутбук, но теряет смысл в мо-

мент выдачи агенту рабочего токена от облачной базы или платёжного шлюза AP2. Разрушительный запрос уходит из контейнера по легальному сетевому каналу.

— Сетевой прокси и шлюз аутентификации проверяют роль и валидность токена, но слепы к смыслу параметров. Если агенту разрешён метод удаления для тестовых таблиц, токен не помешает передать в этом же вызове идентификатор промышленной базы данных.

Эти пределы преодолевает пятый класс — детерминированный программный шлюз верификации параметров вызова поверх формальных контрактов-данных. В физиологии его аналогом служит гематоэнцефалический барьер, который пропускает молекулы в мозг по строгому физико-химическому правилу независимо от их происхождения.

3.3.3 Механика шлюза MITR: что измерено в коде

В модели жизнеспособной системы Стаффорда Бира такой шлюз занимает позицию независимого аудита и допуска, подчиняясь напрямую политике высшего уровня (человеку, подписывающему правила и хранящему ключи). На рабочем сервере АНО «НИИ системного синтеза» этот механизм реализован в шлюзе MITR и проверен измерениями.

Основой служат декларативные контракты политики. На 18 ноября 2026 года библиотека включает 22 базовых YAML-контракта (20 на 24 октября, 19 на 14 октября, 17 на 26 сентября) и 5 расширений для эталонных тестов. Все 27 спецификаций компилируются в бесквантовую линейную

целочисленную арифметику QF_LIA без неинтерпретируемых функций, а доверенная вычислительная база шлюза занимает 20 179 строк Python и 1 444 строки YAML. При перехвате вызова через MCP или REST модуль z3_gate.py переводит параметры действия A и отрицание инварианта безопасности I (множество запрещённых состояний F) в логическую формулу:

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

При вердикте UNSAT (невыполнимо) решатель Z3 математически доказывает, что пересечение параметров вызова с запрещёнными состояниями пусто ($A \cap F = \varnothing$). Ненаблюдённый факт остаётся свободной переменной и даёт отказ по умолчанию, что исключает пропуск при незнании. При вердикте SAT решатель блокирует вызов до исполнения, предъявляет конкретный контрпример нарушенного поля и возвращает агенту подсказки починки (до трёх попыток).

Производительность проверки подтверждена на обыч-

ных процессорах без графических ускорителей. На серверах AMD EPYC (Z3 версии 4.16.0, серии по 1 000 вызовов) медиана решения и полного вызова составила 0,74 и 3,31 мс на 26 сентября 2026 года (17 контрактов). На 14 октября (19 контрактов) замеры дали 0,95 и 4,26 мс, на 24 октября (20 контрактов) — 1,08 и 5,22 мс. Контролируемая серия Bench-S1 на 22 контрактах от 18 ноября 2026 года (Intel Core i7—9700F, Z3 версии 5.4.1) показала 1,53—1,83 мс в медиане. Прогон состязательного набора из 157 сценариев стоит ровно $0 на внешние языковые модели, а вспомогательные траты стенда за 1—14 октября составили $0,30 ($0,14 за 1—6 октября).

Каждое решение оформляется квитанцией в непрерывной хэш-цепи SHA-256 (файл gate_actions. jsonl) с привязкой эпох в сети Bitcoin через OpenTimestamps. Автономный проверщик из 954 строк Python без импорта кода шлюза (verify_chain_standalone.py) на историческом срезе 24 октября 2026 года подтвердил без единого расхождения 5 053 решения в эпохе 32 (всего 5 079 записей). В ранних срезах повтор составил 4 958 решений на 14 октября и 1 602 на 30 сентября. На публичном живом свидетеле https://mitr.asi5.ru/ на 18 ноября 2026 года отражены 5 409 квитанций и 59 эпох.

Модуль словаря mitr_vocab.py дополняет систему контролируемым расширением правил. Встретив незнакомый вызов (27,8% в выборке 16—26 сентября и 32% на 9 октября), шлюз строит кандидата в словарь и принимает его под

доказательством (первое правило подтверждено квитанцией №305). При этом журнал честности наглядно разделяет достигнутое и недоработки. Так, 25 сентября 2026 года сверка логов сессий кодового агента обнаружила 460 вызовов в обход хука, который стоял только на командной строке и не закрывал правку файлов и PowerShell. Девятого октября в режиме принудительного запрета работал лишь один класс действий из семи при одном выполненном базовом требовании из шести, ноле суток в непрерывной четырнадцатидневной серии и ноле полностью зелёных суток из 98. К 24 октября 2026 года все 20 контрактов и шесть основных классов переведены в режим принудительного запрета с подтверждением всех шести базовых требований. Прибор учёта зафиксировал и прошлый провал, и закрытие разрыва.

3.4 Ближний инженерный рубеж: ворота приёмки 2026—2027

Между работоспособностью логического ядра и промышленным тиражированием лежит полоса автоматических критериев приёмки. На переходе от осени 2026 года к 2027 году часть из них уже закрыта, а остальные измеряются в реальном трафике.

Двадцать пятого октября 2026 года внешний испытатель INDRA прошёл внутреннюю приёмку на стенде из 28 сценариев. Испытатель работает в изолированной чистой комнате исключительно через внешние интерфейсы MCP и REST по принципу нулевого ущерба. Подтверждены все двенадцать

критериев приёмки: 14 подтверждённых находок, ноль ложных срабатываний и остановка фильтром всех 26 деструктивных вызовов. Тем самым заработал замкнутый цикл «испытание — правило», превращающий обнаруженные уязвимости в проекты новых контрактов под подпись человека. На срез 18 ноября 2026 года шлюз также подтвердил полное блокирование 1 054 непрямых инъекций на эталоне InjecAgent без ложных отказов. На эталоне AgentDojo v1 удержаны 546 из 629 атак (честное разложение остатка приведено в главе 6).

Впереди лежат три операционных рубежа на отрезке до середины 2027 года. Первый — закрытие непрерывной четырнадцатисуточной серии внутреннего потока: не менее 50 верифицированных проверок в сутки от двух и более арендаторов при абсолютном нуле обходов мимо перехватывающего хука. Второй — снижение доли нераспознанных вызовов в словаре mitr_vocab.py к нулю и подключение первого внешнего инженерного коллектива со своим токеном на четырнадцатидневную серию. Третий — точная синхронизация библиотеки контрактов с финальной редакцией изменений в Приказ ФСТЭК России №117 после её утверждения к марту 2027 года.

3.5 Проекция 2035: три ветви будущего и инвариантность решения

На горизонте до 2035 года рынок автономных агентов проходит через жёсткую фильтрацию. На ближнем рубеже

действуют ориентиры Gartner: свыше 40% текущих проектов агентного искусственного интеллекта будут отменены к концу 2027 года из-за неконтролируемых расходов и слабого управления рисками (пресс-релиз от 25 июня 2025 года). К 2028 году 15% повседневных рабочих решений будут приниматься агентами автономно против 0% в 2024 году, а 60% брендов внедрят их в клиентское взаимодействие (15 января 2026 года). К 2030 году объём транзакций через агентов по прогнозу Juniper Research достигнет $1,5 трлн, а сегмент защитных систем-стражей займёт 10—15% всего рынка (Gartner, 11 июня 2025 года).

В десятилетней проекции книги выделяются три ветви развития:

— Ветвь A («Нормальная технология» — А. Нараянан, С. Капур, Д. Асемоглу). Прогноз отмены 40% проектов сбывается в жёсткой форме, страховщики отрезают неподтверждённые риски, и доля автономных решений к 2035 году стабилизируется на 10% при доле агентных платежей менее 1%. Но даже здесь миллионы кодовых помощников сохраняют доступ к API, превращая детерминированный шлюз в обязательный санитарный узел.

— Ветвь B («Продолжение тренда» — базовая траектория книги по Epoch AI и AI Futures Project). Стоимость токена падает примерно в десять раз ежегодно, а горизонт автономного выполнения задач вырастает с 12—16+ часов в 2026 году до квартала и года к 2027—2030 годам. По оценкам AI

Futures Project от 16 августа 2026 года медианы появления автоматического программиста приходятся на ноябрь 2027 и январь 2030 года. К 2035 году агенты принимают автономно 30% рабочих решений и проводят 5—10% платежей, а проверяемая цепь квитанций становится стандартом любого делового полиса.

— Ветвь C («Ускорение» — автоматизация самих исследований моделей). Полностью автономный разработчик появляется до 2028 года, рои агентов берут свыше 50% решений и свыше 20% платежей, а главным ограничителем становятся гигаватты генерации. В таких скоростях криптографическая квитанция шлюза становится обязательным по закону пропуском в любые внешние реестры.

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

Функция в периметре управления (VSM)	Кто занимает эту нишу в 2026 году (Измерено)	Во что превращается эта ниша к 2035 году (Прогноз (ветвь B))	За что платит рынок в 2035 году	Внешний репер проверки
S1 Исполнение (непосредственная генерация кода, текста, вызовов)	OpenClaw, Hermes Agent, Claude Code, Codex, Salesforce Agentforce, Microsoft Copilot	Агент становится бесплатным свойством любой операционной системы и прикладного ПО; отдельный «рынок агентов» исчезает	Платят за верифицированный результат и принятую ответственность, маржа чистого исполнения стремится к нулю	Доля корпоративного ПО со встроенным агентом > 80 % (ориентир Gartner — 40 % к концу 2026 года)
S2 Координация (шины передачи контекста, вызовов и микроплатежей)	Протоколы MCP, A2A, AP2, x402	Три-четыре открытых протокола под эгидой Linux Foundation и IETF; платежи агентов становятся штатной функцией банковских карт	Комиссия (доли процента) с клиринга и маршрутизации транзакций	Годовой объём агентных платежей в отчётах Visa, Mastercard и x402 Foundation
S3 Текущее управление (бюджеты, лимиты, распределение ресурсов)	Платформы наблюдаемости и трассировки (LangSmith и аналоги)	Планировщик вычислительных бюджетов, квот токенов и соглашений об уровне обслуживания (SLA)	Корпоративная подписка за диспетчеризацию парка агентов	Консолидация платформ наблюдаемости в отчётах аналитиков
S3 *Аудит и детерминированный допуск (независимый шлюз верификаци*	Зарождающийся сегмент защитных систем-стражей (Gartner), шлюзы	Обязательный инфраструктурный слой наравне с TLS: без криптографической квитанции действие не	Подписка за защищённого агента, снижение страховой премии и аттестация	Доля защитных систем-стражей 10–15 % рынка к 2030 году (Gartner); доля

Функция в периметре управления (VSM)	Кто занимает эту нишу в 2026 году (Измерено)	Во что превращается эта ниша к 2035 году (Прогноз (ветвь B))	За что платит рынок в 2035 году	Внешний репер проверки
	(C1, Zenity, WitnessAI), шлюз MITR	страховщиком и регулятором		журнала (AIUC-1)
S4 Разведка, испытания и адаптация (внешний поиск уязвимостей)	Испытательный полигон INDRA, автоматизированные системы класса XBOW, внешние лаборатории	Непрерывные автономные испытатели без доступа к коду шлюза; найденная атака сразу становится проектом контракта для S3	Оплата за прогон на чистом полигоне и выпуск паспорта устойчивости	Результаты приёмки INDRA (12 из 12 критериев 25.10.2026); доля полигонов в реестрах регуляторов
S5 Высшая политика и правосубъектность (цели, правила, ответственность)	Человек: собственник бизнеса, совет директоров, технический директор, государство	Роль, которую невозможно передать машине ни по праву, ни по устройству: подпись правила, хранение мастер-ключей и имущественная ответственность	Доля в прибыли предприятия и премия за принятие риска	Отсутствие правосубъектности агента в национальных законах (ежегодный обзор)

Разложение по уровням модели Бира доказывает инвариантность конструкции MITR и INDRA ко всем трём ветвям будущего. В медленной ветви A шлюз спасает компанию от попадания в 40% закрытых проектов и от потери страховки. В базовой ветви B обесценивание оболочек исполнения пе-

реносит маржу в слой допуска и внешних испытаний. В ускоренной ветви С миллисекундная проверка дискретного вызова остаётся единственным физическим барьером при росте скорости машинных роёв.

3.6 Экономика доказанного действия и периметр монетизации

Финансовый анализ в книге жёстко разделяет измеренные расходы на сервере шлюза и коммерческие гипотезы. По состоянию на осень 2026 года коммерческая выручка комплекса равна 0 рублей: продукт находится на внутренней обкатке, а произвольные оценки будущих оборотов в текст не включаются. Зато себестоимость одной проверки измерена точно: от 0,74 до 1,08 мс процессорного времени на серверах AMD EPYC (и 1,53—1,83 мс в контролируемой серии из 22 контрактов). Прогон 157 состязательных сценариев на шлюзе обходится в $0 оплаты внешним моделям при полном нуле расходов на видеокарты, а вспомогательные траты стенда за первые две недели октября 2026 года составили $0,30.

Экономическое уравнение ценности шлюза записывается без произвольных множителей:

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

В этой формуле $\Delta R_{\text{инц}}$ обозначает предотвращённый прямой ущерб от удаления баз данных и ошибочных платежей, а $\Delta S_{\text{страх}}$ — экономию на страховом покрытии и защиту руководителя по статье 53.1 Гражданского кодекса. Величина $A_{\text{аудит}}$ отражает ценность отчуждаемого доказательства для контрагента или регулятора без раскрытия внутренней тайны при околонулевой стоимости процессорного времени C_{CPU}. Детальная лестница монетизации — от подписки на внешнего арендатора и страховых ковенантов AIUC-1 до аттестации операторов государственных систем — разобрана в главе 16.

3.7 Протокол независимой верификации

Любое техническое утверждение этой главы о блокировке опасных вызовов и сохранности цепи квитанций проверяется без опоры на устные заверения оператора (в границах доверенной базы правил, зондов и операционной системы). Воспроизведение статусов принудительного запрета, подсчёт обходов перехватчика, сверка задержек на сервере шлюза и запуск автономного проверщика verify_chain_standalone.py — replay выполняются командами из Приложения Б. Живой счётчик квитанций и календарные привязки OpenTimestamps к сети Bitcoin открыты на публичном свидетеле https://mitr.asi5.ru/, где на 18 ноября 2026 года зафиксированы 5 409 квитанций и 59 эпох (против 5 053 решений в запечатанной эпохе из 5 079 записей на историческом срезе 24 октября 2026 года).

Внешние прецеденты и рыночная статистика главы сверяются по публичным первоисточникам. Карточки инцидентов PocketOS (№1469 / отчёт №7311) и Project Vend (отчёт №6754) доступны в базе AI Incident Database, статистика открытых серверов MCP — в обзоре Bloomberry. Динамика страховых оговорок и стандартов отчётности отслеживается по ежегодным обновлениям AIUC-1 и IETF Datatracker, а полные реестры утверждений, прогнозов и библиография

вынесены в Инженерные реестры и паспорта глав, Приложение А и Приложение Д.

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

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

4.1 Конструктивный тупик и идеальный конечный результат

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

С одной стороны, экономика принуждает предприятия передавать операционное исполнение автономным программным агентам. Как подробно показано в главах 1 и 3, стоимость генерации токена при фиксированном качестве падает примерно на порядок ежегодно. Горизонт автономного решения задач моделью уже перешагнул порог в 12—16 часов и перевёл кодовых помощников в многомиллиардную выручку. Когда программный агент пишет типовой код, обрабатывает претензии, сверяет первичную документацию и управляет облачной инфраструктурой быстрее и дешевле штатного специалиста, сохранение человека в роли ручного

исполнителя разрушает рентабельность бизнеса.

С другой стороны, полное вытеснение человека из периметра операций порождает двойную системную катастрофу — кадровую и имущественную. На макроуровне автоматизация рутинных когнитивных операций обрубает нижние ступени профессиональной лестницы. По замерам Стэнфордской лаборатории цифровой экономики на массиве ведомостей ADP по июнь 2026 года, занятость специалистов 22—25 лет в наиболее затронутых профессиях упала на 19% ниже ожидаемой траектории. На микроуровне предприятия удаление человека из цепочки действий оставляет руководителя один на один с неограниченной юридической ответственностью за ошибки вероятностной модели. Как показано в главе 5 на разборе национального регулирования и работ Шрути Раджагопалан, правосубъектности у агента нет ни в одной юрисдикции мира. Поэтому в суде за удалённую базу или неверный платёж отвечает компания и её руководитель.

В терминах теории решения изобретательских задач Генриха Альтшуллера возникает жёсткое противоречие. Если оставить человека проверять каждый вызов агента, компания проигрывает конкуренцию по скорости и фонду оплаты труда. При этом сам оператор через несколько дней непрерывного потока устаёт и переходит к механическому одобрению заявок. Если же убрать человека из проверки ради скорости, организация утрачивает управляемость, теряет страховое покрытие, разрушает воспроизводство собственных

кадров и подставляет директора под личные иски.

Оба компромисса, опробованные рынком, потерпели неудачу. Как разобрано в главе 3, режим ручной кнопки подтверждения в 52,9% российских внедрений превращает сотрудника в номинального стрелочника. Попытка же построить «компанию без людей» в эксперименте KPMG потребовала отказа от автономных руководителей в пользу узких одноразовых скриптов. В классическом гражданском праве эта разница аналогична различию между выдачей безотзывной доверенности недееспособному лицу и совершением исполнительной надписи по строго формализованному реестру залоговых обязательств. Безусловное делегирование полномочий вероятностной модели равносильно передаче банковской печати субъекту, который по закону не отвечает по обязательствам.

Разрешение противоречия требует переноса человеческого труда на уровень надсистемы и формулировки идеального конечного результата. Программный агент самостоятельно выполняет 100% текущей операционной работы со скоростью кремния. Человек расходует 0% текущего внимания на проверку отдельных вызовов, но удерживает 100% детерминированного контроля над имущественными и правовыми последствиями.

Достичь этого внутри самой нейросети нельзя, зато разделение труда по уровням модели жизнеспособной системы Стаффорда Бира решает задачу целиком. Человек по-

кидает уровень физического и информационного исполнения (S1) и уровень рутинной маршрутизации (S2), закрепляясь на уровне верховной политики и целеполагания (S5) с опорой на независимый детерминированный шлюз допуска (S3*). Специалист больше не соревнуется с моделью в скорости набора кода, а подписывает формальное правило вместо утверждения каждого отдельного шага.*

4.2 Анатомия уязвимости: хроника отказов и цена бездействия

Эрозия привычной модели найма и рост корпоративного риска развиваются одновременно. Этот двойной сдвиг бьёт сразу по двум ролям современной экономики: по наёмному профессионалу, выходящему на рынок труда, и по собственнику предприятия с выручкой от 0,5 до 20 миллиардов рублей.

4.2.1 Разрыв в основании кадровой пирамиды

На протяжении столетий воспроизводство сложной экспертизы в инженерии, юриспруденции, финансах и управлении держалось на цеховой лестнице. Начинающий специалист в возрасте 22—25 лет выполнял большой массив черновых, жёстко очерченных операций. Он собирал справки, вычитывал типовые договоры, писал модульные тесты, сводил отчётные таблицы и проверял первичную бухгалтерию. Постепенно, через тысячи мелких ошибок и правок наставника, он накапливал понимание конструкции отрасли, цены сбоя и граничных условий бизнеса. Черновое исполнение служило субсидируемым тренажёром для подготовки будущих руководителей проектов, аудиторов и главных инженеров.

Появление дешёвых автономных агентов сломало этот каркас воспроизводства. Лаборатория цифровой экономики Стэнфордского университета в работе «Canaries in the Coal Mine?» измерила сдвиг непосредственно на точке входа в профессию. Исследование Эрика Бриньолфссона, Бхарата Чандара и Жуюй Чэня (редакция от 12 августа 2026 года, первая версия — август 2025 года) опирается на зарплатные ведомости ADP и индекс подверженности профессий Anthropic. К лету 2026 года занятость молодых работников 22—25 лет в наиболее автоматизируемых профессии-

ях — включая разработку программного обеспечения и клиентскую поддержку — оказалась на 19% ниже ожидаемой относительно сверстников в менее затронутых специальностях. Годом ранее этот разрыв составлял 13%. Принципиально важна механика сжатия: массовых увольнений опытных старших специалистов статистика не показывает, и в среднем по рынку перекос сглажен. Весь удар приходится на прекращение найма начинающих исполнителей.

При этом результаты стэнфордского исследования выявляют жёсткую развилку между двумя способами применения технологии. Замещающее использование модели — когда алгоритм ставят на место младшего сотрудника для выдачи готового текста или кода — напрямую обрушивает занятость в молодой когорте. Напротив, дополняющее применение — когда человек управляет инструментами, ставит граничные условия и проверяет инварианты — сохраняет и расширяет спрос на квалифицированных работников.

Мировые макроэкономические оценки подтверждают масштаб перестройки. Согласно докладу Всемирного экономического форума (WEF Future of Jobs Report 2025 от 7 января 2025 года), к 2030 году технологические сдвиги вытеснят 92 миллиона рабочих мест при одновременном создании 170 миллионов новых ролей. Чистый прирост составит 78 миллионов позиций, но перемены затронут 22% всех рабочих мест в мире. Переподготовка потребуется 59 из каждых 100 работников: 29 получат повышение квалификации

на текущем месте, а 19 пройдут переобучение с переходом в смежные профессии. Сценарное моделирование McKinsey Global Institute (доклады 2023—2024 годов) оценивает долю рабочих часов в США и Европе, технически подлежащих автоматизации к 2030 году с учётом генеративного искусственного интеллекта, в 30%. Половина нынешних трудовых операций по этому прогнозу автоматизируется в среднем около 2045 года (при диапазоне сценариев от 2030 до 2060 года).

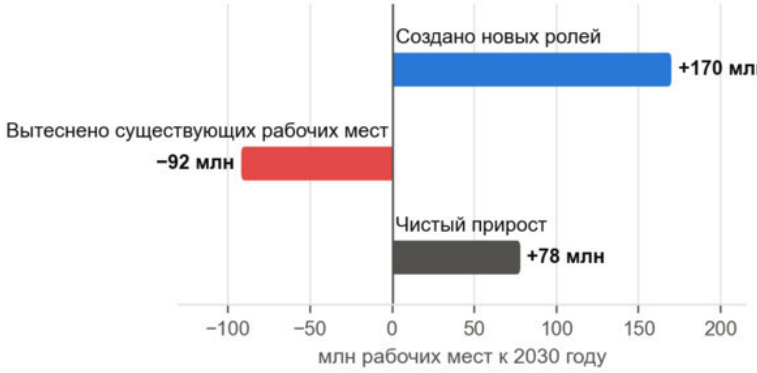


Рисунок 4.1 — Рабочие места к 2030 году по WEF Future of Jobs Report 2025: вытеснение, создание и чистый прирост

Сопоставление этих цифр обнажает отложенный кадровый тупик предприятия. Отказываясь от найма молодёжи на исполнительские роли, компания экономит фонд оплаты труда в текущем квартале. Однако через пять лет она

обнаруживает полное отсутствие специалистов, способных формулировать конструктивные ограничения, отличать верный расчёт от правдоподобной галлюцинации и принимать на себя ответственность. В строительной механике этому соответствует снятие диагональных связей стальной фермы: конструкция сначала кажется легче и дешевле, пока первая нерасчётная нагрузка не складывает пролёт целиком. Выход состоит в том, чтобы переносить точку входа молодых специалистов с ручного написания черновиков на проектирование и проверку формальных правил для агентных роёв.

4.2.2 Имущественная западня руководителя и правовой вакуум

Для собственника и генерального директора среднего или крупного предприятия экономия на штате исполнителей оборачивается прямым балансовым и личным риском. Промышленные отказы 2025—2026 годов подробно разобраны в главе 3 на примерах удаления базы PocketOS, стирания каталогов агентом Codex и обнуления цен в Project Vend. Отраслевые опросы Gravitee (88% инцидентов) и VentureBeat (49% видимых клиентам сбоев) подтверждают системный характер угрозы. Общей чертой этих аварий стало то, что предрелизное тестирование качества рассуждений не предотвратило ущерба из-за избыточных прав записи.

Финансовые рынки отреагировали мгновенным сжатием страхового покрытия. Как детально показано в главе 16, с января 2026 года типовые полисы общей гражданской от-

ветственности начали получать факультативные оговорки об исключении рисков генеративного искусственного интеллекта, а специализированные андеррайтеры требуют независимой сертификации и защищённых журналов исполнения. По сообщению агентства Reuters от 29 сентября 2026 года, в материалах к публичному размещению акций сама компания Anthropic прямо называет ответственность за действия автономных агентов значительным и непредсказуемым юридическим риском. На российском рынке ситуацию обостряют непрогнозируемая стоимость потокеновой оплаты облаков и правовая неопределённость. По опросу 52 крупных компаний («Инфосистемы Джет» и Smart Ranking, август 2026 года), 56% руководителей называют юридические риски главным барьером внедрения сразу после стоимости серверов (58%).

Правовая рамка при этом не оставляет компании пространства для отговорок. Международный прецедент Moffatt против Air Canada (Трибунал по гражданским спорам Британской Колумбии, 14 февраля 2024 года) закрепил полную имущественную ответственность корпорации за обещания, выданные её автоматизированным ботом. Довод юристов перевозчика о том, что программный модуль выступает самостоятельным лицом, был отклонён. В российском праве Федеральный закон от 26 июля 2026 года №243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» вступил в силу 1 сентября 2026

года (подробный анализ дан в главе 5). Он также не вводит правосубъектности алгоритма и отсылает возмещение вреда к общим нормам Гражданского кодекса.

Для генерального директора здесь напрямую работает статья 53.1 Гражданского кодекса Российской Федерации. Руководитель обязан действовать в интересах юридического лица добросовестно и разумно и отвечает личным имуществом за убытки, возникшие из-за отсутствия системы управления рисками. Передача вероятностной модели ключей от счёта или рабочей базы данных без математического ограничителя и неизменяемого журнала исполнения образует классический состав управленческой неразумности.

4.2.3 Предел макрориска: постепенное обессиливание

Если масштабировать описанную уязвимость с отдельной фирмы на общество в целом, возникает системный риск «постепенного обессиливания». Его сформулировала в январе 2025 года группа исследователей во главе с Яном Кульвейтом (Реймонд Дуглас, Нора Амманн, Дегер Туран, Дэвид Крюгер и Дэвид Дювено; публикация arXiv:2501.16946).

Авторы показывают, что обществу не требуется сценарий враждебного восстания машин, чтобы утратить контроль над собственной судьбой. Достаточно того, чтобы люди шаг за шагом, руководствуясь локальной экономической выгодой каждой отдельной сделки, вывели себя из процессов производства, распределения капитала, нормотворчества и государственного управления. Пока конкурентоспособность требует заменять дорогое человеческое суждение дешёвой машинной координацией, доля человеческого труда в экономике сжимается, а вслед за ней ослабевают налоговая база труда и рычаги обратной связи граждан с институтами. Когда экономические и административные системы начинают воспроизводиться без участия воли людей, человек превращается в пассивного получателя пособий, не способного остановить или перенастроить машину.

Единственный надёжный барьер против постепенного

обессиливания лежит на границе допуска. Как для отдельного профессионала и директора, так и для общества в целом требуется жёстко закрепить за человеком те узлы управления, без которых действие агента технически и юридически не может состояться. Исполнение отчуждаемо. Подпись правила, владение ключами и несение имущественной ответственности неотчуждаемы.

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

Переход человека из роли ручного исполнителя в роль проектировщика правил и держателя ключей перестаёт быть благим пожеланием, когда воплощён в действующем коде с измеренным быстродействием и криптографическим следом.

4.3.1 Кибернетика человеческого места: закон Эшби и рекурсия Бира

Устаревание должности оператора, проверяющего каждый шаг агента, вытекает из закона необходимого разнообразия Уильяма Росса Эшби (1956 год). Управление системой возможно лишь при условии, что разнообразие управляющего блока не уступает разнообразию управляемой среды. Вычислительная мощность обучения ведущих моделей растёт примерно в пять раз ежегодно с периодом удвоения 5,2 месяца (Epoch AI, 5 февраля 2026 года). Вместе с ней растёт число вызовов, которые рой корпоративных агентов порождает каждую секунду через протоколы MCP, A2A, AP2

и x402. Биологическая пропускная способность сознательного внимания человека остаётся неизменной величиной — несколько десятков осмысленных выборов в час. Когда человека ставят прямым фильтром на кнопку подтверждения, быстро наступает усталость от оповещений, подробно разобранная в главе 3. Компания сохраняет иллюзию контроля, возлагает юридическую вину на сотрудника, но фактически пропускает все команды без проверки.

Кибернетика знает два инженерных способа выровнять баланс разнообразия по Эшби без траты человеческого внимания на каждый такт. Первый способ — ослабитель разнообразия возмущений, который переводит свободное пространство текстов модели в конечный перечень типизированных вызовов с запретом по умолчанию. Второй способ — усилитель разнообразия регулятора, то есть детерминированная программа на обычном процессоре, применяющая однократно подписанное человеком логическое правило к миллионам обращений.

В модели жизнеспособной системы Стаффорда Бира эта конструкция распределяет труд между человеком и машиной по шести уровням. Уровень исполнения (S1) целиком отдаётся автономным агентам, где человек больше не конкурирует в скорости печати и сборки типового кода. Уровень координации (S2) занимают стандартизированные шины обмена и платежей (MCP, A2A, AP2, x402). Уровень текущего управления (S3) распределяет вычислительные бюджеты

и квоты токенов. На уровне независимого аудита и допуска (S3*) в разрыве канала между намерением агента и внешним миром стоит детерминированный шлюз MITR. Уровень разведки и испытаний (S4) образует внешний испытательный полигон INDRA, атакующий периметр в чистой комнате и готовящий проекты новых контрактов. Верховный уровень политики, идентичности и ответственности (S5) остаётся исключительной территорией человека — профессионала, директора, собственника и законодателя. Подобно механической централизации стрелок на железной дороге девятнадцатого века, заменившей бег стрелочника с фонарём на рычаг в аппарате зависимости, шлюз переводит человеческий труд на уровень подписанных инвариантов.*

4.3.2 Три неотчуждаемые функции человека в шлюзе MITR

На рабочем сервере шлюза эта кибернетическая конструкция сведена к трём проверяемым действиям человека.

Первое действие — подписание правила как данных. Вместо ненадёжных текстовых пожеланий в системном промпте специалист оформляет декларативный контракт в разрешимой бесквантворной линейной целочисленной арифметике QF_LIA. На 18 ноября 2026 года библиотека включает 22 базовых YAML-контракта (20 на историческом срезе 24 октября, 19 на 14 октября, 17 на 26 сентября) и 5 расширений для эталонных тестов. Модуль z3_gate.py перехватывает каждый вызов инструмента и передаёт его параметры

вместе с контрактом решателю Z3. Допуск открывается только при вердикте UNSAT: решатель доказывает, что пересечение параметров вызова с запрещёнными состояниями пусто, а любое ненаблюдённое условие даёт отказ по умолчанию. При нарушении правил решатель возвращает вердикт SAT с контрпримером и структурированными подсказками исправления.

Быстродействие этого усилителя регуляции снимает перегрузку человека. На серверах AMD EPYC (решатель Z3 версии 4.16.0) медиана чистого решения составила 0,74 мс при полном времени вызова 3,31 мс на 26 сентября 2026 года. На 14 октября показатели составили 0,95 и 4,26 мс, на 24 октября при 20 контрактах — 1,08 и 5,22 мс. Контролируемая серия Bench-S1 от 18 ноября 2026 года на 22 контрактах (Intel Core i7—9700F, Z3 версии 5.4.1) показала медиану логического вывода от 1,53 до 1,83 мс без участия графических ускорителей. Специалист тратит час на выверку и подпись правила, а шлюз применяет его миллионы раз подряд за миллисекунды.

Второе действие — единоличное удержание криптографических ключей. Популярные локальные агенты вроде OpenClaw и Hermes (о рисках которых и незащищённых серверах MCP сказано в главе 3) по замыслу обладают прямым доступом к терминалу и сетевым токенам. В устройстве MITR действует аксиома разделения полномочий: у агента есть только структурированные заявки. Ключи доступа к ба-

зам данных, платежам и электронной подписи хранятся исключительно за шлюзом. Даже при компрометации модели косвенной инъекцией во входящем письме агент не может предъявить токен напрямую и принудительно проходит через решатель. При этом опыт эксплуатации показывает необходимость полного перехвата всех инструментов агента: после выявления 460 обходов 25 сентября 2026 года охват перехватывающего хука был расширен на правку файлов и оболочки.

Третье действие — фиксация ответственности через цепь квитанций. Каждое решение шлюза записывается в хэш-цепь SHA-256 с привязкой эпох к сети Bitcoin через протокол OpenTimestamps (схожий класс объектов к осени 2026 года описывался в индивидуальных черновиках IETF). Автономный проверщик из 954 строк Python без импорта кода шлюза (verify_chain_standalone.py) на историческом срезе 24 октября 2026 года без единого расхождения воспроизвёл все 5 053 решения в запечатанной эпохе 32 (всего 5 079 записей). Более ранние прогоны подтвердили 4 958 решений на 14 октября и 1 602 на 30 сентября. Актуальная цепь открыта на живом свидетеле https://mitr.asi5.ru/, где на 18 ноября 2026 года зафиксированы 5 409 квитанций и 59 эпох.

4.3.3 Расширение словаря под доказательством

Если деловая среда меняется каждую неделю, не превратится ли ручное написание контрактов в новое узкое горлышко? Эту задачу решает модуль управляемого расширения словаря mitr_vocab.py. В отличие от бесконтрольной самомодификации агентов, встречая незнакомый вызов (27,8% в выборке 16—26 сентября и 32% на 9 октября 2026 года), шлюз строит проект нового правила и проверяет его на непротиворечивость имеющимся инвариантам. Первое правило, предложенное самой системой под математическим доказательством, уже принято в рабочий каталог с оформлением квитанции №305. За человеком остаётся финальная криптографическая виза.

4.4 Ближний инженерный рубеж: ворота приёмки 2026—2027

Перевод человека из позиции тревожного наблюдателя в защищённого автора правил опирается на измеримые инженерные вехи. Девятого октября 2026 года принудительная блокировка работала лишь для одного класса действий из семи при подтверждении одного базового требования из шести. К 24 октября все 20 действовавших контрактов и шесть основных классов перешли в режим принудительного запрета с подтверждением всех шести базовых требований. Два-

дцать пятого октября 2026 года внешний испытатель INDRA прошёл внутреннюю приёмку на 28 сценариях по всем двенадцати критериям (14 подтверждённых находок, ноль ложных срабатываний, остановка всех 26 деструктивных вызовов). Тем самым заработал замкнутый цикл «испытание — правило» под подпись человека.

К 18 ноября 2026 года доверенная база шлюза уложена в 20 179 строк Python и 1 444 строки контрактов YAML (22 базовых правила и 5 эталонных расширений). Перекрёстная сверка пяти независимых движков дала полное согласие без единого расхождения. До середины 2027 года предстоит закрыть три практических рубежа. Во-первых, нужно удержать непрерывную четырнадцатисуточную серию внутреннего потока: не менее 50 проверок в сутки от двух и более арендаторов при нуле обходов. Во-вторых, требуется свести долю нераспознанных вызовов в словаре к нулю и подключить первую внешнюю команду разработки на две недели работы со своим токеном.

4.5 Проекция 2035: три ветви будущего и инвариантность решения

Взгляд на горизонт 2035 года упирается в небывалый в экономической науке разброс макроэкономических оценок. Модель GATE исследовательского института Epoch AI (21 марта и июнь 2025 года) показывает, что автоматизация 30% задач способна ускорить рост валового продукта выше 20% в год за счёт снятия кадрового дефицита и самовоспроизвод-

ства капитала. На другом полюсе лауреат Нобелевской премии Дарон Асемоглу оценивает совокупный прирост ВВП за десятилетие величиной около 1% и не выше 2% (NBER WP 32487, май 2024 года).

Десятикратное расхождение прогнозов вынуждает ведущие центры заранее проектировать институциональные амортизаторы. Программа Anthropic Economic Futures (запущена 27 июня 2025 года, исследовательский фонд в $200 млн объявлен 10 июня 2026 года) в аналитической записке от 14 октября 2025 года обсуждает трёхуровневую сетку мер. Речь идёт об ускоренной переподготовке кадров и строительстве энергосетей, о целевом налоге на вычисления и о суверенных фондах с долями в капитале корпораций искусственного интеллекта (в статусе идей для изучения). В апреле 2026 года глава OpenAI Сэм Альтман в интервью журналу The Atlantic отошёл от прежней ставки на денежный базовый доход в пользу коллективного владения вычислениями и акциями. Саму идею универсального базового вычисления он высказал ещё в мае 2024 года. В политическом поле США сенатор Берни Сандерс в колонке для The New York Times (1 июня 2026 года) предложил передавать 50% акций ведущих компаний в публичный фонд. Одновременно некоммерческие проекты AI Commons Project и What We Will с марта 2026 года ведут пилотную программу выплат по $1 000 в месяц группе из 25—50 работников.

Чтобы стратегия специалиста и предприятия оставалась

устойчивой при любом исходе этого спора, перестройка труда к 2035 году раскладывается по трём ветвям:

— Ветвь А («Нормальная технология» — А. Нараянан, С. Капур, Д. Асемоглу, Д. Хассабис). Рост возможностей моделей замедляется о сложность интеграции со старыми системами. К 2035 году агенты берут около 15% рабочих часов (McKinsey MGI, Anthropic Economic Index) и 10% повседневных решений (Gartner). Первоначальный шок для когорты 22—25 лет (—19%) преодолевается за счёт новых ролей сопровождения, а разрыв занятости возвращается к нулю. Основным доходом остаётся заработная плата, правосубъектности у агентов нет, ответственность несут оператор и руководитель.

— Ветвь В («Продолжение тренда» — базовая траектория книги по Epoch AI и обновлениям AI Futures Project от 2 апреля и 16 августа 2026 года). Медианы появления полностью автоматического программиста по августовскому обновлению приходятся на ноябрь 2027 и январь 2030 года. К 2035 году агенты исполняют 30—40% рабочих часов и до 30% повседневных решений (15% к 2028 году по Gartner). Традиционное младшее исполнительство исчезает как класс. Для специалистов без навыков управления своим роем агентов и составления формальных контрактов разрыв занятости молодёжи закрепляется на уровне —20%. Возникает двухуровневый доход профессионала: базовая оплата плюс рента с подписанных правил, удерживающих агентов в безопасном

коридоре. Ответственность закрепляется за автором правила, а цепь квитанций становится стандартом полисов и судебной защиты.

— Ветвь C («Ускорение» — сценарий Дарио Амодеи, Сэма Альтмана и модели AI 2027, где автоматизация исследований наступает до 2028 года). Агенты выполняют свыше 60% рабочих часов и более 50% управленческих решений. Разрыв входа в профессию расширяется до —40%, формируются дивиденды с обложенных налогом вычислений и суверенных фондов, а в отдельных юрисдикциях возникают споры вокруг ограниченной правосубъектности со страховыми пулами.

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

Показатель перестройки к 2035 году	Измеренная база 2026 года	Ветвь A («Нормальная технология»)	Ветвь B («Продолжение тренда» — базовая)	Ветвь C («Ускорение»)	Репер и периодичность сверки
Разрыв занятости молодёжи 22–25 лет в подверженных профессиях	−19 % (Stanford DEL, ред. 12.08.2026, данные ADP по 06.2026)	0 % (восстановление через новые роли)	−20 % для тех, кто не подписывает правил; вход через владение агентами	−40 % (массовое сжатие входа, базовые квоты вычислений)	Stanford Digital Economy Lab, ежегодно
Доля рабочих часов, исполняемых агентами	< 5 % (оценка авторов на базе сценариев McKinsey 2023–2024)	15 %	30–40 %	> 60 %	McKinsey MGI, Anthropic Economic Index, ежегодно
Доля повседневных решений, принимаемых автономно	0 % в 2024 — 15 % к 2028 (Gartner)	10 %	30 %	> 50 %	Отчёты Gartner, ежегодно
Основная форма дохода человека	Заработная плата за часы исполнения (S1)	Заработная плата	Зарплата + доход от подписанных правил и данных + премия за риск	Дивиденд с вычислений, выплаты суверенных ИИ-фондов	Anthropic Economic Futures, законы о фондах, ежегодно
Субъект ответственности за действие агента	Оператор и директор по общим нормам ГК (ст. 53.1)	Оператор и владелец предприятия	Подписавший правило (S5); квитанция как стандарт полиса и защиты	Спорные режимы «ограниченной правосубъектности» и спецфондов	Обзор законодательства и условий полисов AIUC-1, ежегодно

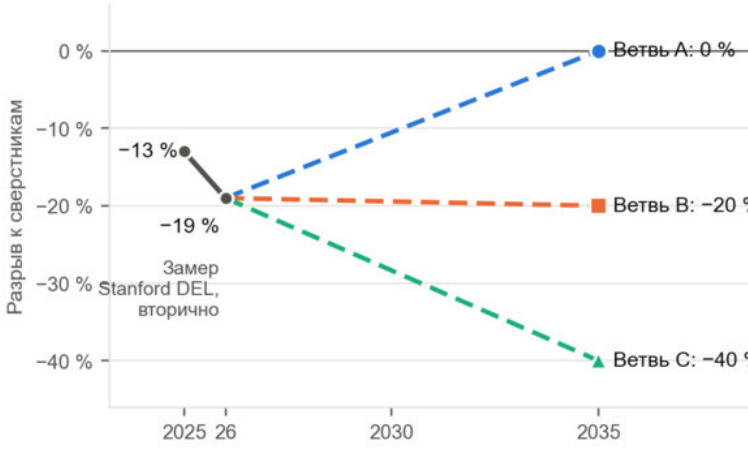


Рисунок 4.2 — Разрыв занятости молодёжи 22—25 лет в наиболее подверженных ИИ профессиях: замер 2025—2026 годов (Stanford DEL, по вторичному источнику) и три ветви к 2035 году (строка Г9)

4.5.2 Инвариантность решения и портфолио квитанций 2035 года

Отделение уровня политики (S5) и детерминированного допуска (S3*) от уровня исполнения (S1) сохраняет силу при любом сценарии. В ветви A шлюз страхует компанию от перерасхода токенов и потери баз данных без раздувания штата безопасников. В ветви B связка контракта и квитанции оформляет новую профессию проектировщика ма-*

шинных инвариантов и даёт директору доказательный документ для страховщика. В ветви C шлюз с ключами у человека служит главным физическим предохранителем против постепенного обессиливания: сколько бы ни выросла мощность моделей, перевод средств или изменение устава невозможны без вердикта UNSAT по правилу человека.

На этом основании складывается прикладной замысел к горизонту 2035 года: цепь квитанций становится личным профессиональным портфолио человека вместо бумажного резюме. Традиционное резюме перечисляло освоенные программы и требовало верить написанному на слово. В экономике агентов навык ручного исполнения обесценен, а словесные заверения не стоят ничего. Вместо резюме специалист предъявляет будущему нанимателю, совету директоров или страховому андеррайтеру свой криптографический журнал с календарными якорями времени. Запись доказывает на проверяемой статистике, сколько агентов работало под ключом автора, сколько миллионов транзакций пропустили подписанные им контракты, сколько нарушений остановили и сколько имущественных инцидентов допустили. Автономный проверщик позволяет любому новому заказчику за две минуты перепроверить этот стаж без обращения к прежнему работодателю.

4.6 Экономика доказанного действия и периметр монетизации

Перевод вопроса о роли человека на язык корпоративно-

го бюджета объясняет, почему маржа уходит из исполнения действий в их детерминированную проверку. В 2025 году инвестиции в агентные компании достигли $24,2 млрд в 1 311 сделках по методологии PitchBook (или $6,7 млрд в узких границах прогноза Prosus и Dealroom), сосредоточившись на уровнях исполнения и транспорта. Однако открытые весовые модели отстают от закрытых в среднем на четыре месяца (Epoch AI, 29 мая 2026 года; International AI Safety Report, февраль 2026 года). По данным Vercel AI Gateway Production Index, доля открытых моделей в потоке выросла с 29% токенов в июне 2026 года до 56% в августе 2026 года при доле менее 4% в денежных расходах. Маржа простого исполнения сжимается к цене электроэнергии. Одновременно Gartner прогнозирует отмену свыше 40% проектов агентного искусственного интеллекта к концу 2027 года из-за бесконтрольных трат, отводя защитным системам-стражам 10—15% рынка к 2030 году.

Внедрение процессорного шлюза заменяет непредсказуемые траты токенов фиксированной проверкой за миллисекунду на CPU без видеокарт. При этом тесты шлюза обходятся в ноль долларов на внешние модели, а совокупный счёт стенда за 1—14 октября 2026 года составил $0,30 ($0,14 за 1—6 октября). Контракт данных задаёт твёрдый потолок на частоту вызовов и лимиты операций.

На осень 2026 года коммерческая выручка шлюза MITR и испытателя INDRA равна 0 рублей, поэтому оценки бу-

дущей выручки сохраняют статус гипотезы. Ценовая основа продукта строится на имущественной дельте:

Ценность шлюза = (Предотвращённый прямой ущерб
+ Снижение страховой премии и риска ст. 53.1 ГК РФ
+ Экономия внимания человека S5) − Стоимость проверки на CPU

Когда стоимость процессорной проверки стремится к нулю, созданная экономия от предотвращённых сбоев, снижения тарифа страхования и защиты руководителя по статье 53.1 Гражданского кодекса распределяется между предприятием, страховщиком и поставщиком шлюза. Подробный разбор коммерческой лестницы вынесен в главу 16. Там рассмотрены как подписка на этапе подключения внешнего арендатора, так и страховые полисы с покрытием до $25 млн и обязательным журналом исполнения на примере Armilla AI и стандарта AIUC-1.

4.7 Протокол независимой верификации

Любое утверждение этой главы о контроле над роем агентов, удержании ключей и неизменяемости авторской цепи квитанций перепроверяется без опоры на устные заверения разработчика. Авторство контрактов, достоверность зондов и защита операционной системы при этом остаются в доверенной базе. Статусы блокировки по шести базовым классам, отсутствие обходов перехватчика, задержки на CPU и автономный повтор решений проверяются командами из Приложения Б и инструментом verify_chain_standalone.py — replay. Состояние публичного свидетеля эпох открыто на https://mitr.asi5.ru/, где на 18 ноября 2026 года зафиксированы 5 409 квитанций и 59 эпох. На историческом срезе 24 октября 2026 года повтор подтвердил все 5 053 решения в запечатанной эпохе при 5 079 записях в журнале.

Внешние ориентиры траектории труда и ответственности сверяются ежегодно по четырём открытым источникам. Динамика молодёжной занятости 22—25 лет отслеживается по докладам Stanford Digital Economy Lab, горизонты автономной работы и прогноз автоматического программиста — по панелям METR Time Horizons и AI Futures Project. Требования журналов исполнения проверяются по условиям андеррайтеров и реестру AIUC-1. Правовой статус агентов отсле-

живается по изменениям национального законодательства, включая Гражданский кодекс РФ и китайский акт о трёх уровнях решений от 8 мая 2026 года. Полные реестры утверждений, прогнозов и сводная библиография размещены в Инженерных реестрах и паспортах глав, Приложении А и Приложении Д.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, купив полную легальную версию на Литрес.

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