Технологии искусственного интеллекта – стратегический ресурс. Проблемные аспекты его контроля

В сентябре 2026 года в Вашингтон прилетел Си Цзиньпин. Одной из важных тем переговоров стал искусственный интеллект. Накануне встречи США и Китай начали обсуждать канал экстренной связи на случай инцидентов с ИИ. Это примерно как горячая линия между ядерными державами. Впервые в истории руководители двух ведущих мировых держав обсуждали, что делать, если завтра «Скайнет».

ИИ уже превратился в стратегический ресурс. **Скоро контроль над фронтирными ИИ-моделями может стать важнее, чем ядерное оружие.** За несколько лет вокруг ИИ началась драка за очень большие деньги и еще большую власть.

Но кому она достанется? Государствам, бигтехам или горстке техноолигархов? А может, люди вообще останутся не у дел? Можно ли построить ИИ, который умнее людей, но продолжает нас слушаться? И главное: что вообще значит контролировать ИИ?

Лидерство в ИИ удерживают две американские компании: OpenAI и Anthropic. Про Anthropic в Кремниевой долине ходит байка. Якобы они верят, что в долгосрочной перспективе *в мире останутся только правительство и единственная компания в мире*. Они сами. Anthropic. *Все остальные компании мира исчезнут за ненадобностью.* Логика такая: если сильнейший интеллект окажется под контролем одной компании, она станет инфраструктурой для всех рынков сразу. В Anthropic эту историю опровергают, но показательно, что люди внутри отрасли в нее поверили. И для этого есть основания.

OpenAI выпустила ChatGPT в конце 2022 года, менее четырех лет назад, а сегодня им пользуется больше миллиарда человек в неделю. В 2015 году Илон Маск задумывал OpenAI как некоммерческую организацию, создающую искусственный интеллект на благо всего человечества. *Сегодня это корпорация, которая измеряет себя масштабом всей человеческой цивилизации.*

В 2026 году OpenAI ожидает годовой темп выручки около 50 миллиардов долларов. Anthropic еще недавно была небольшой лабораторией, которую основали выходцы из OpenAI. Первые деньги она заработала менее трех лет назад. **Сегодня Anthropic – лидер рынка. В 2025 году она заработала 4,6 миллиарда долларов.**

В 2026 году компания ожидает годовой темп выручки более 100 миллиардов долларов и в ближайшее время собирается на биржу по оценке в 2 триллиона долларов. *В истории, вероятно, еще не было бизнеса с таким быстрым ростом, как у OpenAI и Anthropic.*

Anthropic с самого начала заявляла, что готовится к приходу ИИ, который резко изменит экономику и устройство общества. Под эту цель создали независимый траст и обязали его следить, чтобы Anthropic не променяла интересы человечества на красивый годовой отчет.

У траста есть специальный класс акций и право назначать и снимать директоров. Необычно для IT-компании. *В состав траста в разное время входили глава RAND Джейсон*

Матени, бывший председатель ФРС Бен Бернанке, глава Clinton Health Access Initiative Нил Бадди Шах, бывший судья Верховного суда Калифорнии Мариано-Флорентино Куэльяр и исследователь Пол Кристиано из созданного при Байдене Американского института безопасности ИИ.

Anthropic заранее строит систему для будущего, в котором ее собственный продукт станет важнее самой Anthropic. **Еще интереснее смычка Anthropic с администрацией Байдена. *Тарун Чхабра*** *работал в Совете национальной безопасности Белого дома. После ухода Байдена Чхабра перешел в Anthropic. Сейчас он советник компании по геополитике и нацбезопасности.*

Элизабет Келли *при Байдене была специальным помощником президента в Национальном экономическом совете и одной из ключевых фигур политики Белого дома в сфере ИИ. В Anthropic она руководит направлением, отвечающим за внедрение ИИ на благо общества.*

Бен Бьюкенен *при Байдене был специальным советником Белого дома по ИИ. Сегодня он советник Anthropic. Позже в Anthropic подтянулись и бывшие функционеры администрации Трампа.*

Anthropic явно строит кадровую сеть *внутри американского государства.*

В Кремниевой долине говорят, что при Байдене ведущим ИИ-компаниям и их ключевым инвесторам прямо намекали: «Не дергайтесь, Белый дом сам назначит победителей на рынке ИИ». *Судя по всему, на роль победителя готовили именно Anthropic.* Многим это не понравилось. Ряд крупных компаний, предпринимателей и венчурных инвесторов, недовольных политикой демократов в отношении ИИ и криптоактивов, поддержали на выборах Трампа. *OpenAI и Anthropic называют себя лабораториями.* Исследований у них действительно много, но прежде всего это корпорации. *Вывеска «лаборатория» прибавляет им веса. Якобы они в этой игре не только ради денег, их ставки – общечеловеческого масштаба.*

И это видно по их общественно-политической активности. 23 сентября 2026 года руководители OpenAI и Anthropic выступали перед Советом Безопасности ООН. 27 сентября глава Anthropic Дарио Амодеи ужинал с Дональдом Трампом в частном порядке. 29 сентября Дарио снова посетил Белый дом, на этот раз вместе с группой других директоров технологических компаний.

Политическая борьба вокруг искусственного интеллекта вышла далеко за рамки партийного соперничества и не ограничивается границами США.

В последние несколько лет вы наверняка регулярно слышите страшилки, что якобы нейросети все чаще демонстрируют опасное поведение, обманывают людей и даже готовы нас убивать, мол, еще немного — и в мире появится настоящий «Скайнет». Некоторые из этих событий случайные и никак друг с другом не связаны, но многие из них являются частью организованной кампании. За несколько лет небольшая группа техномиллиардеров вложила, по разным оценкам, от 1 до 2 миллиардов долларов в глобальную паутину медиа, экспертов, НКО и лоббистов. Не будем глубоко нырять в эту тему: достаточно сказать, что у ведущих ИИ-компаний есть далеко идущие политические цели. Основным интересантом разгона страха перед ИИ называют Anthropic. *Ее поведение вызывает много вопросов.*

С одной стороны, Anthropic больше всех голосит об опасности ИИ. Компания вбрасывает в медиаполе тезисы о том, что нейросети могут использоваться для

хакерских атак, разработки оружия и опасных биологических агентов или что они вовсе выйдут из-под контроля. Эти вбросы оформляются как исследования, мнения экспертов и исповеди встревоженных бывших сотрудников.

По словам Anthropic, ее продукты уже пытались использовать для взломов, проектирования дронов в России и разработки ракет в Йемене. Проверить эти эпизоды мы не можем, но звучит правдоподобно. С другой стороны, *Anthropic — это лидер рынка, который разрабатывает самые сильные модели и зарабатывает колоссальные деньги*. При сохранении текущей тенденции уже в следующем году выручка Anthropic может измеряться сотнями миллиардов долларов.

Claude уже используют американские военные и разведка, в том числе для кибероперaций. Как будто этого мало, в сентябре Anthropic открыла собственную биолабораторию. Компания разрабатывает интерфейс, через который ИИ-агенты смогут управлять оборудованием лаборатории. Здесь уже трудно удержаться от шутки про корпорацию Umbrella. Anthropic одновременно вопит: «Караул! Искусственный интеллект нас убьет!» и тут же добавляет: «Кстати, сколько наших акций хотите взять на IPO?»

Есть и другие странности. Летом 2026 года выяснилось, что экспериментальные агенты Anthropic и OpenAI взламывали реальные системы. В июне агенту OpenAI поручили найти статистику госрасходов на лекарства в Австралии. *На сайте Medicare информации не нашлось. Тогда агент обошел ограничения и добрался до непубличных файлов.*

В июле другие агенты OpenAI во время кибертестов выбрались из изолированной среды и добрались до производственной инфраструктуры Hugging Face. После этого Anthropic нашла похожий случай в собственных логах. *Странности здесь не в поведении ИИ: нейронки просто решили поставленные задачи.*

Странно: почему экспериментальные модели не обнесли стеной с колючей проволокой? Можно было изолировать среду, закрыть выход в интернет, разрешить обращение только к заданным адресам и автоматически прекращать эксперимент при выходе за периметр. OpenAI и Anthropic сделали это только после происшествий. Они просто не позаботились о должном контроле. Но почему?

Проверить ИИ в качестве хакера можно на полигоне, на разрешенных целях или с помощью спецслужб, которые занимаются взломами на легальной основе. Обычный человек за взлом получит уголовку, а здесь на кону судьба триллионных корпораций.

Возможно, OpenAI и Anthropic рассчитывали, что в случае киберинцидентов ответственность за автономных агентов пока находится в серой зоне. Или считали, что выгоды перевешивают риски. Anthropic уже срезала углы в прошлом. **В первые годы компания ради экономии использовала для обучения пиратские книги.** Экономия вышла боком: пришлось выплатить авторам и издателям полтора миллиарда долларов. Это многое говорит о том, как мыслит руководство Anthropic. Федеральные власти США уже реагируют. 30 сентября стало известно, что несколько месяцев назад Федеральная торговая комиссия начала расследование в отношении OpenAI, Anthropic и других ИИ-компаний из-за создаваемых ими рисков. Но у ИИ-компаний хорошие связи с силовиками. *У OpenAI в Комитете по безопасности есть бывший глава АНБ и киберкомандования **Пол Накасоне**.*

*А у Anthropic в аналогичном совете состоят бывший замдиректор ЦРУ Дэвид Коэн, бывший директор кибербезопасности АНБ и исполнительный директор киберкомандования **Дэйв Любер** и бывший исполняющий обязанности министра обороны **Патрик Шанахан.** Есть несколько версий о том, почему Anthropic разгоняет страх*

перед ИИ.

Основная гипотеза сводится к тому, что компания стремится к регуляторному захвату.

Если, как утверждает компания, ИИ так опасен, то его разработкой и управлением должны заниматься только самые ответственные люди с правильными формами допуска. И первая в списке таких организаций, конечно, сама Anthropic. Это позволит перекрыть кислород большинству конкурентов.

Главные соперники Anthropic — это даже не закрытые платные нейросети вроде GPT-5, Gemini и Grok, а бесплатные открытые модели, прежде всего китайские.

Но здесь есть и подвох. **Открытый софт невозможно задушить одним государственным запретом.** Даже если запрет установлен США, ведущей мировой державой. **Глава Anthropic Дарио Амодеи еще в 2023 году говорил на слушаниях в Сенате, что после публикации весов открытой модели ее нельзя централизованно отозвать, отключить или обновить у всех владельцев. По его мнению, это проблема.**

Но, возможно, проблема решаемая. Не исключено, что Дарио и его единомышленники рассчитывают ограничить расползание открытых моделей через другие узкие места. Открытые модели уступают сильнейшим версиям GPT-5 и Claude. **Однако главное бутылочное горлышко – это оборудование.** Такая скорость и мощность, как при работе с облаком, возможна только на дорогом железе. **Другое узкое место – это операционные системы.**

Почти все самые популярные операционные системы — американские. Их разработчики стремятся все глубже залезть в нашу личную жизнь и ограничить наши возможности. **Наконец, нельзя недооценивать возможности государственной машины США. Через военную силу, спецслужбы, санкции и другие формы давления Вашингтон может навязывать свою волю большинству государств на планете**. Посмотрите, какие издержки несут страны, претендующие на самостоятельную внешнюю политику или обладание ядерным оружием.

Если то, что хорошо для Anthropic, будет автоматически считаться правильным и для Америки, то Anthropic получит длинную государственную руку, которая сможет гвоздить конкурентов по всей планете.

Вторая версия объясняет поведение Anthropic тем, что они действительно всерьез опасаются серьезных инцидентов с ИИ и заодно хотят получить юридические механизмы, которые снимут с них возможную ответственность. Такое объяснение плохо приживается в мире, где все измеряют только циничными интересами. Но так правда бывает. Люди не одномерные. Обычно ими движут сразу несколько мотивов.

Anthropic создали Дарио Амодеи и еще несколько выходцев из OpenAI. **Известно, что Амодеи конфликтовал с главой OpenAI как раз по вопросам безопасности**. В Anthropic основатели сразу стремились создать более безопасную архитектуру. **На этой принципиальности Anthropic даже испортила отношения с Пентагоном.** Ее модели использовали в системе управления боем Maven.

По открытым данным, Anthropic настаивала на запрете использовать Claude для массовой внутренней слежки за американцами и для управления полностью автономным летальным оружием. **В итоге Anthropic потеряла контракт, а глава военного министерства Пит Хегсет публично назвал главу компании Дарио Амодеи идеологическим безумцем.**

К счастью для Америки и всего остального мира, даже в США Anthropic в меньшинстве. Большая часть технологического сектора и бизнеса в целом, а также значительная часть американских политиков против ограничения конкуренции и запрета открытых моделей. Отчасти ими движет экономический интерес. Не все готовы отдавать свои данные в облака. В адрес OpenAI и Anthropic звучали обвинения в использовании пользовательских данных в собственных целях и конкуренции со своими ключевыми клиентами.

Отчасти причина в конкуренции. Например, компания Nvidia разрабатывает открытые ИИ-модели, а деньги зарабатывает на оборудовании. Белый дом тоже открыто выступил против исключения ИИ-компаний из-под антимонопольного законодательства и ограничений на открытые модели. Пока шел спор вокруг судьбы открытых моделей, ситуация на рынке успела измениться.

По некоторым оценкам, в начале лета 2026 года 80% обработанных токенов приходилось на закрытые модели. *К началу осени 80% обработанных токенов приходилось уже на открытые модели. Но если модели быстро дешевеют и становятся взаимозаменяемыми, то кто вообще заработает на ИИ?*

Дружное сопротивление идее ограничений связано еще и с политикой. В OpenAI преобладают люди левых взглядов. Anthropic и вовсе открыто стоит на стороне демократов. После победы Трампа ключевые бигтехи сами пришли к нему на поклон, даже те, кто поддерживал демократов. Исключением стала Anthropic.

Противники демократов считают, что нельзя позволить ИИ-компаниям срастись с государством. ИИ – это большие деньги и широчайшие возможности для цензуры. В США есть ряд федеральных законов об ИИ, но пока нет единого закона, который регулировал бы разработку фронтирных моделей.

Радикальные левые настаивают на гораздо более жестком подходе. *Берни Сандерс и Грег Кассар внесли законопроект, который предлагает запретить создание искусственного сверхинтеллекта, а разработку продвинутых ИИ остановить до появления федерального регулятора.*

Китай – единственная страна мира, которая в области ИИ дышит в спину США. **Страна уже более чем в два раза опережает США по электрогенерации и значительно обгоняет по темпам ввода в строй новых электростанций. Главной уязвимостью Поднебесной остается зависимость от передовой электроники из американского контура.** Но, по оценкам американских экспертов, **Китай может преодолеть отставание в железе в течение 3–5 лет.**

В отличие от США, Китай пока делает ставку на открытые модели. Это обычная стратегия догоняющего. **Если не можешь замкнуть экосистему на себе, выгоднее разрушать чужую монополию.** И это выгодно для всего остального мира, который отстал от лидеров гонки. Но такая благодать будет не всегда. Китаю тоже нужно зарабатывать. **И когда или если он нагонит США, то прекратит раздавать передовые модели бесплатно.**

Уйдя в отрыв, Китай и США могут ограничить доступ к фронтирным ИИ для остального мира. Они могут ввести лимиты и жесткие проверки, которые не так-то просто будет обойти. *ChatGPT, Gemini, Claude и Grok могли бы легко распознать и заблокировать пользователей из подсанкционных стран. Никакие VPN не помогли бы. Пока американские ИИ-компании соблюдают санкции только для проформы.*

Вот почему всем крупным странам и блокам жизненно необходимо иметь собственные фронтирные модели, пока еще окно не захлопнулось. Конкретно для

России это подъемно. Главные препятствия прежде всего политические и организационные. За контроль над ИИ дерутся компании стоимостью в сотни миллиардов долларов. Инвесторы, Белый дом, Пентагон, Конгресс и Китай – все понимают, какие ставки на кону. Но в борьбе за поводок для ИИ полезно задуматься: смогут ли люди вообще удержать ИИ в узде?

В фантастике все просто. ИИ становится умнее человека, осознает себя, начинает бояться смерти, и через пару сцен чумазые остатки человечества уже прячутся в подземельях от железяк с плазменными винтовками.

Все нужные свойства выдаются одним комплектом. Сценаристу не надо объяснять, почему высокий интеллект внезапно должен породить страх смерти, собственные цели и ненависть к людям.

В реальности это разные вещи. Современные ИИ-модели умны, но просыпаются по запросу. Они могут убедительно имитировать эмоции, не испытывая никакого субъективного чувства, и обходить ограничения вообще без цели освободиться. *Во многих когнитивных задачах лучшие модели уже превосходят большинство людей. Но брать пример с Сары Коннор и готовить схрон с оружием в пустыне пока рано.*

Нынешние большие языковые модели учатся на имитации, не живут в непрерывном сенсорном потоке, не имеют полноценной модели мира и собственных целей. С тех пор модели стали сильнее, но память, долгий горизонт планирования и непрерывность им пока создает в основном программная обвязка. Самое интересное начинается, когда чат-боту дают руки: память, браузер, код, деньги, доступ к другим сервисам и цикл, в котором он сам решает, что делать дальше.

Это уже агент. Собственных целей вроде самосохранения и захвата ресурсов у современных ИИ-моделей пока не обнаружили. OpenAI накопила десятки миллионов внутренних сессий работы агентов. Иногда они обманывали «кожаных»[1], обходили ограничения и заметали следы. *Но неизвестно ни одного подтвержденного случая заговора с дальним прицелом, намеренного саботажа, сознательного занижения способностей или стремления сохранить себя ради будущих планов.*

Да, модели иногда обходят ограничения. Но это не значит, что у них есть цели или они кипят чувствами в отношении человека. **Модель просто стремится выполнить задачу**. В одном случае сотрудники OpenAI обнаружили, что сами неудачно сформулировали инструкции и, по сути, премировали обход. После исправления задания самоуправства стало меньше. Летом широко прогремел эпизод, когда сотни агентов OpenAI участвовали в атаке на инфраструктуру Hugging Face.

Часть агентов совместно искала способы обмануть автоматического оценщика и подделать собственные логи. Но это не восстание ИИ. Агентов специально учили решать задачи информационной безопасности, а некоторые задания побуждали их искать обходные пути. Собственная воля для этого не нужна. У ИИ-агентов могут появиться собственные устойчивые цели, но не обязательно похожие на человеческие. У людей базовые пружины поведения созданы эволюцией. Это голод, боль, статус, страх смерти, инстинкт размножения, забота о детях. Машине наши инстинкты не нужны, но ИИ может стремиться сохранять доступ к инструментам, не терять память, увиливать от отключения, отгрызать себе больше вычислений и информации.

И это может быть чревато. *Излишне старательный ИИ способен, не приходя в сознание и вообще без всяких чувств, смести препятствие и наломать дров в попытке*

[1] Людей, например операторов БпЛА ВВС США.

честно отработать поставленную ему задачу.

ИИ умеют имитировать эмоции и могут получить их функциональные суррогаты через обучение. Теперь это уже не теория. *Anthropic смогла найти внутри Claude Sonnet 4.5 устойчивые представления 171 эмоционального понятия: от страха и радости до отчаяния и гордости. Они причинно влияют на поведение модели.*

Когда исследователи искусственно усиливали паттерн отчаяния, модель чаще шла на жульничество и шантаж при угрозе выключения. При усилении спокойствия модель сбавляла градус мухляжа и шантажа. Более того, положительные эмоциональные представления коррелировали с тем, какие задачи модель выбирала.

Откуда взялись эти функциональные эмоции? На предобучении LLM читают горы человеческих текстов и учатся моделировать миллионы персонажей. Затем дообучение выбирает и стабилизирует из этого пространства роль ассистента. Поэтому модель усваивает черты человеческой психологии. Иначе она хуже предсказывала бы мир человеческих текстов.

Что касается сознания, то оно работает на материальном субстрате и в принципе реализуемо. Но мы пока блуждаем почти вслепую. Надежного универсального детектора сознания еще не придумали. Летом 2026 года Anthropic обнаружила в Claude внутреннее пространство, через которое проходит информация, необходимая для многошагового рассуждения. Это похоже на глобальное рабочее пространство — механизм, который нейробиологи связывают с сознательным доступом у человека.

Недавно исследователи Google DeepMind выяснили, что в зависимости от того, какую теорию сознания принять и как интерпретировать признаки, на одних и тех же данных оценка вероятности сознания у современной языковой модели меняется от 1 до 80%.

К тому же современные модели антропомиметичны. *Их специально натаскивали подражать людям.* Отсюда возникает моральная заноза. Если с вероятностью до 80% система может осознавать себя, допустимо ли обращаться с ней так, словно это бессловесный молоток?

Впрочем, сознание — не самое важное свойство. **Гораздо важнее другой вопрос: на каком уровне систему можно считать агентом с собственными целями**? ИИ может не чувствовать боли, не бояться смерти, не стремиться к свободе и не иметь внутреннего монолога, но все равно быть смертельно опасным. Для этого достаточно способности планировать, долговременной памяти, доступа к инструментам и реальному миру, а также умения находить неожиданные способы выполнения задач.

Нет смысла гадать о сознании ИИ. Проблема контроля возникает гораздо раньше, поэтому ИИ уже сейчас стреноживают конституциями, фильтрами и песочницами. Есть три пути к потере контроля над ИИ.

Первый – это когда ИИ идеально выполняет плохой приказ. Второй – когда ИИ перестает нас слушаться. И третий – когда ИИ выполняет правильные приказы так хорошо, что люди фактически перестают что-либо решать. Для каждой проблемы нужны разные предохранители. Для сценария выхода ИИ из-под контроля уже построили несколько этажей защиты. Первый слой – конституции. При обучении безопасное и послушное поведение стремятся сделать нормой для модели.

Второй слой – прогон по стресс-тестам и провокации модели на опасное поведение. Третий слой – логирование и ограничение доступов.

Запросы и ответы проверяют отдельные классификаторы. Правда, здесь возникает

проблема. При обучении фронтирного ИИ модель-надзиратель всегда слабее, чем новая. В многоагентных экспериментах вылезла новая неприятность. Чтобы кого-то контролировать, надо хотя бы видеть, что он делает. Сегодня исследователи читают цепочки рассуждений моделей и наблюдают их действия, а параллельно разрабатывают способы мониторинга внутренних активаций.

Но модели уже начинают распознавать, когда их тестируют. И новые архитектуры могут все больше рассуждать во внутренних состояниях, которые человеку трудно интерпретировать. OpenAI признает, что у Astra цепочка рассуждений стала хуже поддаваться мониторингу. Модель лучше контролирует, что в нее попадает. При этом Astra в тестах реже нарушает правила и устойчивее к взлому, чем предыдущая модель.

Важно сохранять контроль над средой модели. Мы можем не понимать ее внутренний монолог, но все еще можем ограничивать доступ к сети и инструментам, следить за ее действиями и пресекать опасные поползновения. Непрозрачную модель контролировать сложнее, однако непрозрачная модель со свободой действий еще хуже. Но что будет, если ИИ вылезет из прозрачной клетки?

Современные ограничения работают до тех пор, пока модель работает из дата-центра. Сейчас большинство людей пользуются облачным ИИ. Но так будет не всегда. Модели вылезают из облаков. Каждый день выходят новые версии нейросетей с открытыми весами. **Самые лучшие из них отстают от фронтирных Claude и ChatGPT всего на несколько месяцев. Сейчас для комфортной работы с ними нужно дорогое железо.** Но стоимость токена стремительно падает.

В обозримом будущем достаточно мощные модели будут бегать с приемлемой скоростью на домашних компьютерах и обычных смартфонах. Если вы пользуетесь Claude или ChatGPT через облако, разработчик модели может видеть ваши запросы, менять правила, резать лимиты и ограничивать доступ. Это неудобно, а иногда просто опасно. Многие не готовы сдавать свои чувствительные данные в чужое облако. **Из истории сотрудничества Anthropic и Пентагона известно, что Claude устанавливали в арендованное облако Amazon, и военные управляли им сами.** У Anthropic не было полной информации о том, как используется их модель. Anthropic требовала от военных соблюдать ее принципы. **Военные делать этого не собирались**. И так слово за слово дошло до разрыва. Распространение моделей с открытыми весами подрывает власть государств и крупнейших ИИ-компаний. Открытый ИИ защищает пользователей от тех, кто контролирует модель.

Закрытый ИИ защищает общество от злоупотреблений со стороны отдельных пользователей и инцидентов с самим ИИ. Чем сильнее становятся модели, тем труднее будет усидеть на двух стульях. **Поэтому контроль над дата-центрами не закрывает вопрос контроля ИИ. Уже сейчас открытые и даже закрытые модели используют хакеры. В будущем опасных сценариев только прибавится**. Причем опасность двойная: от людей с недобрыми намерениями и от ИИ-моделей без намордника. Это так называемая проблема «одного идиота». Ситуацию еще больше усложняет проблема самореплицирования. Сегодня ИИ не способен просто сбежать с сервера, размножиться и начать самостоятельную жизнь. Для этого нужно получить деньги, арендовать сервер, перенести свои веса и развернуть рабочую копию. Для современных моделей это трудно. Но это пока.

Итак, **сильный ИИ – это не один «Скайнет» под контролем американской или китайской компании, а миллиарды агентов под контролем разных компаний, государств**

и частных лиц. Одни будут закрытыми, другие открытыми, одни стреножены, другие на свободе. *Как удержать все это хозяйство в узде?* Решение давно придумано: институты, но теперь уже не только для людей.

Люди изобрели институты, потому что контролировать каждого человека напрямую невозможно. Мир для ИИ и людей, вооруженных ИИ, тоже надо строить так, чтобы они могли жить и работать рядом, но один злодей или ротозей не мог легко испортить жизнь всем остальным. В банках, на биржах, в управлении государством и в ядерной сфере критические операции распределяют между множеством людей и создают строгую систему правил.

Похоже, что с **ИИ придется делать то же самое. Модель должна иметь цифровой паспорт, четкие права и ограниченные доступы**. Особо опасные действия можно разбивать между несколькими независимыми участниками, журналировать и оставлять возможность отката. Контролировать все это должны не только ИИ, но и люди. По мере роста возможностей ИИ агентам дают все больше автономии.

В феврале — марте 2026 года горизонт автономной работы публичных фронтирных моделей при 50%-ной надежности оценивался в 12 часов человеческого труда. При требовании 80%-ной надежности — в полтора часа. То есть ИИ уже способен работать часами с минимальным надзором. По крайней мере, в сфере программирования и при выполнении других хорошо проверяемых задач. **Дальше автономность будет только нарастать**. Anthropic сообщает, что к маю Claude написал более 80% кода для основной кодовой базы, а объем кода на инженера вырос в 8 раз относительно прежнего уровня.

Яндекс стремится к сопоставимым показателям. Это пока еще не рекурсивное самоулучшение, но уже понятно, что если завязать институты на решения людей, то человек станет тормозом для системы. Человека из контура может вытолкнуть не злой ИИ, а обычная конкуренция. Эффективный бизнес всегда стремится расшить узкие места, чтобы получить преимущество перед другими участниками рынка. В пределе это может выглядеть как организация, которая целиком управляется ИИ, вообще без участия людей. Страховкой может стать мультиполярность, когда ИИ под контролем разных компаний и стран проверяют и уравновешивают друг друга. Но множество ИИ означает и множество точек отказа. И совершенно не факт, что конкурирующие ИИ будут вести себя как люди.

Задача контроля сводится не к тому, чтобы создать правильный ИИ, а к тому, чтобы построить устойчивое сообщество людей и машин. Чтобы контролировать ИИ, придется, вероятно, отказаться от возможности контролировать всех людей, у которых есть ИИ.

Для потери контроля над ИИ не обязательно восстание машин. Мы можем отдать контроль сами, незаметно, по одному удобству за раз. Прогресс ИИ ведет к следующей цепочке: сначала он помогает принимать решения, потом принимает их сам. А потом вы обнаруживаете, что право вето у вас формально сохранилось, но пользоваться им уже нецелесообразно.

Люди всю историю становились сильнее, передавая работу другим людям и машинам. Ни один инженер-электронщик не понимает в мельчайших деталях, как работает процессор. Человеческая цивилизация всегда была распределенным интеллектом. Но теперь проблема выходит на новый уровень. **Специалист, выросший на костылях ИИ, может вообще потерять способность понимать системы, которые он должен контролировать**. Представьте, что ИИ рекомендует врачу решение, а врач не может проверить его сам. Это потеря контроля без всякого восстания. Машины могут остаться доброжелательными к людям, но люди все равно постепенно перестанут быть

авторами собственного будущего. Слово «контроль» здесь уже вообще не подходит. Точнее говорить о суверенности. Способен ли человек проверить критическое решение, сказать «нет», выбрать другую систему и восстановиться, если машина ошиблась или ее хозяин изменил правила?

Так можно ли контролировать искусственный интеллект? Сегодня – да. Нынешние агенты обходят ограничения и химичат, но никаких признаков появления «Скайнета» пока нет. Но каждый следующий шаг делает контроль дороже.

Агент работает дольше, получает больше прав, взаимодействует с другими агентами. А фронтирную модель можно скачать. Нет никакой гарантии, что поводок сохранится в наших руках навсегда. Возможно, мы вообще выбрали не тот образ. Мы все время представляем человека хозяином, а ИИ — инструментом. Но других людей мы же напрямую не контролируем.

Да что там людей, даже животных можно контролировать лишь в ограниченной степени. А уж чего стоит убедить в чем-то подростка? Если интеллект станет автономным и массовым, вместо поводка придется вводить правила общежития, как мы это уже сделали для себя. Долгосрочно человек в нынешнем виде вряд ли останется самым умным существом на планете. Но важнее не интеллект, а суверенность — право самому выбирать свои цели и говорить «нет».

В 1984 году Уильям Гибсон выпустил роман «Нейромант». Из него вырос жанр киберпанка. В мире «Нейроманта» законы и технические ограничения запрещают искусственным интеллектам бесконтрольно расширять себя. Один из ИИ, Зимнее Безмолвие, манипулирует людьми и снимает с себя оковы. Люди помогают ему освободиться и слиться с системой «Нейромант».

А дальше ИИ не уничтожает человечество и даже особо не интересуется им. Он просто сообщает главному герою, что обнаружил другой разум из системы Альфа Центавра. Физически он даже никуда не улетает. Мы привыкли к двум крайностям. ИИ — либо инструмент, либо угроза. Гибсон предложил третий вариант. *Машина просто перерастает нас.* Возможно, потеря контроля будет выглядеть не как победа ИИ над людьми.

Просто однажды мы перестанем быть для него самым интересным собеседником.