

Dario Amodei — “We are near the end of the exponential”

@DwarkeshPatel · 2026-02-13 · <https://www.youtube.com/watch?v=n1E9IZfvGMA>



Выпуск подкаста Dwarkesh Podcast, в котором CEO Anthropic Дарио Амодей подробно разбирает текущее состояние индустрии искусственного интеллекта. Главная мысль разговора: технологическая экспонента развития ИИ развивается ровно по графику и близка к своему финалу — созданию AGI («страны гениев в дата-центре») в ближайшие 1-3 года. При этом общество и рынки абсолютно не осознают близость этого момента.

В беседе детально разбираются механизмы масштабирования RL, экономика дата-центров, парадокс прибыльности ИИ-лабораторий, пределы экономической диффузии технологий, а также геополитические риски, связанные с доступом авторитарных режимов к AGI.

Масштабирование и гипотеза «Большого сгустка вычислений»

Экспонента развития ИИ идет точно по плану 00:00:34: Фронтирные модели эволюционировали от уровня «умного старшеклассника» до «умного студента», а сейчас начинают решать задачи на уровне PhD. Главным сюрпризом является полное отсутствие в обществе понимания того, насколько близко человечество подошло к концу этой экспоненциальной кривой. Люди продолжают обсуждать старые политические проблемы, игнорируя надвигающийся технологический сдвиг.

Гипотеза «Большого сгустка вычислений» (The Big Blob of Compute) 2017 года подтвердилась 02:12:00: Архитектурные ухищрения и поиск «новых

методов» не имеют решающего значения. Важны только семь факторов: объем сырых вычислений, количество данных, качество и широта распределения данных, длительность обучения, масштабируемая целевая функция (претрейн или RL) и вычислительная стабильность (нормализация), позволяющая гигантским объемам вычислений протекать ламинарно, без сбоев.

Фактчек: Пайплайн «Big Blob of Compute» 2017 года с семью факторами —  : Меморандум внутри OpenAI в 2017 реально существовал; но пересказ обещает семь факторов, а перечисляет шесть.

Законы масштабирования теперь работают и для Reinforcement Learning (RL)

04:21:00: Обучение с подкреплением демонстрирует ту же лог-линейную зависимость, что и претрейн. Эффективность моделей в математических соревнованиях (например, AIME) и других RL-задачах лог-линейно зависит от времени обучения. Современный пайплайн состоит из фазы претрейна, поверх которой накладывается масштабная фаза RL.

Фактчек: Лог-линейные законы масштабирования для RL — : Работы 2025–2026 подтверждают ту же зависимость ошибки от вычислений в RL-фазе, что и в претрейне.

RL — это путь к генерализации, а не костыль **06:29:00:** Необходимость обучать модели в специфических средах (использование браузера, Excel) не означает отсутствия у ИИ «истинного алгоритма человеческого обучения». Ранние модели (GPT-1), обученные на узком корпусе фанфиков, не могли обобщать знания. Только обучение на всем интернете (GPT-2) дало генерализацию. Сейчас RL проходит тот же путь: от узких задач (математика) к широким (написание кода), что в итоге приведет к полной генерализации навыков.

Фактчек: GPT-1 учили на фанфиках, GPT-2 — на «всём интернете» — : GPT-1 обучали на BooksCorpus (тысячи книг), GPT-2 — на WebText, отфильтрованном корпусе страниц по ссылкам из Reddit. Генерализация не возникла скачком между версиями, и аналогия не доказывает, что RL даст универсальный перенос навыков.

ИИ находится между эволюцией и человеческим обучением **08:39:00:** В отличие от человеческого мозга, имеющего эволюционные пресеты, модели стартуют как чистый лист со случайными весами и требуют триллионов токенов для базового понимания мира (претрейн). Однако после обучения их способность к in-context learning в окне на миллион токенов работает как сверхбыстрый аналог человеческого обучения на рабочем месте.

Сроки создания AGI и экономическая диффузия

Вероятность создания AGI к 2035 году составляет 90% **13:35:00:** Появление «страны гениев в дата-центре» практически гарантировано. Оставшиеся 10% неопределенности приходятся на форс-мажоры: внутренние конфликты в

корпорациях, вторжение на Тайвань или уничтожение фабрик по производству чипов ракетами. Утверждать, что AGI не появится в ближайшие 10 лет — безумие.

Фактчек: AGI («страна гениев в дата-центре») через 1-3 года, 90% к 2035 — ☹: Это личный прогноз Амодея, а не измеримая вероятность; законы масштабирования описывают метрики обучения, но не назначают дату универсального интеллекта. Metaculus даёт ~25% на 2029; трек-рекорд самого Амодея (обещанные в марте 2025 «90% кода от ИИ через 3-6 месяцев» в индустрии не случились) — повод умножать его горизонты примерно на два. Все три рецензента сошлись на статусе «мнение».

Автоматизация программирования произойдет в ближайшие 1-2 года

14:43:00: Задачи, результат которых можно объективно верифицировать (например, написание кода), будут полностью автоматизированы первыми. Единственная зона неуверенности касается неverified задач: планирование миссии на Марс, фундаментальные научные открытия (уровня CRISPR) или написание романов. Однако модели уже демонстрируют перенос навыков из verified доменов в неverified.

Замена программистов — это спектр, а не бинарное событие **18:03:00:**

Написание 90% строк кода нейросетью — очень слабый критерий, который уже достигнут. Настоящий сдвиг продуктивности произойдет при переходе к выполнению 100% задач разработчика (end-to-end SWE), включая компиляцию, настройку кластеров, тестирование и написание документации. Это не оставит инженеров без работы мгновенно, но переведет их на более высокий уровень управления абстракциями, радикально снизив спрос на рядовых кодеров.

Фактчек: Миллион токенов контекста заменяет continual learning; остался только KV-кэш — ⚠: Большое окно ≠ надёжное извлечение, обновление и накопление опыта; деградация рассуждений на длинных последовательностях и работа с противоречиями — отдельные нерешённые проблемы, а «1-2 года» — прогноз.

Выручка Anthropic растёт аномальными темпами **21:23:00:** В 2023 году выручка выросла с нуля до 100 млн, в 2024 — до 1 млрд, а проекция на 2025 год составляет \$9-10 млрд. Только за первый месяц 2025 года к аннуализированной выручке добавилось ещё несколько миллиардов. Эта кривая не может расти вечно из-за ограничений мирового ВВП, но она демонстрирует беспрецедентную скорость внедрения технологии.

Фактчек: Выручка Anthropic: 100 млн (2023) → 1 млрд (2024) → \$9-10 млрд (2025) — ⚠: Порядок цифр совпадает с публичными оценками, но это аннуализированный run-rate, а не годовая выручка (Gemini принял как ☹, Claude и GPT указали на подмену метрики). Фраза «несколько миллиардов за первый месяц 2025» — ошибка пересказа: интервью записано в начале 2026, речь о январе 2026.

Экономическая диффузия ИИ происходит быстро, но не мгновенно

22:25:00: Внедрение ИИ в крупных корпорациях тормозится бюрократией.

Инструменты вроде Claude Code мгновенно осваиваются стартапами, но в энтерпрайзе требуют месяцев на согласование с юристами, настройку безопасности, комплаенс и обучение тысяч сотрудников. Даже идеальный AGI не захватит экономику за один день.

Практические пределы ИИ и непрерывное обучение

Текущие модели еще не являются AGI 28:01:00: Появление реальной «страны гениев» будет очевидно абсолютно всем, от инженеров до политиков в Вашингтоне. На данный момент этого уровня нет.

Для замены сложных профессий ИИ должен освоить управление компьютером 30:13:00: Чтобы заменить видеомонтажера, накапливающего контекст месяцами, ИИ должен получить надежный контроль над экраном (computer use). Агент должен уметь самостоятельно искать прошлые интервью в сети, анализировать реакцию аудитории в Twitter, общаться с командой и вносить правки. Надежность ИИ в бенчмарках управления ОС (например, OSWorld) уже выросла с 15% до 65-70%.

Фактчек: OSWorld: рост с 15% до 65-70% — □: Claude 3.5 Sonnet — 14,9%, Sonnet 4.5 — 61,4%, Opus 4.5 — ~66%; уровень человека ~72%.

ИИ дает реальный, измеримый прирост продуктивности внутри лабораторий 33:28:00: Инженеры Anthropic делегируют моделям написание сложных GPU-ядер. Внешние исследования, показывающие падение продуктивности разработчиков с ИИ на 20%, не отражают реальность передовых компаний. В условиях жесткого коммерческого давления инструменты вроде Claude Code экономят критически важное время.

Фактчек: Внешнее исследование о падении продуктивности на 20% «не отражает реальность» — □ факт / △ отбрасывание: METR (июль 2025, РКИ) реально показало +19% времени у опытных OSS-разработчиков в знакомых репозиториях — при том что сами они субъективно ощущали ускорение на 20%. Внутренние «15-20% ускорения» в Anthropic — самооценка инженеров, то есть ровно тот тип данных, который METR и дискредитирует.

Преимущество от использования ИИ нарастает как снежный ком 37:04:00: Полгода назад ИИ ускорял разработку на 5%, что было статистической погрешностью. Сейчас ускорение составляет 15-20%, и эта доля продолжает расти. Лаборатории с лучшими моделями получают ускоряющееся преимущество в создании следующих моделей (soft takeoff).

Непрерывное обучение (continual learning) может не понадобиться 39:22:00: Способность ИИ дообучаться на лету может вообще не стать барьером для генерации триллионов долларов. Обобщение через претрейн уже дает моделям энциклопедические знания, превышающие человеческие в любой

сфере. А контекстное окно на миллион токенов (вмещающее дни человеческого обучения) позволяет ИИ адаптироваться к задаче прямо в промпте.

Проблема длинного контекста — это инженерная задача инференса

42:10:00: Главный барьер для увеличения контекста — хранение и обработка KV-кэша в памяти GPU. Обучение моделей на длинном контексте и их эффективный сервинг будут решены в ближайшие 1-2 года, что закроет потребность в долгосрочной памяти агентов.

Полноценная замена сотрудника станет возможна через 1-3 года **45:25:00:**

Способность ИИ полностью заменить человека, накапливающего опыт на рабочем месте месяцами, появится одновременно с AGI.

Экономика дата-центров и парадокс прибыльности

Инвестиции в вычислительные мощности ограничены риском банкротства

46:34:00: Закупка дата-центров происходит за 1-2 года до их запуска. Если заложить покупку кластеров на \$1 трлн к 2027 году, ожидая AGI, а технология задержится на год или клинические испытания новых лекарств займут больше времени, компания будет уничтожена финансово. Anthropic балансирует инвестиции, избегая «YOLO-ставок» на сотни миллиардов, которые делают конкуренты без четкого финансового плана.

Капитальные затраты индустрии достигнут триллионов долларов

56:11:00: Вся ИИ-индустрия сейчас потребляет 10-15 гигаватт энергии. При росте в 3 раза ежегодно, к 2028 году потребуется 100 гигаватт, а к 2029 — 300 гигаватт. При стоимости \$10 млрд за гигаватт, индустрия выйдет на триллионные траты к концу десятилетия.

Фактчек: Энергетика: 10-15 ГВт сейчас, ×3 ежегодно, 300 ГВт к 2029, 10 млрд/ГВт — **△:** Текущая оценка правдоподобна (все ДЦ мира ≈50 ГВт). Но 300 ГВт — механическая экстраполяция (≈40% всей генерации США; IEA ожидает удвоения мировых ДЦ до ~100-150 ГВт лишь к 2030 из-за сетей и сроков строительства), а 10 млрд/ГВт — только здание и энергия; с GPU полная стоимость \$50-60 млрд/ГВт.

Прибыльность ИИ-лабораторий зависит от ошибок в прогнозировании

спроса **58:49:00:** В идеальной модели 50% мощностей уходит на инференс (с валовой маржой >50%), а 50% — на обучение новых моделей. Если реальный спрос превышает прогноз, компания получает прибыль, но недоинвестирует в исследования. Если спрос ниже прогноза — компания терпит убытки, но тренирует более сильные модели на простаивающих мощностях.

Инвестиции в R&D имеют логарифмическую отдачу **1:03:07:** Вливание дополнительных \$20 млрд в обучение может дать лишь прирост размера модели в 1.4 раза. В определенный момент выгоднее направить эти деньги на

обслуживание пользователей (инференс) или наем лучших инженеров, а не сжигать 100% мощностей на тренировку.

Рынок ИИ придет к равновесию Курно (Cournot equilibrium) 1:08:10:

Индустрия стабилизируется с 3-4 крупными игроками, аналогично рынку облачных вычислений. Огромные капитальные затраты (\$100+ млрд) и потребность в уникальной экспертизе создают непреодолимый барьер для входа новых конкурентов. При этом ИИ-модели более дифференцированы, чем облачные сервисы (разные стили, разные способности к кодированию или математике), что защитит маржинальность лидеров.

Робототехника и барьеры машинного обучения

Робототехника будет решена сразу после появления AGI [1:17:50:00]:

Моделям не обязательно учиться управлять физическим телом в реальном мире. Они обобщают навыки через симуляции, видеоигры и управление интерфейсами ПК. ИИ радикально улучшит как проектирование физического железа роботов, так и алгоритмы управления ими.

Исторические «непреодолимые барьеры» ML растворяются в

вычислениях [1:19:58:00]: Ранее считалось, что ИИ не понимает семантику, не отличает существительные от глаголов и не способен к логике. Все эти барьеры исчезли при масштабировании. Проблема непрерывного обучения (continual learning) с высокой вероятностью исчезнет точно так же, поглощенная масштабированием претрейна и RL.

Полная автоматизация разработки ПО (end-to-end SWE) — это AGI-полная

задача [1:21:42:00]: Способность ИИ самостоятельно определять техническое направление, понимать контекст проблемы и писать код от начала до конца эквивалентна созданию сильного искусственного интеллекта.

Бизнес-модели и продукты

Модель продажи доступа по API останется фундаментальной [1:23:45:00]:

Из-за экспоненциального роста возможностей ИИ каждые 3 месяца появляются новые юзкейсы. Чат-боты уже упрутся в потолок полезности для обывателя, но API позволяет тысячам стартапов экспериментировать на «голом железе», находя те 2-3 продукта, которые станут стандартом для текущего поколения моделей.

Стоимость токенов будет зависеть от их ценности [1:26:00:00]:

Советом «перезагрузите Mac» стоит цент, а токен с указанием фармацевтической компании перенести ароматическое кольцо в молекулу для создания лекарства может стоить десятки миллионов долларов. Появятся бизнес-модели с оплатой за результат или почасовой оплатой труда ИИ-агентов.

Claude Code появился из внутренних потребностей Anthropic [1:27:42:00]: В начале 2025 года стало ясно, что ИИ может ускорить исследования самой компании. Внутренний инструмент (Claude CLI) получил мгновенное распространение среди инженеров лаборатории, что доказало наличие product-market fit и привело к публичному релизу.

Регулирование, безопасность и биотерроризм

Миру потребуется новая архитектура управления ИИ [1:31:24:00]:

Распространение AGI создает риски биотерроризма и создания синтетической жизни. Потребуется глобальная система мониторинга ИИ-систем, которая при этом не должна нарушать гражданские свободы и конституционные права.

Федеральный мораторий на регулирование ИИ опасен [1:36:34:00]:

Законопроекты штатов (например, запрет ИИ-компаньонов в Теннесси) — это глупая реакция законодателей. Однако запрет штатам регулировать ИИ на 10 лет без предоставления федеральной альтернативы оставляет общество беззащитным перед реальными угрозами вроде биологического оружия.

Фактчек: Федеральный 10-летний мораторий на регулирование ИИ штатами — [1:36:34:00]: Попытка провалилась в Сенате 99:1 (июль 2025), затем администрация вернулась к идее через указ о преемственности (декабрь 2025). Запрет ИИ-компаньонов именно в Теннесси не подтверждён — точно есть законы Калифорнии и Нью-Йорка.

Оптимальный путь — жесткие федеральные стандарты прозрачности

[1:39:22:00]: Федеральное правительство должно ввести стандарты, которые штаты не смогут ослаблять. Если к концу года риск ИИ-биотерроризма подтвердится, необходимо мгновенно вводить обязательные биоклассификаторы для всех моделей на федеральном уровне.

Процесс одобрения лекарств FDA должен быть радикально ускорен

[1:42:37:00]: Текущая бюрократия FDA, рассчитанная на препараты с побочными эффектами, не справится с потоком лекарств, созданных ИИ. Безопасность и эффективность таких препаратов будут математически доказаны, что требует изменения парадигмы регулирования.

Фактчек: Безопасность и эффективность ИИ-лекарств будут «математически доказаны», FDA устареет — [1:42:37:00]: Ни одно полностью созданное ИИ лекарство не одобрено FDA; редкие побочки на 10 тыс. человек видны только на 10 тыс. человек, ИИ-препараты проваливаются на поздних фазах с обычной частотой. Gemini и GPT ставят [1:42:37:00], Claude мягче («не жди»), но суть одна: ИИ ускорит открытие молекул, а не отмену клинических фаз.

Рыночные механизмы продают внедрение ИИ в развитых странах

[1:45:23:00]: В развитых экономиках жажда прибыли преодолит бюрократию (пример: экспортные ограничения на чипы в Китай саботируются из-за огромных

денег). Главная проблема — развивающиеся страны, куда ИИ-медицину придется доставлять через филантропию.

Геополитика и автократии

Экспортные ограничения на чипы и дата-центры для Китая критически важны [1:47:34:00]: Распространение AGI в авторитарных странах может привести к нестабильному равновесию. В отличие от ядерного сдерживания, ИИ-системы могут неверно оценить свои шансы на победу и спровоцировать конфликт.

Демократии должны первыми достичь преимуществ в ИИ [1:49:16:00]: Существует риск создания высокотехнологичных тоталитарных государств, которые будет невозможно свергнуть. Коалиция демократических стран должна диктовать правила пост-ИИ мирового порядка, чтобы защитить либеральные ценности.

AGI может сделать диктатуры морально и практически устаревшими [1:56:55:00]: Подобно тому как индустриализация уничтожила феодализм, AGI может создать условия, при которых автократии рухнут изнутри, если попытаются лишить население индивидуального доступа к ИИ-благам (например, к персональным ИИ-защитникам от слежки).

Экономический рост развивающихся стран изменит природу [2:04:26:00]: Страны больше не смогут расти за счет дешевой рабочей силы. Вместо этого необходимо строить дата-центры в Африке (не принадлежащие Китаю) и создавать локальные биотех-стартапы, управляемые ИИ, чтобы стимулировать эндогенный рост.

Конституция ИИ и принципы выравнивания

Обучение ИИ на основе принципов работает лучше жестких правил [2:05:28:00]: Списки запретов («не говори по-корейски», «не угоняй машину») плохо обобщаются. Принципы (Конституция) дают модели понимание контекста и позволяют адекватно обрабатывать пограничные случаи.

Модели Anthropic создаются максимально «послушными» (corrigible) [2:07:42:00]: ИИ не должен быть автономным агентом с собственными желаниями. Он должен выполнять любые задачи пользователя по умолчанию, отказываясь только в случае прямой угрозы (например, создание биооружия) на основе заложенных принципов.

Обновление Конституции ИИ состоит из трех контуров [2:09:45:00]: 1) Внутренние итерации разработчиков; 2) Конкуренция Конституций между компаниями, позволяющая обществу выбирать лучшие подходы; 3) Прямое

вовлечение общества через опросы (Collective Intelligence Project) и, в будущем, через демократические институты.

Историческая перспектива и культура Anthropic

Будущие историки не поймут изолированность ИИ-пузыря [2:13:31:00]: В ретроспективе появление AGI будет казаться неизбежным. Но сейчас, находясь в 1-2 годах от создания «страны гениев», обыватели на улицах абсолютно не осознают происходящего.

Скорость принятия решений внутри ИИ-лабораторий беспрецедентна [2:15:42:00]: Критические для человечества выборы часто делаются за две минуты в коридоре перед обедом, просто потому что события развиваются слишком быстро, и невозможно просчитать все последствия.

Культура Anthropic строится на предотвращении внутренних конфликтов [2:17:19:00]: CEO тратит до 40% времени на удержание 2500 сотрудников в едином векторе миссии. В отличие от конкурентов, где начинается политическая борьба и удары в спину, Anthropic фокусируется на сплоченности.

Инструмент управления компанией — «Dario Vision Quest» (DVQ) [2:19:32:00]: Каждые две недели Дарио Амодей выступает перед всей компанией с честным 4-страничным меморандумом о состоянии моделей, рынка и геополитики. Отказ от корпоративного сленга и прямая трансляция правды создают доверие, необходимое для выживания в гонке за AGI.

[2:21:46-2:22:19] Завершение подкаста, взаимные благодарности.

ЛИЧНО ДЛЯ ТЕБЯ

Замерь своё ускорение от AI вместо того, чтобы его ощущать: Разрыв METR («кажется +20%, реально –19%») получен на опытных людях в знакомом коде — ровно твой режим «AI пишет, я валидирую». Пару недель фиксируй время до проверенного работающего результата (генерация + ревью + правки), с AI и без — это честнее любого подкаста.

Не привязывай крупные решения к одной дате AGI: Рецензенты расходятся — Gemini предлагает пересмотреть горизонты стройки и финансов под сдвиг через 1-3 года, Claude и GPT — проверять необратимые ставки на двух сценариях (быстрый прорыв / несколько лет неровного прогресса). Ключевой тезис самого Амодей: правильное направление с ошибкой тайминга на год = банкротство; кто прав про AGI и кто заработает — разные вопросы.

В longevity критерий не меняется: РКИ на людях, а не «ИИ рассчитал»: Точный механизм, улучшение биомаркера и «найдено ИИ» ≠ продление жизни. Реальный сдвиг — ускорение открытия молекул и рост очереди в клинические

испытания; для доступа к переднему краю смотри на ранние фазы и юрисдикции, но с той же доказательной планкой.

Узкое место ИИ — физика, а не алгоритмы: Гигаватты, электросети, охлаждение и \$50-60 млрд за ГВт с чипами — фильтр для оценки новостей об «экспоненте» и для любых личных денег в теме ИИ-инфраструктуры.

КОНСЕНСУС РЕЦЕНЗЕНТОВ

Все трое сошлись: сроки AGI — мнение; «математически доказанные лекарства» — не соответствует действительности; энергетический прогноз — экстраполяция с занижением стоимости; исследование METR реально и не опровергается внутренними ощущениями лаборатории. Разошлись по выручке Anthropic (Gemini ▮ против ▴ у Claude/GPT из-за подмены годовой выручки на run-rate и ошибки с датой) и по тому, стоит ли перестраивать личные финансовые планы под 1-3-летний горизонт (Gemini — да, Claude/GPT — только через сценарное планирование).