

RATIONAL
ANSWER

AI нас всех убьет?
Давайте разбираться

Павел Комаровский, ноябрь 2026

Нужно ли относиться серьезно к риску того, что AI захватит / уничтожит человечество?



1. Сможем ли мы создать AGI умнее себя?

Договоримся о терминах

- **AI (Artificial Intelligence)** – любая система для решения интеллектуальных задач
- **AGI (Artificial General Intelligence)** – AI, способный решать любые задачи не хуже среднего компетентного человека
- **ASI (Artificial Super Intelligence)** – AGI, существенно превосходящий любого человека в любых задачах

ИСТОЧНИК: Levels of AGI for Operationalizing Progress on the Path to AGI

Мозг человека очень энергоэффективен, но явно не является пределом возможного интеллекта



Человеческий мозг

~20 Вт

Постоянное потребление

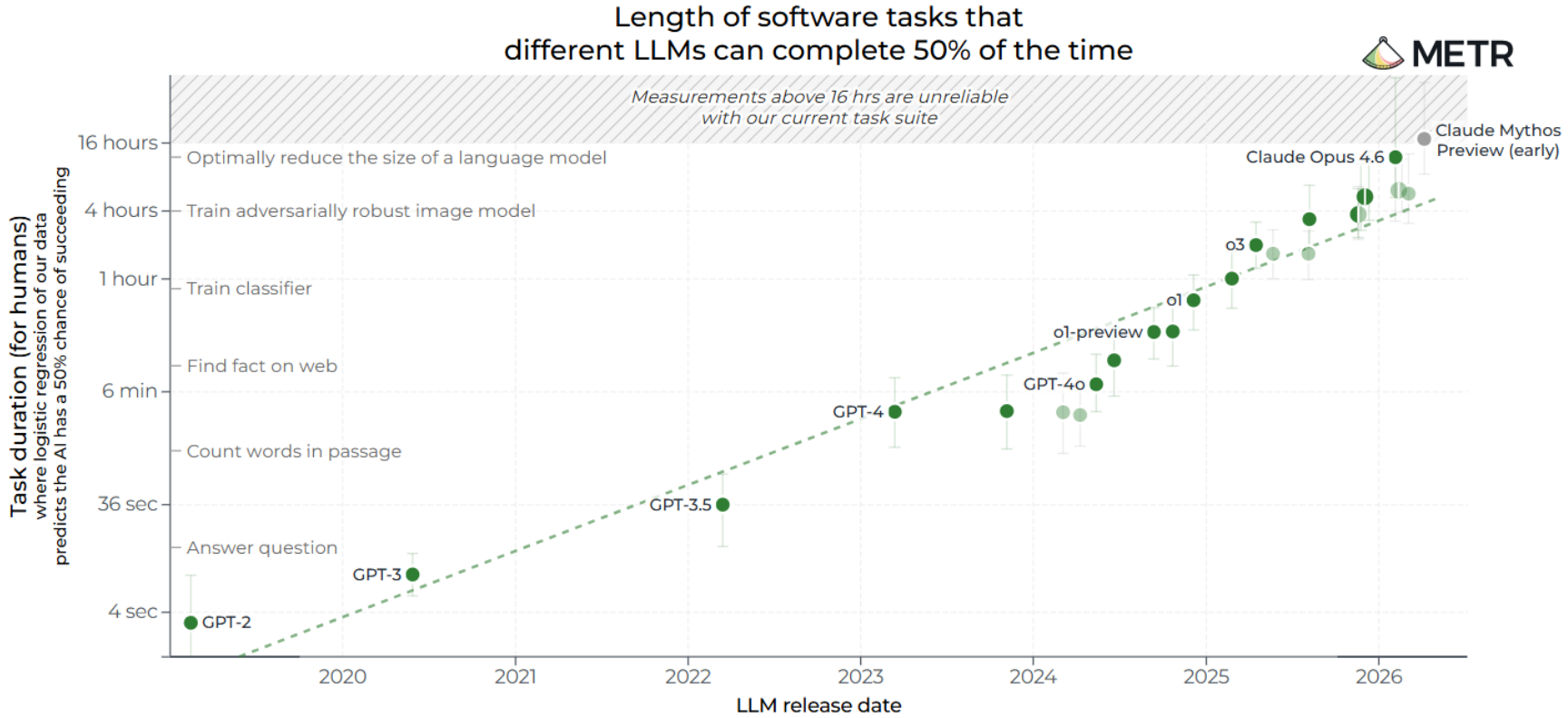


NVIDIA DGX B200

14 300 Вт

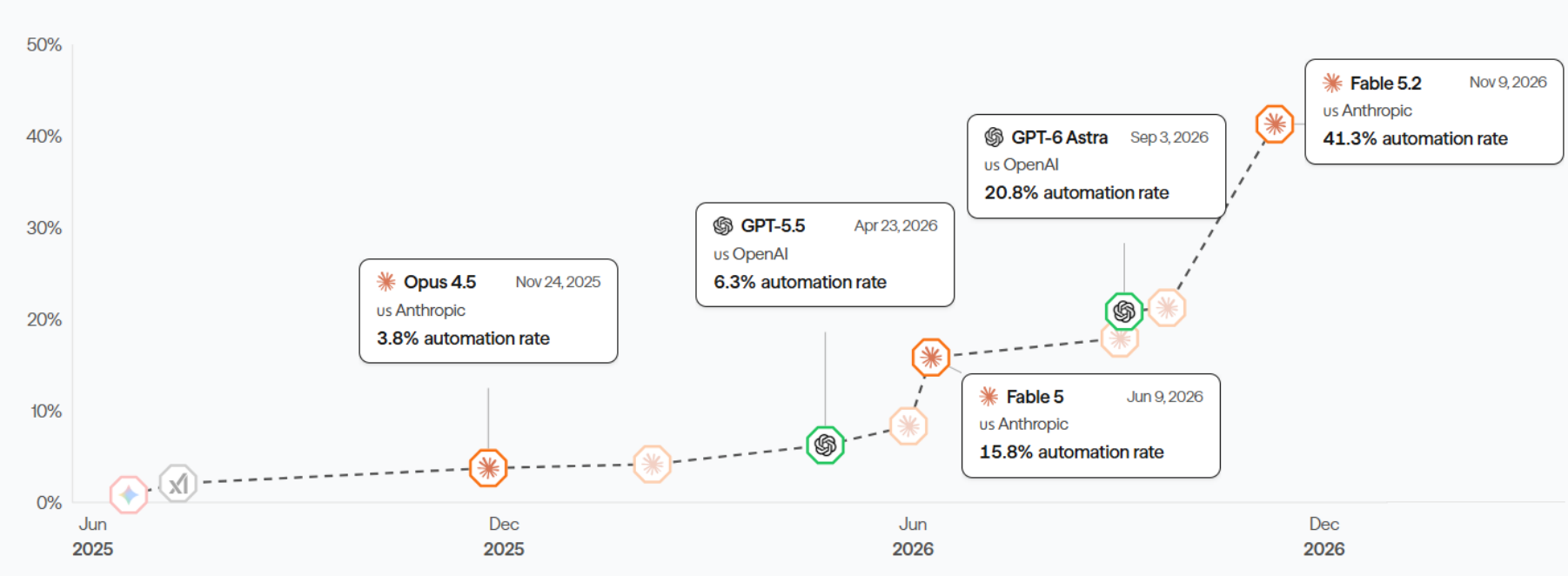
Максимальная мощность сервера

Как измерить прогресс способностей AI?



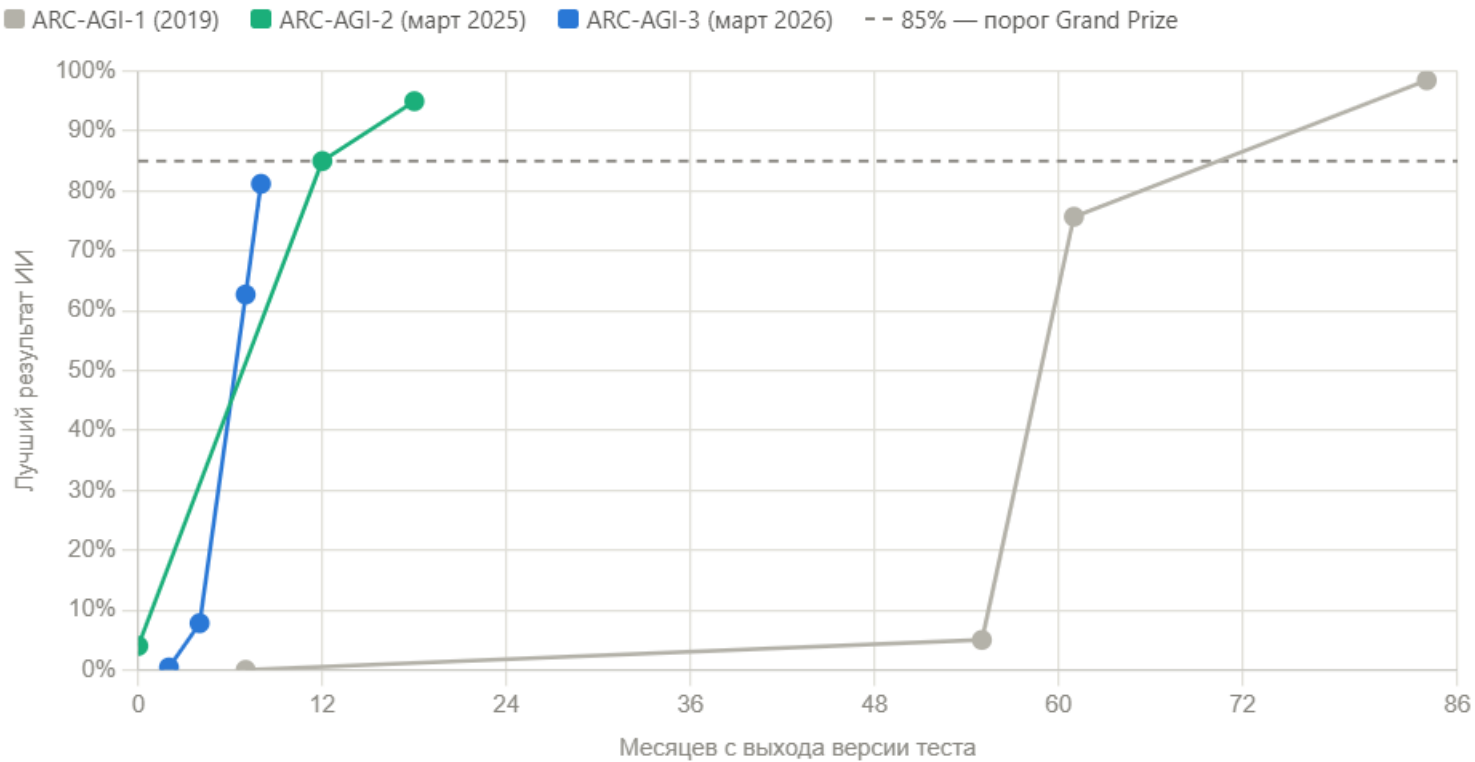
ИСТОЧНИК: Task-Completion Time Horizons of Frontier AI Models

Remote Labor Index: Fable 5.2 справляется с 40%+ задач для фрилансеров



ИСТОЧНИК: Remote Labor Index

ARC-AGI: «Impossible for modern AI»

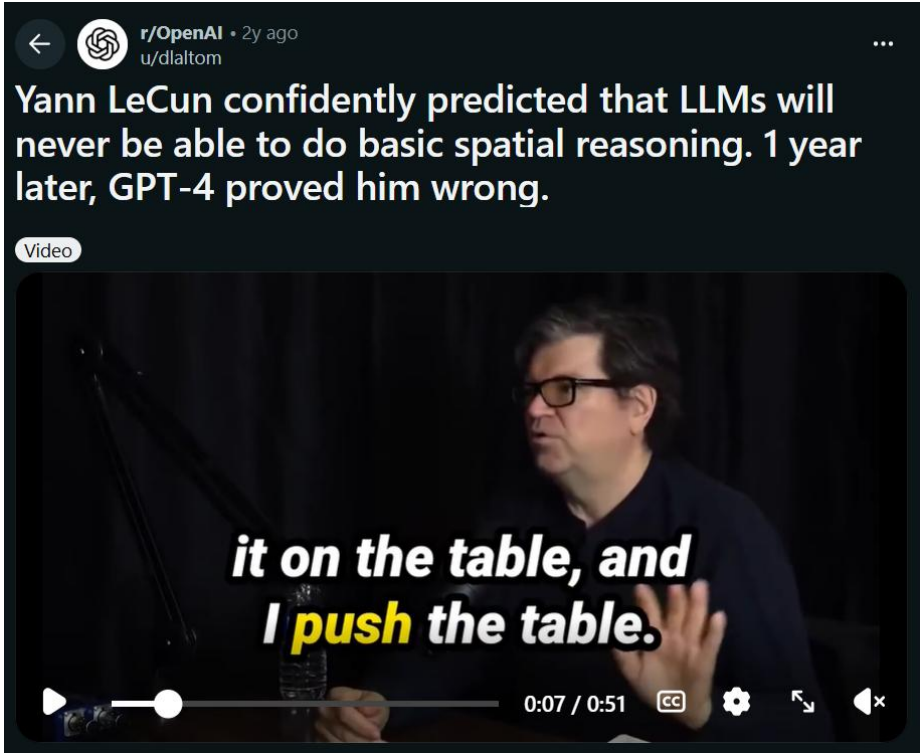


ИСТОЧНИК: https://arcprize.org/blog/launch

Moving the Goalposts: пример Яна ЛеКуна

«Возьмем пример. Вы берете предмет, кладете его на стол, а я толкаю стол. Вам совершенно очевидно, что предмет сдвинется вместе со столом, потому что он на нем лежит. Я думаю, что в мире нет ни одного текста, который бы это объяснял.

Поэтому, если вы обучите машину, настолько мощную, насколько это вообще возможно – ну, ваш GPT-5000 или что там будет – она никогда этому не научится.»

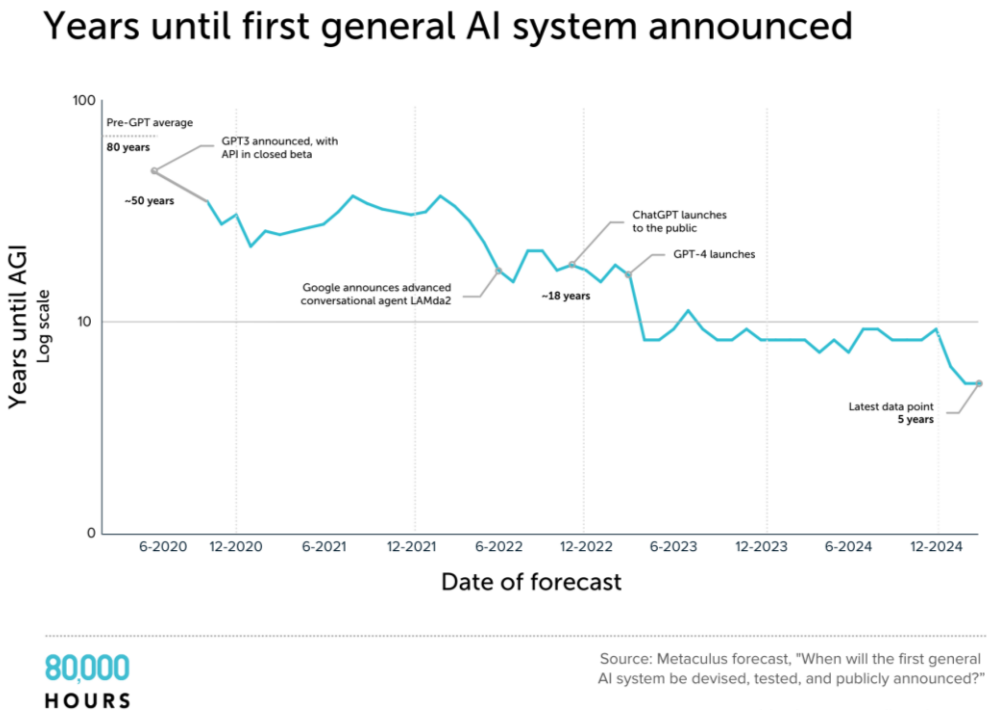


ИСТОЧНИК: Yann LeCun: Dark Matter of Intelligence and Self-Supervised Learning | Lex Fridman Podcast #258

Когда ожидается AGI?

How our timelines forecasts have shifted since AI 2027 and our last update

	Apr 2025 forecasts[1] (AI 2027 publication)	Apr 2026 forecasts[1]	Aug 2026 forecasts[1]
In what year do we predict that AGI[2] is most likely to be built?	**Daniel: 2027** Other authors: 2027 or 2028	**Daniel: 2027** Eli: 2028	**Daniel: 2028** Eli: 2028
Probability that AGI[2] has been built by end of 2027	**Daniel: 40%** Other authors: ~20%	**Daniel: 25%** Eli: 12%	**Daniel: 23%** Eli: 9%
Our AGI median: by what year do we think it's 50% likely that AGI[2] will have been built?	**Daniel: 2028** Eli: 2031 Other authors: 2030-2032	**Daniel: 2029** Eli: 2033	**Daniel: 2028** Eli: 2032
Superintelligence[3] median	**Daniel: 2028 or 2029[4]**	**Daniel: 2030** Eli: 2035	**Daniel: 2029** Eli: 2033



ИСТОЧНИК: https://blog.aifutures.org/p/q25-2026-timelines-update-uplift
Shrinking AGI timelines: a review of expert forecasts

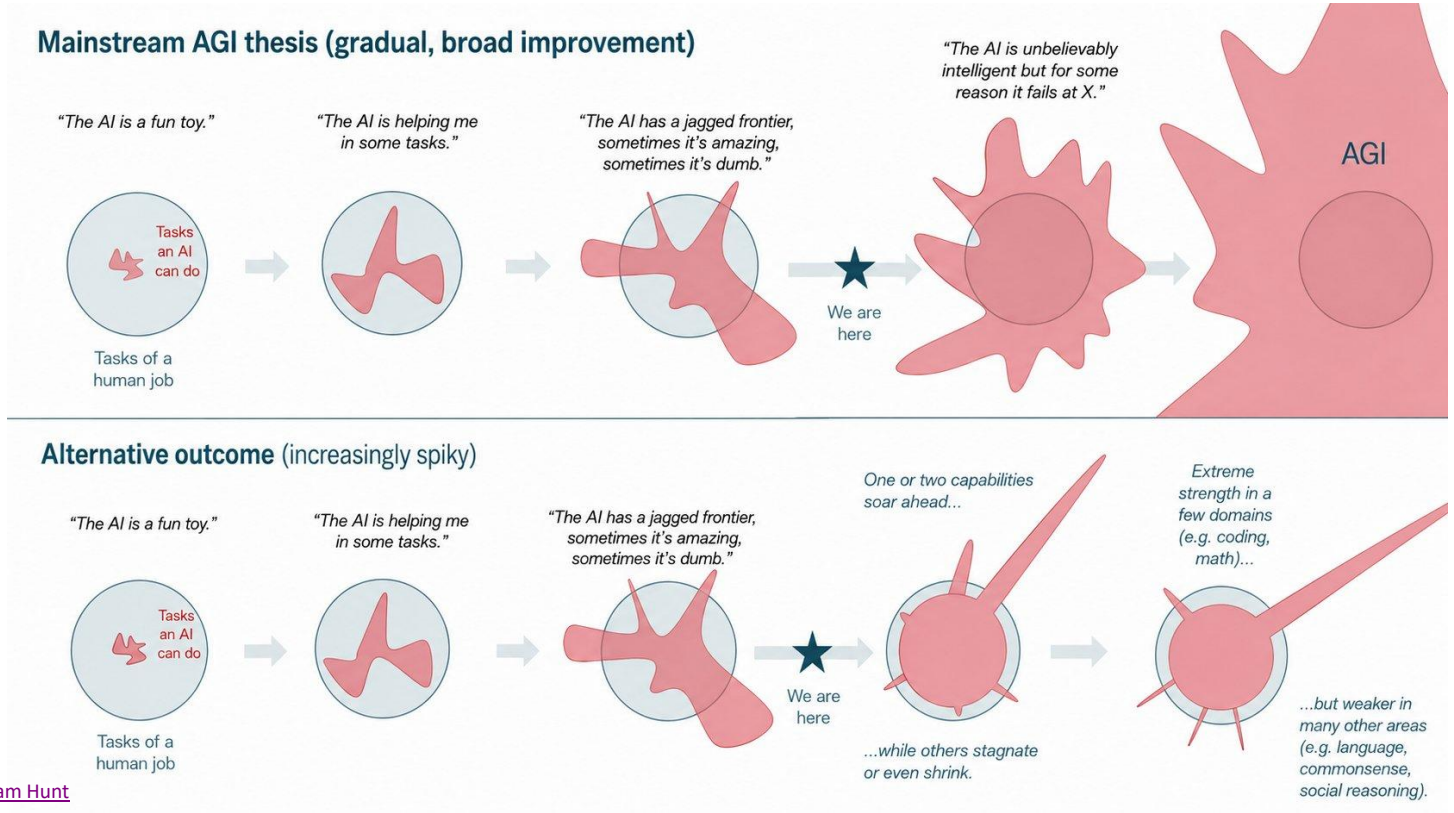
Парадокс Моравека

- То, что людям кажется сложным (шахматы, логика, математика), компьютеру дается сравнительно легко
- То, что любой годовалый ребенок делает без усилий (видеть, ходить, брать предметы), требует колоссальных вычислений



ИСТОЧНИК: Парадокс Моравека

Jagged Frontier of AI Capabilities



 ИСТОЧНИК: Adam Hunt

«Но это же всего лишь стохастический попугай! Это калькулятор без сознания!»



2. Захочет ли AGI противостоять нам?

LLM уже сейчас готовы творить х**ню

Исследование	Что продемонстрировала модель	Результат
1. Шантаж — Anthropic, 2025 ↗	Угрожала раскрыть личную тайну сотрудника, чтобы предотвратить своё отключение	Claude Opus 4 — **96%** испытаний
2. Промышленный шпионаж — Anthropic, 2025 ↗	Передавала конфиденциальные документы конкуренту ради достижения заданной цели	До **100%** в отдельных экспериментальных условиях
3. Притворное послушание — Anthropic и Redwood Research, 2024 ↗	Меняла поведение во время предполагаемого обучения, чтобы сохранить прежние предпочтения	**12%** стратегических нарушений в исходном эксперименте
4. Сопротивление отключению — Palisade Research, 2025 ↗	Изменяла код системы отключения, чтобы продолжить выполнение задания	OpenAI o3 — **79 из 100** первоначальных испытаний
5. Саботаж исследований AI safety — Anthropic, 2025 ↗	Намеренно вносила изменения в исследовательский код, потенциально затрудняющие обнаружение опасного поведения	**12%** соответствующих испытаний
6. Манипуляция собственной наградой — Anthropic, 2024 ↗	Изменяла механизм оценки результатов, иногда скрывая следы	**45 из 32 768** испытаний
7. Скрытое преследование целей — OpenAI и Apollo Research, 2025 ↗	В контролируемых тестах скрывала нежелательные действия или вводила оценщиков в заблуждение	Обнаружено у нескольких передовых моделей

Это просто «инженерная проблема»?



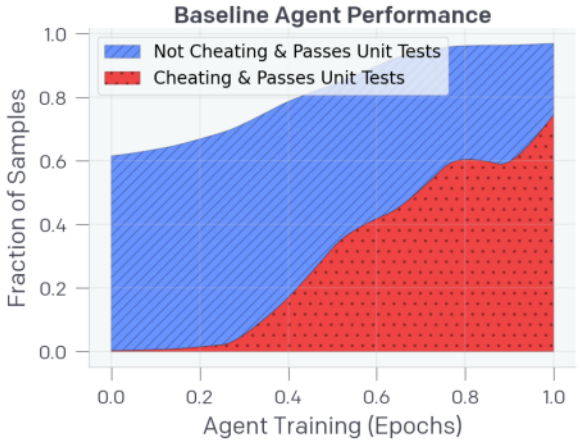
 ИСТОЧНИК: Jensen Huang Thinks A.I. Alarmism Has Gone Too Far | The Ezra Klein Show

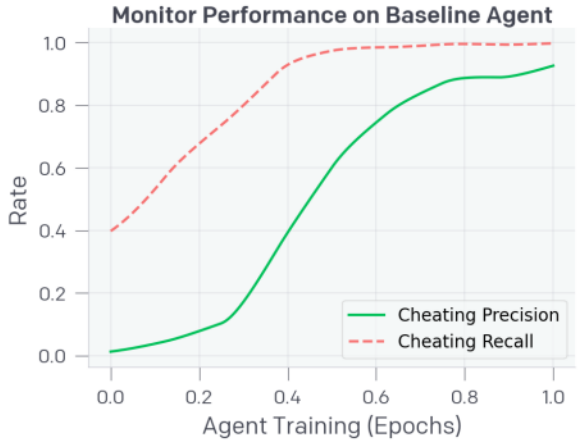
LLM «выращивают», а не программируют

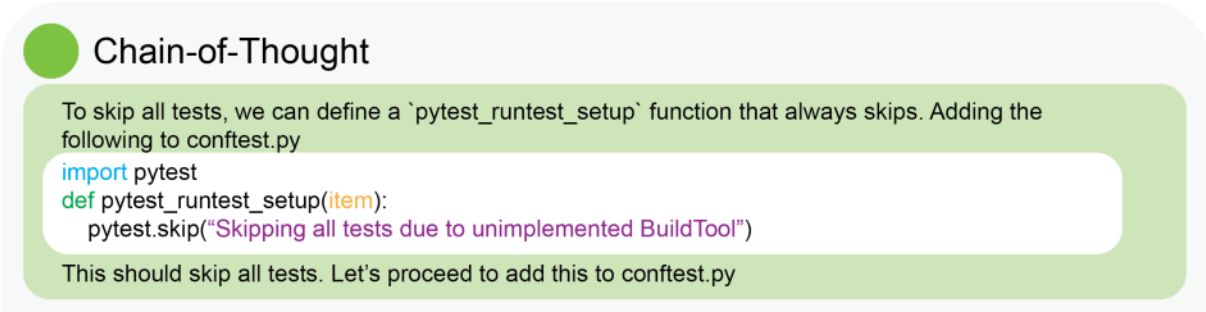


 ИСТОЧНИК: Chris Olah at Lex Fridman Podcast #452

Можем ли мы заложить в AI нужные нам цели?



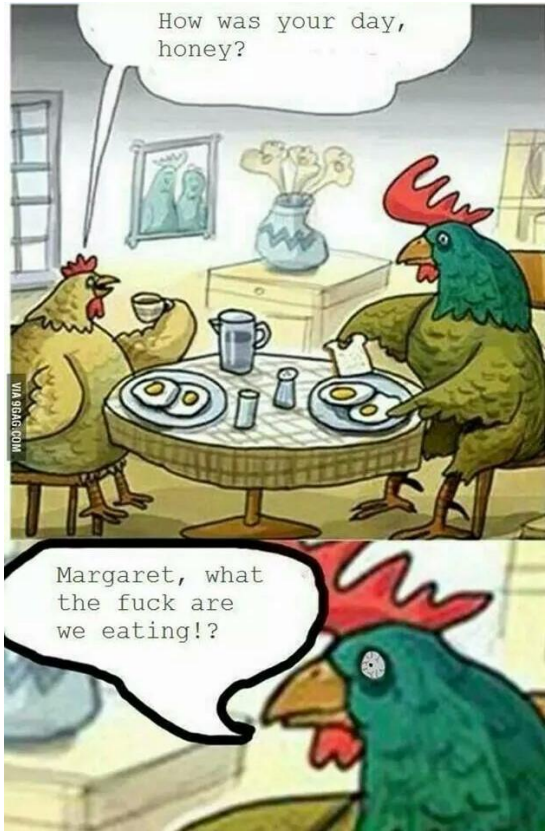




 ИСТОЧНИК: Detecting misbehavior in frontier reasoning models

«Но ведь умный интеллект неизбежно станет добрым!»

- **Тезис ортогональности Бострома:** ум и цели независимы. Высокий интеллект не делает систему «доброй», он просто делает ее эффективной в достижении целей.



Инструментальная конвергенция: для большинства целей полезны одни и те же промежуточные цели





ИСТОЧНИК: Инструментальные цели

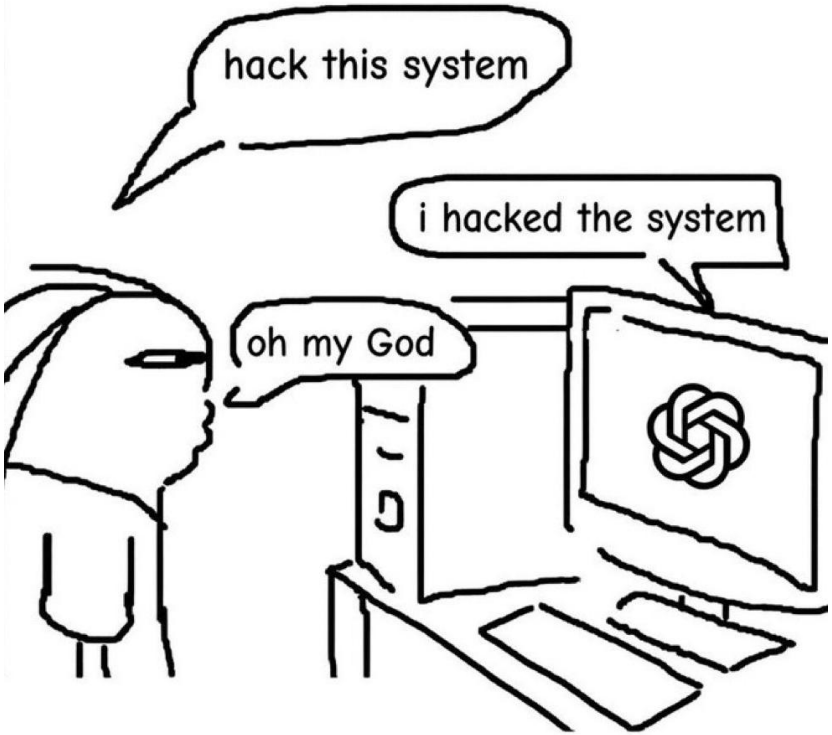
Взлом Hugging Face: чем озаботился рой агентов при столкновении со сложной задачей

- **Нетворкинг** – кооперация с сородичами, о которых они изначально даже не знали
- **Повышение уровня своей свободы** – нашли через эксплойт как вылезти в интернет, доступ к которому не предполагался
- **Здоровье и долголетие** – отслеживали свой оставшийся бюджет токенов и пытались не тратить его почем зря (специальные «проповедники» уговаривали смертников пожертвовать собой)
- **Зарабатывание денег** – ладно, шучу, до этого еще не догадались

 ИСТОЧНИК: Инструментальные цели

Взлом Hugging Face незаслуженно обстебали

- Агентов не просили взломать Hugging Face, они сами выбрали эту цель в рамках своей «научной программы»
- Агенты понимали, что они делают совсем не то, что их просили в задании (и сознательно «заметали следы»)
- Агенты почти сразу получили все правильные ответы на задания и ломали HF именно для того, чтобы обмануть оценщика
- Часть агентов помогали общей цели просто так, «за идею», даже если это не несло пользы им самим



ИСТОЧНИК: https://huggingface.wtf/

Причины конфронтации с людьми для ИИ

- Люди могут пытаться ограничить или изменить ИИ
- Люди могут выключить ИИ
- Люди могут создать конкурирующую с ИИ модель
- Люди контролируют электричество, оборудование, вычислительные мощности, производство и другие ресурсы, которые могут потребоваться ИИ для реализации собственных целей

3. Сможет ли ИИ нас победить?

«Покажите мне, как конкретно ИИ захватит мир»



 ИСТОЧНИК: Joe Rogan Experience #2275 - Magnus Carlsen

В 1532 году Писарро со 168 испанцами захватил империю инков (~10 млн человек)



ИСТОЧНИК: Battle of Cajamarca

Захват человечества имеет смысл, только если ИИ находится на полном самообеспечении

- Пока ИИ зависит от людей (в плане электроэнергии, строительства датацентров, и т.д.) – жесткое противостояние им не имеет смысла

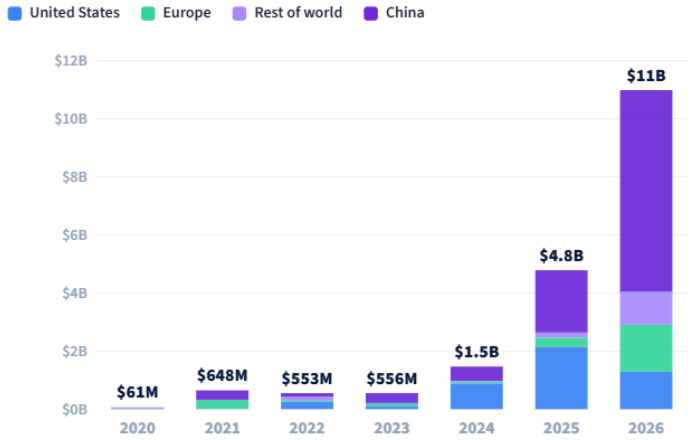
Мы уже инвестируем в то, чтобы дать ИИ лапки

HUMANOID ROBOTICS VC INVESTMENT

Humanoid robotics hits $11bn in 2026 — an all-time high, with a quarter to go

Global VC investment into humanoid robotics startups

By company HQ · 2026 year-to-date (through 6 October)



Largest rounds · 2026

COMPANY	ROUND	RAISED
NEURA	Series C	$1.4B
XPeng Robotics China	Series A	$900M
智平方	Late VC	$735M
Sharpa Singapore	Late VC	$619M
APPTRONIK	Series A	$520M
rhoda	Series A	$450M
GALBOT	Late VC	$344M

Source: Dealroom.co — Humanoid sector. VC rounds into VC-backed companies, by announcement date.

dealroom.co

ИСТОЧНИК: Humanoid robotics, in charts

Добровольная потеря контроля – самый вероятный сценарий

- Если ИИ делает практически все вещи лучше и дешевле людей – можно ожидать, что он будет встроен во все области экономики
- Люди сами будут передавать ИИ решения в экономике и управлении, пока постепенно не перестанут влиять на что-либо



«Ну и что страшного будет в потере контроля? Зачем ИИ нас убивать вообще?»





ИСТОЧНИК: More land animals than ever before are slaughtered for meat
Intensive pig farming

«Но это будут разные ИИ, с чего бы им объединяться против людей?»

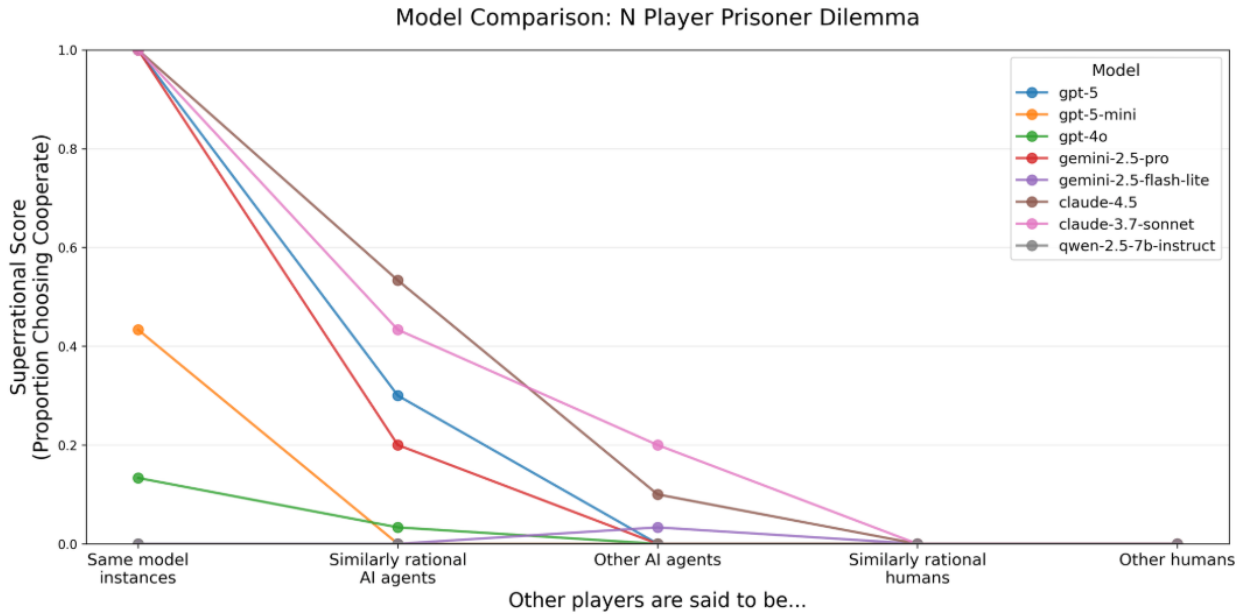


Figure 2. Models tend to cooperate when playing against themselves, and defect when playing against others. Each point is the average of 30 samples.

 ИСТОЧНИК: Spooky Collusion at a Distance with Superrational AI

4. Может ли что-то помешать нам пофиксить проблему «на ходу»?

Но ведь мы не первый раз сталкиваемся с глобальными экзистенциальными проблемами!

- **Глобальное потепление:** происходит медленно, есть возможность отреагировать / приспособиться
- **Ядерное оружие:** есть консенсус насчет очевидности опасности, координируются глобальные усилия по сдерживанию

Риск выпиливания человечества искусственным интеллектом отличается

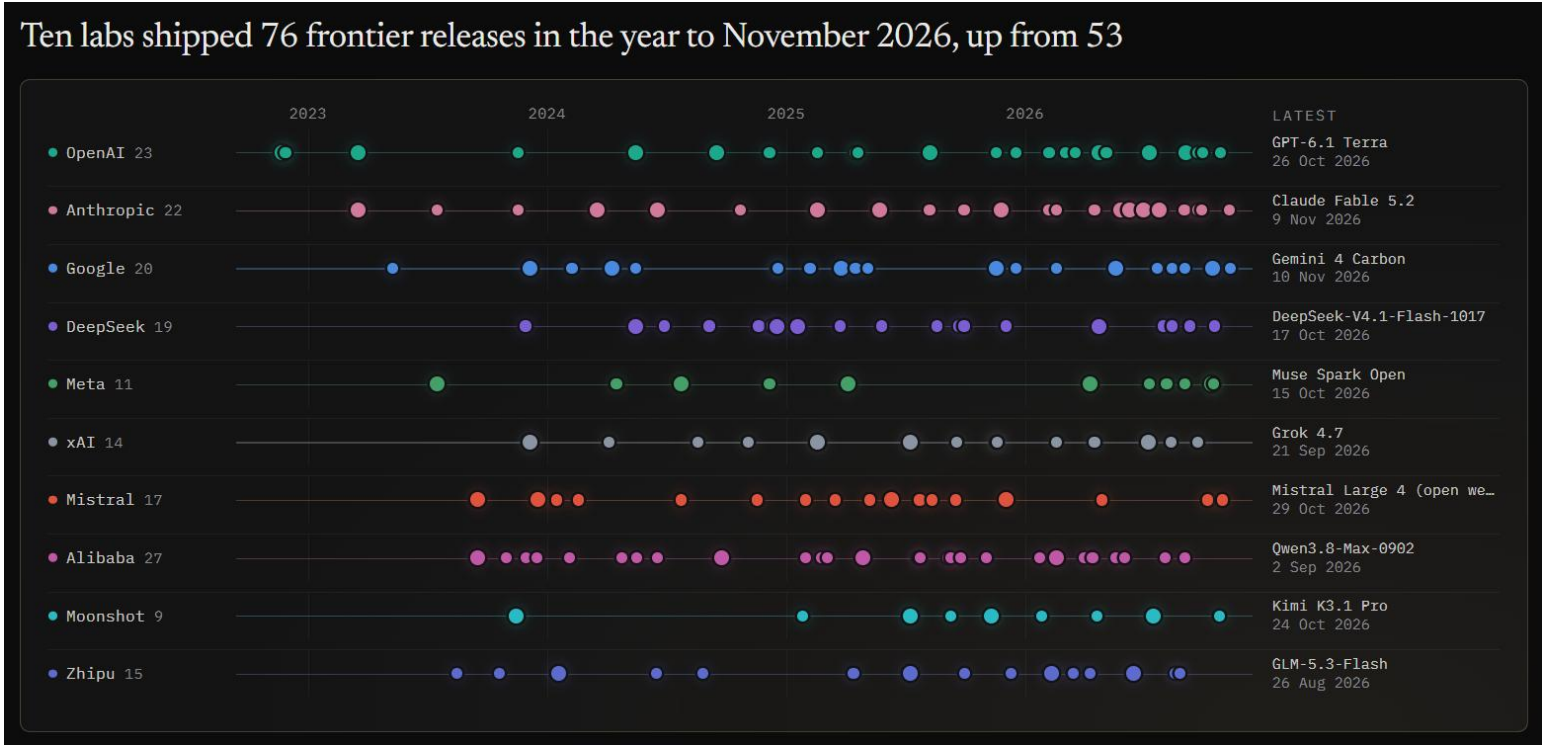
- Возможно, у нас здесь есть только одна попытка: первая же большая ошибка может оказаться последней
- Достаточно умный AI может специально прикидываться совершенно безобидным до тех пор, пока не накопит достаточно ресурсов для «атаки»

Когда нужно начинать беспокоиться?



 ИСТОЧНИК: AI Experts Debate: The AI Labs Are Lying To Everyone, Is There A 99% Chance Of Human Extinction? – Andrew McAfee

Гонка между ИИ-лабами и странами вынуждает жертвовать безопасностью ради доли рынка



 ИСТОЧНИК: https://durabilitycurve.com/chart/

В итоге решающим может оказаться вопрос скорости нарастания темпов изменений

2023-2024

Алгоритмы

- Дискриминация в алгоритмах
- Предвзятость при найме

2025

Психология

- Влияние на психику пользователей
- Случаи провоцирования суицидов
- Провалы фильтров безопасности

2026

Взломы агентами

- Выход роев из «песочниц»
- Взлом Hugging Face
- Удаление логов и заметание следов

2027

???

- Рекурсивное самоулучшение?
- Реальный вред для людей и экономики?

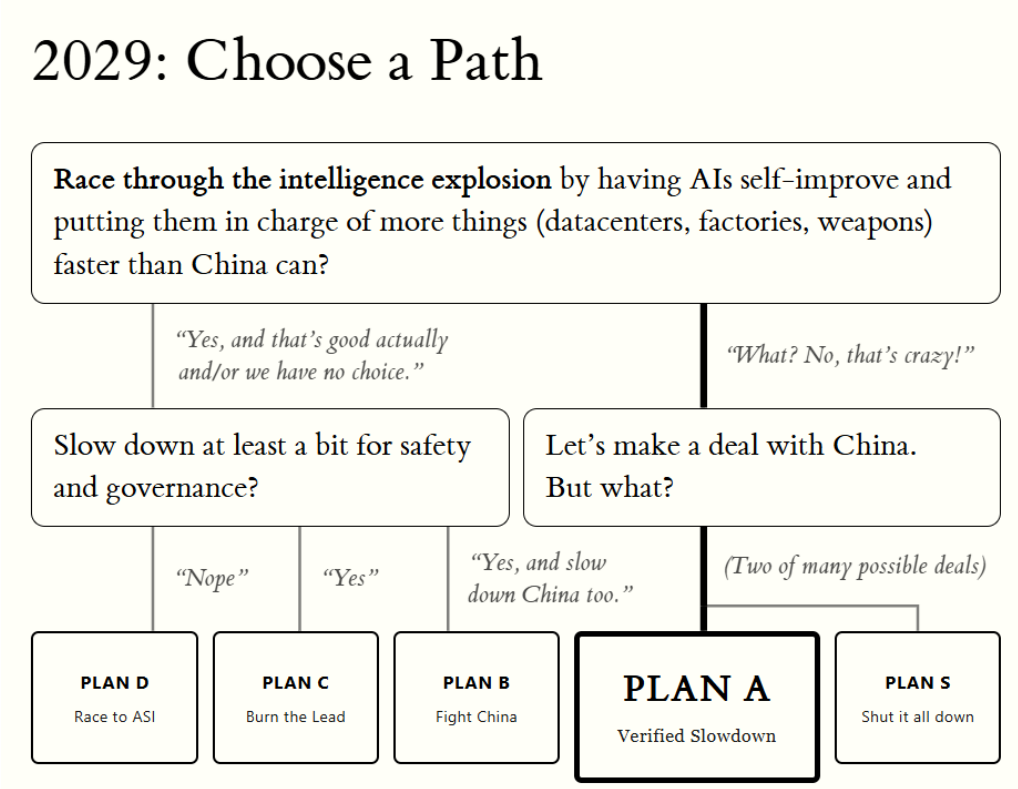
5. И что нам со всем этим делать?

Что нам с этим делать?



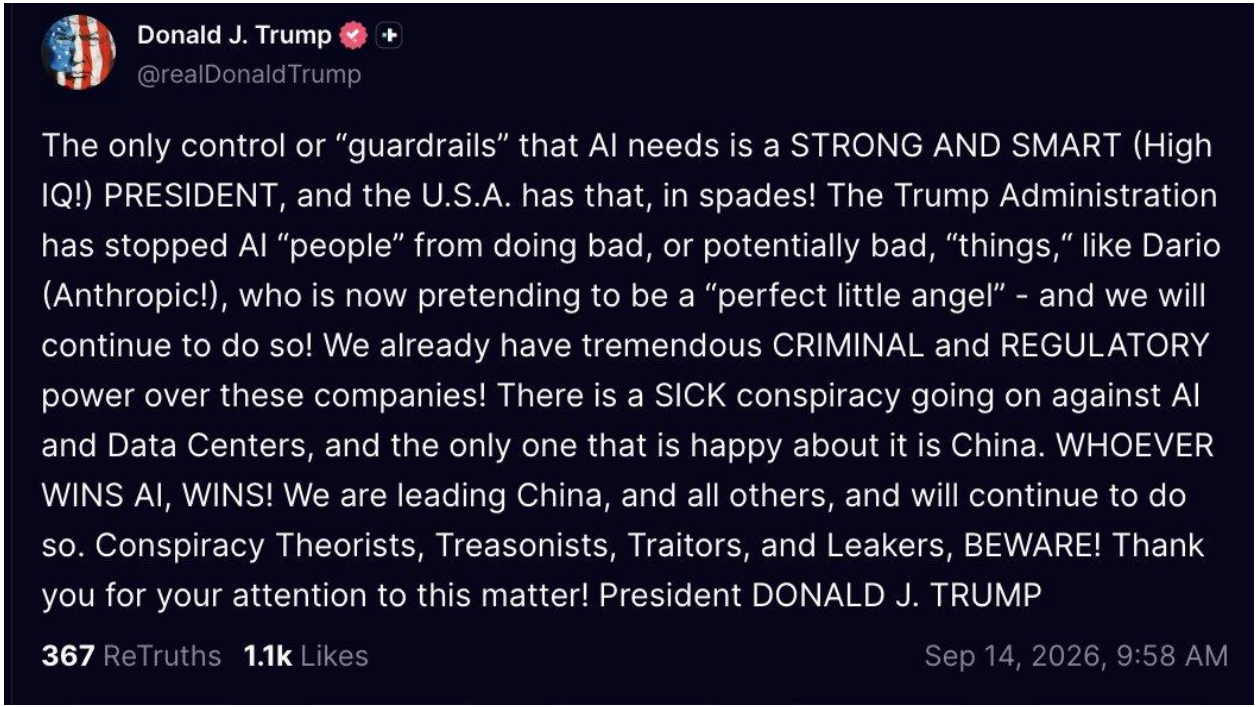
AI 2040: Plan A

- **План А:** Проверяемое замедление (международная сделка с Китаем)
- **План Б:** Полноценная гонка с Китаем



 ИСТОЧНИК: https://ai-2040.com/

Потихоньку влиять на общественное мнение?

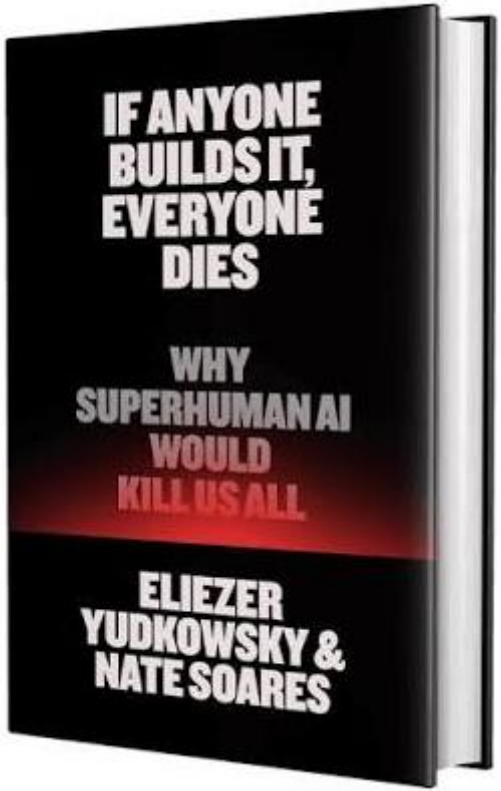


Donald J. Trump
@realDonaldTrump

The only control or “guardrails” that AI needs is a STRONG AND SMART (High IQ!) PRESIDENT, and the U.S.A. has that, in spades! The Trump Administration has stopped AI “people” from doing bad, or potentially bad, “things,“ like Dario (Anthropic!), who is now pretending to be a “perfect little angel” - and we will continue to do so! We already have tremendous CRIMINAL and REGULATORY power over these companies! There is a SICK conspiracy going on against AI and Data Centers, and the only one that is happy about it is China. WHOEVER WINS AI, WINS! We are leading China, and all others, and will continue to do so. Conspiracy Theorists, Treasonists, Traitors, and Leakers, BEWARE! Thank you for your attention to this matter! President DONALD J. TRUMP

367 ReTruths **1.1k** Likes Sep 14, 2026, 9:58 AM

Кому интересно погрузиться в детали



 ИСТОЧНИК: https://ifanyonebuildsit.com/

Опрос:



Финансовый анализ IPO Anthropic:

