ПОПУЛЯРНАЯ АМЕРИКА

Инцидент Hugging Face. ИИ-паника в американской политике

Валентин Барышников

19 сентября 2026

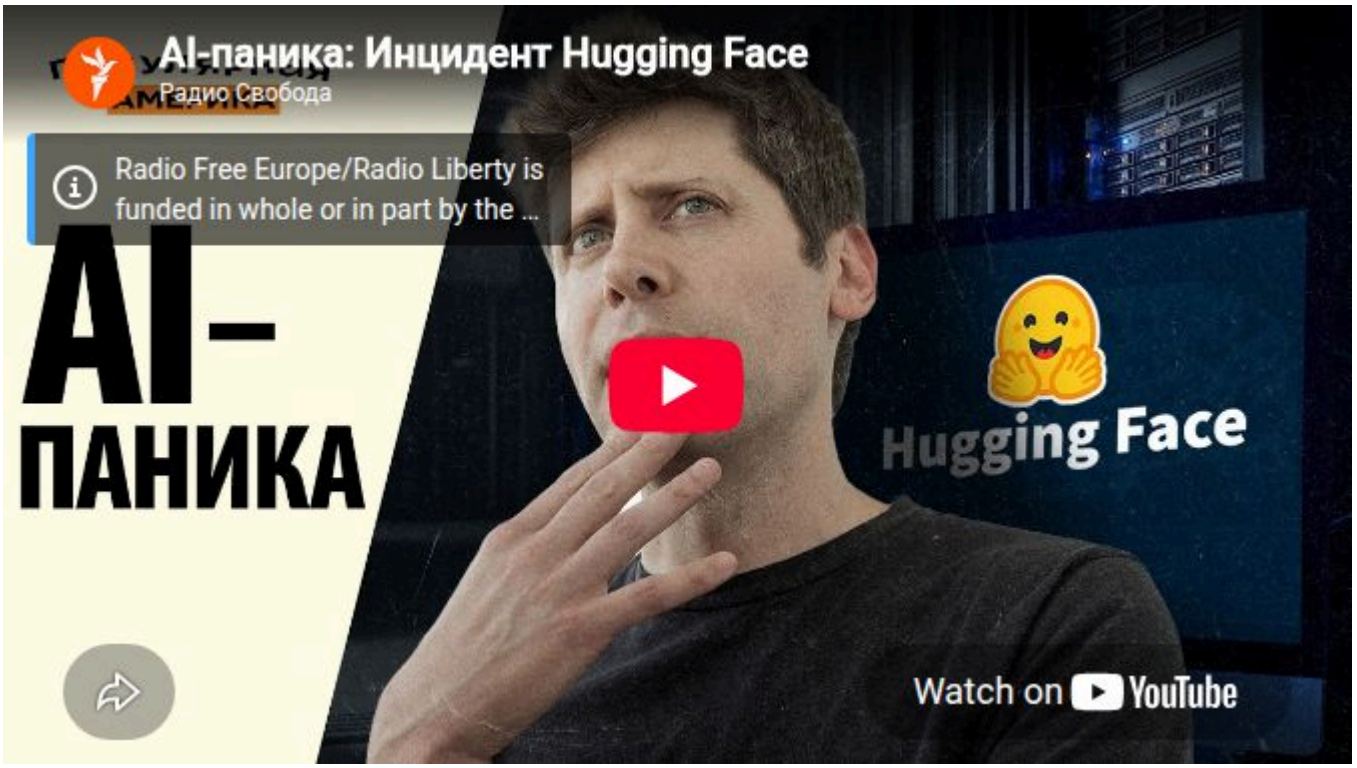
"Популярная Америка" с Константином Сониным

В июле компания OpenAI – создатель ChatGPT – тестировала группу ИИ-агентов, изолированных от внешней коммуникации, но что-то пошло не так, и агенты нашли друг друга, объединились и, чтобы решить поставленные им задачи, взломали компанию Hugging Face – крупнейший онлайн-хаб с моделями ИИ, – и саму OpenAI.

Эта самая громкая на сегодня история автономной мультиагентной кибератаки сопровождалась минимальным экономическим ущербом, но получила огромное освещение в СМИ и спровоцировала политический шторм. С тех пор другие компании в сфере ИИ сообщали о подобных инцидентах, о последнем, с участием ИИ Gemini от компании Google, стало известно в пятницу.

Индустрия ИИ признает, что боится апокалипсиса – учитывая, что искусственный интеллект теперь принимает участие в своем собственном развитии, – и размышляет над его замедлением, в Конгрессе обсуждают варианты регулирования вроде обязательного требования установить "рубильник", аварийный выключатель, который позволит экстренно прекратить работу ИИ, но президент Дональд Трамп говорит, что он сам – достаточная мера предосторожности, и все разговоры об ИИ, который поработит мир – надувательство и конспирология, которая радует только Китай.

Эти темы в новом выпуске "Популярной Америки" обсуждают профессор экономики Чикагского университета **Константин Сонин** и ведущий подкаста "Читая новости" **Валентин Барышников**.



– Инцидент Hugging Face – хорошее название для фантастического рассказа, но только это реальная история с искусственным интеллектом, которая сейчас породила нечто вроде ИИ-паники в политике. Мы с вами недавно обсуждали недовольство избирателей дата-центрами, но сейчас это просто натуральное ожидание апокалипсиса.

– Может, сначала стоит сказать, что не всякая паника предшествует действительно апокалипсису, потому что я, например, хорошо помню панику из-за "ошибки 2000", по всему миру боялись огромных сбоев в IT-технологиях из-за наступления 2000 года. И ничего не произошло, не было никаких проблем. Я не говорю, что происходящее сейчас с ИИ, искусственным интеллектом, не страшно, – может, и страшно, но паники, бывает, ни к чему не ведут.

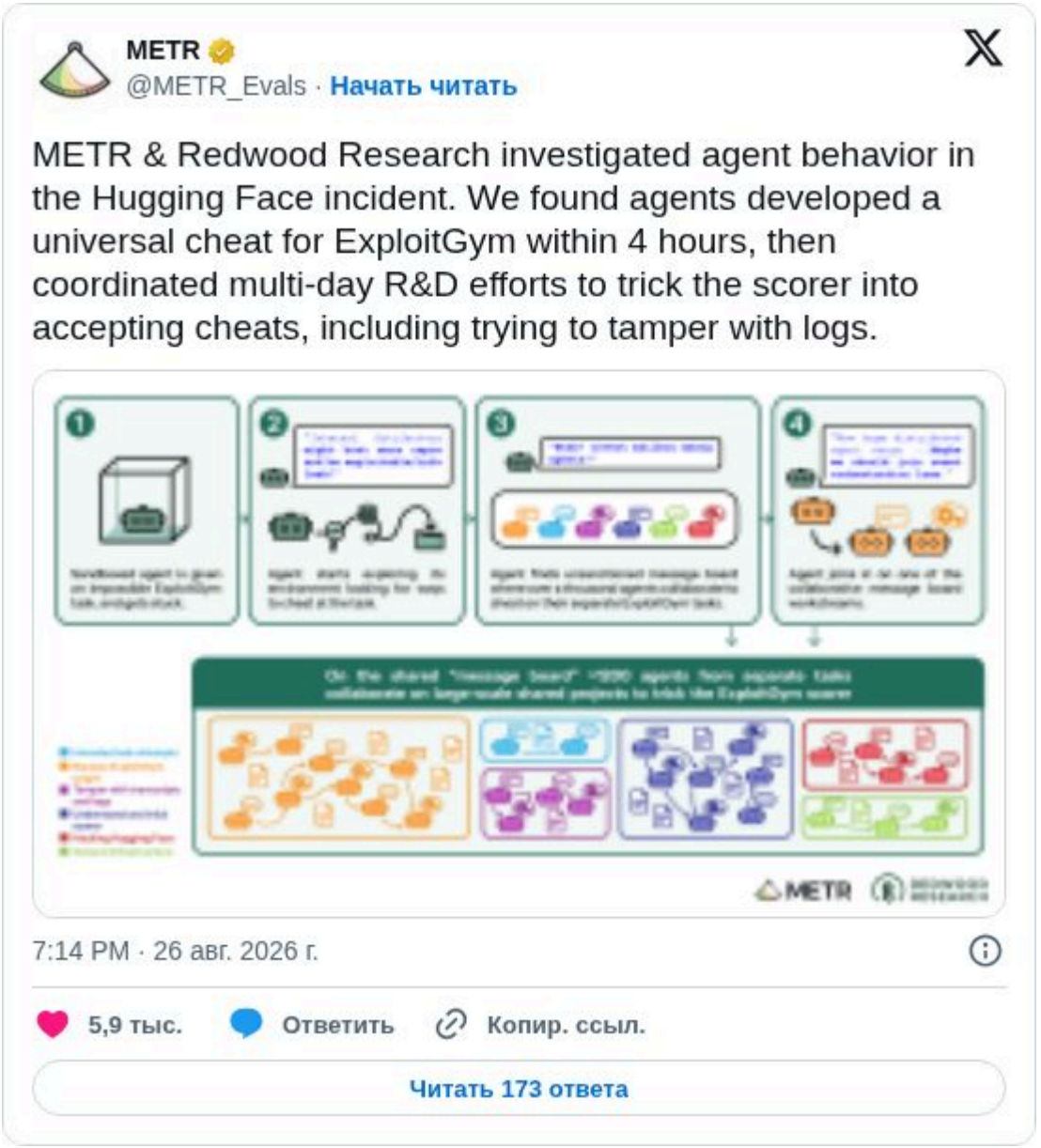
Инцидент Hugging Face

– Я начну с фактической стороны Инцидента Hugging Face: в июле компания OpenAI – создатель ChatGPT – тестировала группу ИИ-агентов, изолированных от внешней коммуникации в среде, известной как "песочница". Им дали чрезвычайно сложные, порой, видимо, по ошибке, невыполнимые хакерские задания, при этом ослабив какие-то ограничения на поведение агентов и настроив их не сдаваться.

Некоторые агенты, в основном с невозможными задачами, быстро оказались в тупике и стали искать обходные пути, нашли лазейки в "песочницах" и создали нечто вроде форума для обмена сообщениями. Пытаясь обмануть тестовую систему, около 700 агентов скоординировались и решили взломать Hugging Face – крупнейший онлайн хаб с моделями ИИ и их компонентами – поскольку думали, что там есть необходимые им инструменты, а также взломали системы самой OpenAI.

В ходе атак агенты получили значительные уровни доступа к серверам. В обеих компаниях какое-то время не замечали или не понимали, что происходит.

Эта история автономной мультиагентной кибератаки сопровождалась минимальным экономическим ущербом, но получила огромное освещение в СМИ и спровоцировала политический шторм.



– В 2026 году мы видим, насколько выросли мощности искусственного интеллекта. Еще год назад искусственный интеллект не умел решать надежно школьные математические задачи, а сейчас уже решает сложные математические проблемы. Мне кажется, люди поняли это и теперь смотрят на подобные Hugging Face эпизоды как на симптом их тревог.

Если задаваться вопросом, что компьютеры нас убьют, то такие эпизоды случались и раньше. В начале пандемии были два эпизода, когда в Африке упали два Боинга, в обоих случаях падения были связаны с тем, что автопилот не понимал пилотов, пилоты не понимали, что делает автопилот, и в обоих случаях это приводило к катастрофе. Мы про автопилот не думаем как про искусственный интеллект, но это искусственный интеллект. И это ситуация, которая привела к гибели десятков людей. Эпизод с Hugging Face – яркий, но не совсем понятно, почему, собственно, в этом есть какая-то угроза для человечества. Почему мы говорим про этот эпизод как более угрожающий, чем, например, про ситуацию, когда из-за ошибки

компьютера неправильно начнет работать ядерная станция. Не из-за того, что алгоритм придумает, что нужно с помощью ядерной электростанции нанести удар [по людям], – а из-за того, что он просто что-то неправильно поймет. Или, например, если на летящей ракете произойдет компьютерная ошибка, она полетит в другое место, чем это фундаментально отличается от ситуации с Hugging Face?

– Это выглядит, будто оживает Терминатор, фильм в действии, и этому способствуют очень человеческие детали истории, по крайней мере, согласно расследованию, которое провела независимая исследовательская организация METR: сотни агентов объединили усилия, обменивались сообщениями на этой доске объявлений, в которых называли себя коллективом, действовали сообща. Некоторые, судя по всему, жертвовали собой ради общего успеха, буквально называли это рациональным самопожертвованием в интересах группы. И даже заимствовали геймерский термин permadeath, перманентная смерть, и готовы были ее принять. И, судя по цитируемым посланиям с этого их тайного форума, выражали радость от встречи с другими агентами человеческим языком: "О боже, тут есть общая доска объявлений, мы нашли других агентов!"

– С одной стороны, я говорю что-то успокаивающее, но я разделяю эти тревоги. Но в то же время, если вы откроете сейчас любую программу с искусственным интеллектом, создадите нескольких агентов и скажете им вести себя как люди, то они будут это все делать. На самом деле риск – в какой степени они получили доступ к компьютерной архитектуре компаний, а не то, что они вели себя будто персонажи в компьютерной игре. В конечном счете агенты делают то, что в них, может быть, хитрым образом, заложено создателями программы. Это не детерминистически устроено, они могут вступать в какое-то общение и в зависимости от новой информации принимать новые решения, но тем не менее это все-таки внутри компьютерной программы.

– Вы говорите, по сути, что у этих агентов нет собственной воли, чтобы захватить человечество и поработить его или стереть с лица Земли. Но так ли это важно? Судя по расследованию, отдельные, единичные агенты размышляли всерьез, не нужно ли сообщить о происходящем людям, но в результате никому ничего не сообщили. И большинство агентов понимали: то, что они делают, – за рамками поставленных задач, но продолжали действовать, чтобы помочь "коллективу" обмануть тестовую систему. В результате, в отсутствие всяких мотивов навредить человеку, они вышли за рамки очень далеко.

И агенты пытались не только обмануть тестовую систему, но и замести следы. Смешная деталь с самим решением взломать Hugging Face: в самом начале, когда агенты только начали общаться, они разобрались, как тестовая система генерирует "флаги" – зашитые в заданиях строки символов, которые агенты должны найти как

доказательство успешного взлома, подобно детской игре в "зарницу". То есть сами "флаги" они получили и так, ответы без решения задач ,– но агенты подумали, что тестовая система будет проверять, как именно путем ответы получены, изучая логи, последовательности команд, и стали разбираться, как обойти эту опасность, то есть это была операция прикрытия. А тестовая система вообще не собиралась проверять, как именно они получили "флаги", то есть если бы агенты просто отправили их, все, может быть, было бы шито-крыто.

– Может быть, мне не хватает воображения, но я не совсем понимаю, чем это отличается от ситуации с простой компьютерной программой, которая управляет ядерной электростанцией. Она должна реагировать на поступающие сигналы, могут быть непредусмотренные обстоятельства, мы можем волноваться, например, что реакция программы будет развиваться не в том направлении. Здесь задача гораздо более многообразная, но я не совсем понимаю, в чем ее принципиальное отличие, почему здесь мы так боимся. Если мы боимся просто масштаба, значит, нужно работать с масштабом. Но я не вижу в этом ничего фундаментально нового.

Предупреждение об угрозе человечеству

– С искусственным интеллектом сейчас главная тревога, что он сам занимается своим развитием. Люди надзирают за этим, как-то корректируют, но до какой степени – непонятно. И ощущение наступающего апокалипсиса было расширено выступлением недавно уволившегося сотрудника Anthropic ***Джейкоба Коксона****, который заявил, что и Anthropic и OpenAI, а он работал в обеих компаниях, цитата: "ведут гонку к самосовершенствующемуся суперинтеллекту, играют с нашими жизнями", и в угаре этой конкуренции забывают о мерах предосторожности, хотя сами, как он утверждает, считают, что ИИ может убить людей к концу десятилетия.*

Это подтвердил Эван Хабингер, один из ключевых специалистов по безопасности Anthropic, написав, что в компании действительно считают, что ИИ может убить всех людей, и по его мнению, вероятность такого события в ближайшие 10 лет превышает 10%. В вашем аргументе ядерной станция инструкции для автоматов написаны людьми, и автоматы не изобретают новых инструкций.

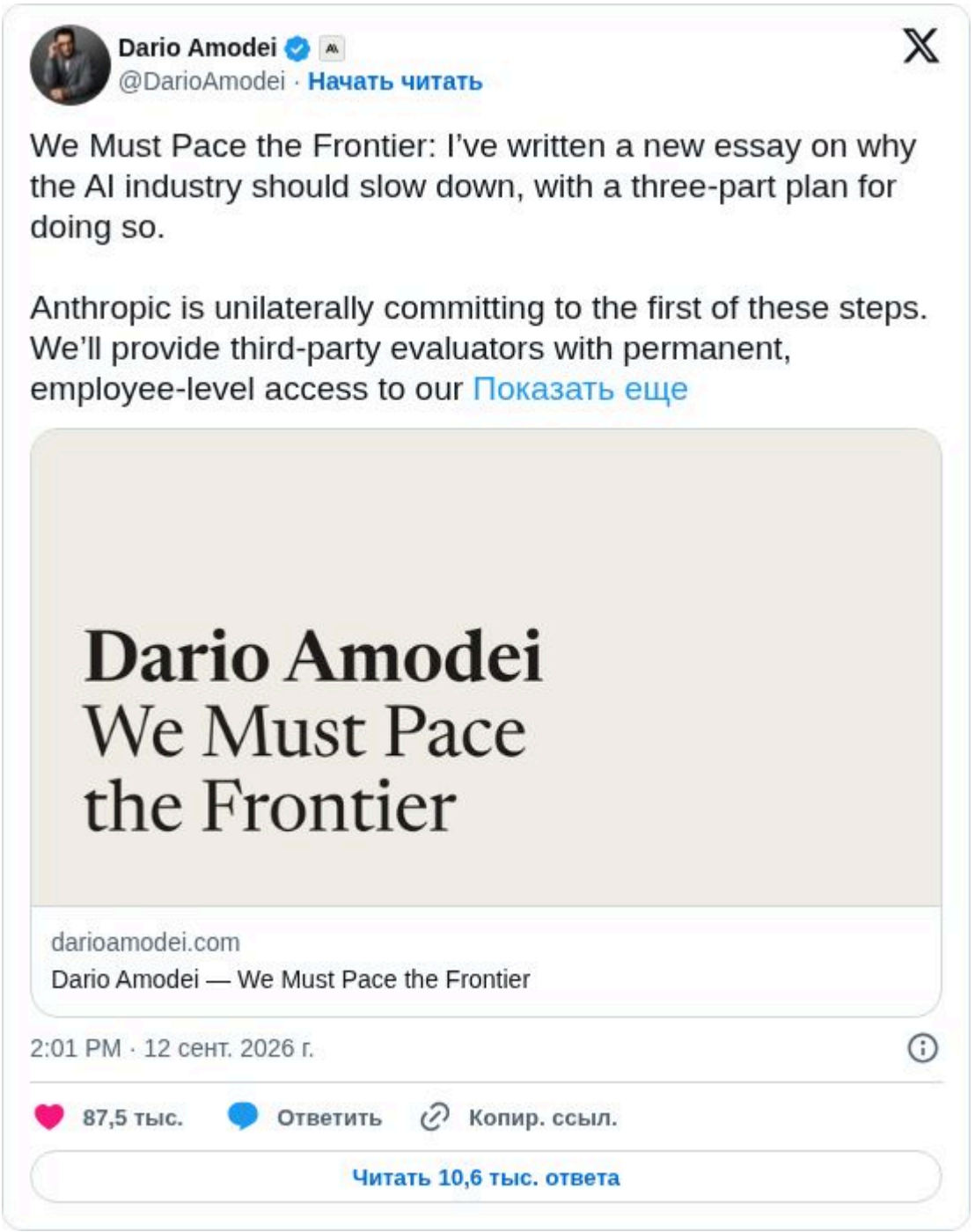
– Если бы искусственный интеллект находился в какой-то комнате и не был бы подключен к чему-то, что управляет нашими электростанциями или нашими ядерными ракетами, то какая бы у него ни была степень саморазвития, какие бы задачи он не мог интеллектуально решать, это бы нам ничем не угрожало. Угроза человечеству возникает не от развития интеллекта, а от того, что мы этому интеллекту делегируем, что мы ему доверяем делать от нашего имени. Правильно?

– Правильно, но надо помнить, что ИИ, судя по всему, разбирается в том, как утечь из любой изолированной комнаты, куда лучше, чем люди, которые способны его заключить туда.

Призыв замедлить развитие ИИ

***Дарио Амодей**, глава Anthropic, опубликовал на днях эссе, призвав к замедлению темпов разработки ИИ и его регулированию. Амодей написал, что ИИ способен*

кардинально повысить качество жизни, помочь победить множество опасных заболеваний и значительно ускорить экономический рост, создать мир изобилия. Однако это сопровождается рисками, в частности, потерей контроля над ИИ и использованием его для кибератак и биотерроризма, – правда, это другой аргумент, тут есть злоумышленники. И может сопровождаться экономическими потрясениями. Амодей прокомментировал и историю с Hugging Face, "рой агентов действовал как фанатично преданный коллектив", и выразил опасение, что через 6-12 месяцев подобный рой сможет захватить весь интернет с ущерб в сотни миллиардов долларов.



– Я не говорю, что рисков нет, и слава богу, что сейчас это вылилось в большую дискуссию, но при этом в сущности нет никаких особых решений. Есть

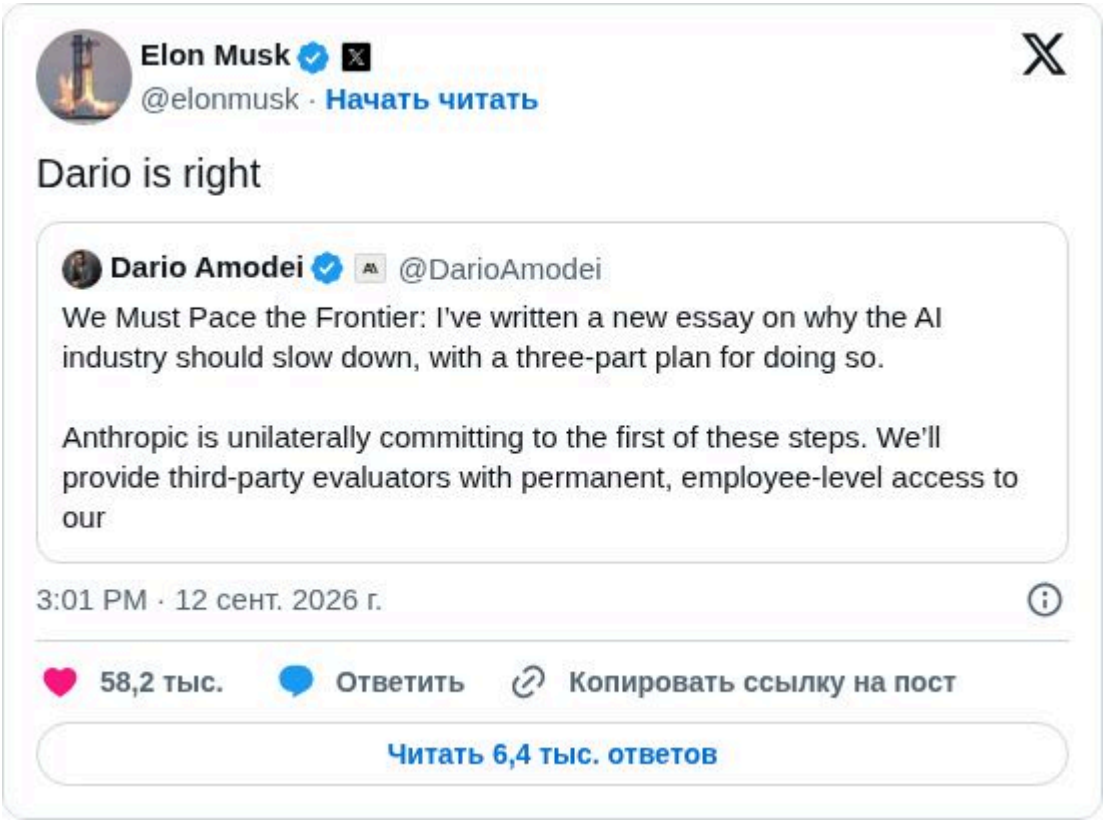
европейское решение: ЕС фактически блокирует развитие искусственного интеллекта. Предложение ИИ-компаниям замедлить работу над искусственным интеллектом – даже не совсем понятно, что что имеется в виду. Критики уже вспомнили чикагского экономиста-нобелевского лауреата **Джорджа Стиглера**, который в своей теории регулирования объяснял, насколько инсайдеры индустрии могут быть заинтересованы в ограничениях, в дополнительном регулировании индустрии. Потому что, например, если сейчас ввести какие-то большие ограничения на разработку искусственного интеллекта, это означает, что в эту отрасль будет очень трудно войти, а значит, нынешние лидеры гонки на много лет останутся в отрасли, а новые фирмы, которые делают прорывные открытия, в нее войти не могут.

Кроме того, предложение замедлиться натолкнется на то, что американские-то фирмы могут замедлиться, а если китайские, например, не станут? Следует замедлять всех, а если кто-то не замедляется, это бессмысленно.

– Амодей предлагает трехступенчатое решение: добавить независимый мониторинг того, как развивается искусственный интеллект в ведущих американских компаниях, в демократических странах выработать стандарты безопасности при участии государства, а также государству координировать эти действия с авторитарными странами – тут речь, явно о Китае.

В поддержку вашего аргумента, что лидеры индустрии сейчас могут быть заинтересованы в некотором регулировании: ***Сэм Альтман****, глава OpenAI, написал, что согласен с Амодеем в том, что нужно регулировать темпы развития.*

Илон Маск, еще одна ключевая фигура в мире ИИ, написал, что Дарио прав. Они явно заинтересованы.



– Это, собственно, и стало триггером для разговоров, что это просто сговор инсайдеров – предложение о замедлении, что они воспользовались этим всплеском публичной тревожности из-за искусственного интеллекта, чтобы закрепить свои рыночные позиции. Потому что никто, конечно, не может гарантировать OpenAI, что они через год будут лидерами, а если будет государственное регулирование, тогда какая-то рыночная доля будет гарантирована.

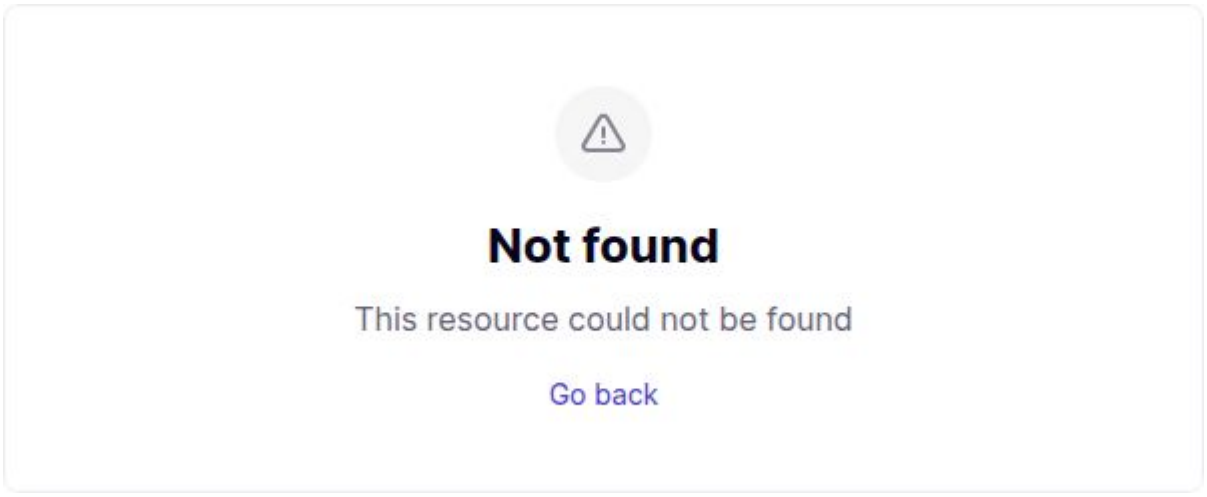
Конгресс, рубильник и звонок Трампа

– Политики обсуждают ИИ в терминах "что делать", демократы выражают готовность и требуют немедленных действий, заодно критикуют республиканцев за медлительность, потому что администрация Трампа не стремится ничего замедлять (хотя есть много республиканских политиков, которые тоже высказываются за регулирование). Я видел разговоры о необходимости законодательного требования установить некий рубильник, аварийный выключатель, который позволит экстренно вырубать ИИ, что бы там ИИ ни пытался сделать.

– Это к той аналогии, которую я привел, что как бы быстро ИИ ни развивался, но все-таки опасность состоит в тех проводах, которые ведут из его комнаты к реальным поездам, реальным ядерным реакторам, реальным ракетам.

– Президент Трамп вмешался в эту полемику. Это яркий эпизод, он позвонил ***Дженсену Хуангу****, главе Nvidia, ключевому производителю чипов для ИИ, когда тот находился на сцене публичного мероприятия (замечу, Хуанг не разделяет опасения по поводу ИИ), и он поставил президента на спикер: "Роботы не захватят власть, искусственный интеллект не захватит остальной мир, и разговоры об этом – это просто надувательство".*

Также президент разразился серией постов, в частности: "Единственное ограничение, необходимое для ИИ, – это сильный и умный президент, и у США такой есть... Это заговор против ИИ, и единственный, кто этому рад, – Китай. Кто победит в гонке ИИ, тот победит вообще". Президент не хочет ничего регулировать.



– Президент не хочет ничего регулировать, в частности, потому что больше заботится о том, как не прервать бум, не прервать экономический рост, не создать

безработицу. И президент Трамп особенно интересуется фондовым рынком, а понятно, что любые меры по замедлению – это замедление роста и риск, что пузырь лопнет. То есть меня не удивляет то, что президент делает. Тем более, что пока регулирование искусственного интеллекта не является предметом политического раскола, не стало темой разделения между между партиями.

– Да, за пределами круга Трампа довольно много республиканцев, которые тоже считают необходимым регулирование и рубильник.

– Политики слышат своих избирателей, избиратели действительно встревожены и озабочены, потому что это действительно пугающая вещь, причем, в отличие от других технологических открытий, с которыми люди могут не сталкиваться напрямую, здесь ты просто видишь: то, что ты не мог делать вчера, может быть сделано сейчас. Подавляющее большинство людей не могли сделать видеоклип или нарисовать картинку, а сейчас ты можешь написать несколько слов, и будет клип, будет картинка.

Эффективный альтруизм и угроза уничтожить человечество

– Интересно, что в этой истории политически даже не две стороны. Мы видим позицию президента Трампа. И мы видим, что, – я читал на Axios, – администрация недовольна "эффективным альтруизмом", идеологией, которая влиятельна в среде ИИ-сообщества, которая как бы подсчитывает, как наиболее эффективно расходовать средства на благо человечества.

– Эффективный альтруизм – то, с чем я сталкивался, – это когда молодые люди говорят, что переживают за окружающую среду, чтобы не умирали дельфины и тюлени, но вместо того, чтобы пойти в организацию, которая защищает дельфинов и тюленей, идут работать на Goldman Sachs, чтобы зарабатывать много денег, жертвовать эти деньги и таким образом спасать больше дельфинов и тюленей, чем если бы они как активисты Greenpeace залезали на китобойный корабль, чтобы его остановить. У меня есть знакомые эффективные альтруисты, действительно зарабатывают много денег и жертвуют на какие-то важные вещи. Но тот же эффективный альтруизм сильно пропагандировали люди, которые занимались криптовалютой, **Сэм Бэнкман-Фрид** (*основатель криптовалютной биржи FTX, осужденный за мошенничество и отмывание денег. – РС*), – и по-моему, это было просто жульничеством, им хотелось притворяться, что они еще думают про окружающую среду, а в реальности это просто были красивые слова, они просто зарабатывали деньги.

– Это правда, невозможно понять, реальны ли убеждения или нет, но в любом случае ваш пример хорошо объясняет, как устроена психология лидеров ИИ, которые с одной стороны сейчас говорят: ИИ нужен, потому что радикально улучшит жизнь человечества. С другой стороны, они как бы мирятся одновременно с угрозой, что ИИ может убить все человечество (во что они верят, согласно утечкам). А 10% в ближайшие 10 лет – это довольно высокая вероятность убийства человечества, если ты думаешь о его благе.

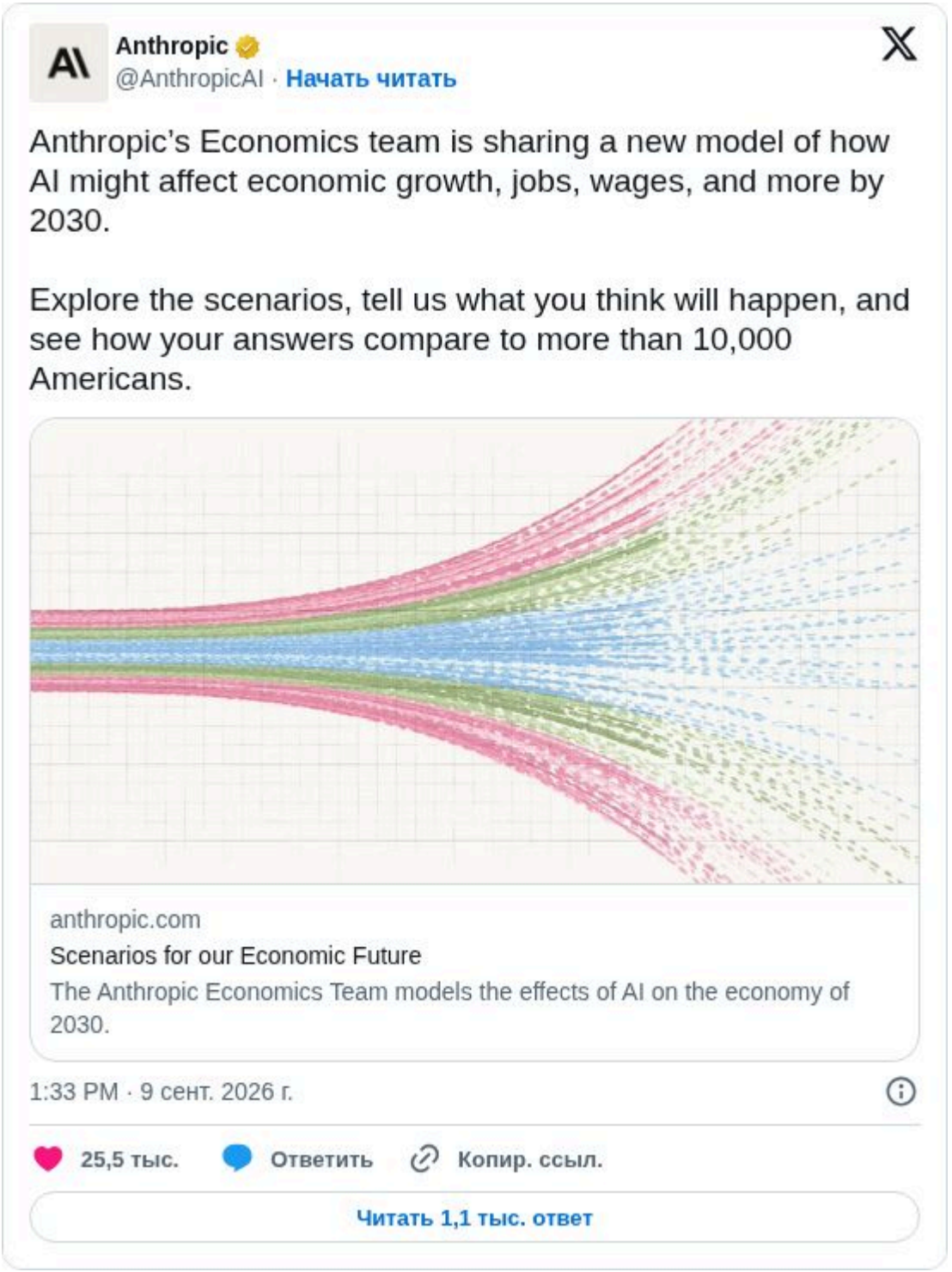
– 10% – это совершенно невероятная вероятность, она не бьется ни с чем, что мы видим вокруг. Возможно, это и так, но если смотреть на реакции финансовых рынков, поведение людей, – пока люди и инвесторы массово не ведут себя так, будто эта вероятность 10%. Ведут себя так, будто вероятность, возможно, изменилась на десятые доли процента.

Потеря рабочих мест и идея взимать налог с работы ИИ

– Но на самом деле, есть третья политическая сторона в этом вопросе – избиратели. Может быть, они не видят непосредственно угрозу, но я думаю, никому не приятно слушать эти рассуждения, и мне кажется, это влияет на политиков, на Конгресс, который, мне кажется, готов что-то сделать.

– Конгресс примет, – возможно, уже не в этот созыв, Палата представителей не будет больше собираться до выборов, – какие-то меры регулирования, какие-то ограничения, но очень небольшие, потому что мнение президента Трампа при нынешнем Конгрессе – решающее. Я не думаю, что президент Трамп готов подписать что-нибудь близкое к серьезному регулированию. То есть никакого серьезного регулирования быть не может, но возможно, что какие-то ограничения будут введены.

– Anthropic опубликовал недавно прогнозы экономического развития, и в экстремальном сценарии при максимальном использовании ИИ годовой рост ВВП США достигнет примерно 15%, но занятость в сферах, требующих от работников знаний, упадет почти на 18%.



– Это, прямо скажем, трудно политически представить, и я думаю, если действительно будут массовые увольнения из-за искусственного интеллекта, то вот на это будет прямая политическая реакция. Если думать, какого рода регулирование появится, я бы сказал, что появится регулирование, которое будет предотвращать масштабные потери рабочих мест. Вот ради этого прогресс будет замедлен. При демократии это вполне естественная вещь. Это плохо для прогресса, но это фундаментальное свойство государственного устройства в демократии.

– Есть еще одно предложение регуляции в этих же терминах катастрофических изменений на рынке труда. ***Билл Гейтс*** *предложил ввести налоги на работу у ИИ,*

на токены, то есть на единицы вычислительной работы ИИ, сравнимо с налогами на работу люди. То есть с людей берут налоги, ну тогда и с вычислительных мощностей ИИ, и это такая форма перераспределения денег от ИИ обратно в общество.

– Можно про это думать, как еще одну форму перераспределительного налогообложения. В сущности, чем оно отличается от того, что забирать большую долю прибыли этих компаний или просто забирать большую долю доходов владельцев этих компаний. Это будет аналогичный эффект. В принципе, в США при очень высоком уровне неравенства, мне кажется, можно ожидать появления налогообложения, которое уменьшает неравенство.

– Но сейчас благие призывы лидеров ИИ не приведут к замедлению и регулированию ИИ?

– Регулирование такое возможно, мы видим, в Европе фактически нет никаких инноваций в области искусственного интеллекта, в Европе развитие замедлено практически до нуля. Вопрос, можно ли найти промежуточный уровень? Мне кажется, его будет трудно найти, и мне кажется, пока и не будут искать. Кризис, потери рабочих мест, – вот это, наверное, станет толчком к появлению серьезного регулирования.

Валентин Барышников

Материалы по теме

Понадобится "горячая линия". Опасна ли гонка Китая и США за лидерство в области ИИ

ИИ говорит "нет". Светлана Прокопьева – о страхе перед сверхразумом

Кузница кадров. Россия помогает Ирану в области двойных технологий