ЭЛИЕЗЕР ЮДКОВСКИЙ
НЕЙТ СОАРЕС

ЕСЛИ КТО-ТО ЕГО СОЗДАСТ — ВСЕ ПОГИБНУТ

ELIEZER YUDKOWSKY AND NATE SOARES

IF ANYONE BUILDS IT, EVERYONE DIES

WHY SUPERHUMAN AI WOULD KILL US ALL

ЭЛИЕЗЕР ЮДКОВСКИЙ, НЕЙТ СОАРЕС

ЕСЛИ КТО-ТО ЕГО СОЗДАСТ — ВСЕ ПОГИБНУТ

ПОЧЕМУ СВЕРХЧЕЛОВЕЧЕСКИЙ **ИИ** УНИЧТОЖИТ НАС ВСЕХ

Перевод с английского
Евгения Поникарова

издательство **АСТ**
Москва

УДК 004.8
ББК 32.813
Ю16

Художественное оформление и макет Андрея Бондаренко

Юдковский, Элиезер.
Ю16 Если кто-то его создаст — все погибнут. Почему сверхчеловеческий ИИ уничтожит нас всех / Элиезер Юдковский, Нейт Соарес; пер. с англ. Е. Поникарова. — Москва : Издательство АСТ : CORPUS, 2026. — 256 с.

ISBN 978-5-17-182679-6

В 2023 году сотни видных деятелей в области искусственного интеллекта подписали открытое письмо, предупреждающее о серьезной угрозе для человечества. Но компании и страны все равно спешат создать машины умнее человека. И мир совершенно не готов к тому, что произойдет дальше.
Несколько десятилетий двое специалистов, подписавших то письмо, — Элиезер Юдковский и Нейт Соарес — изучали, как сверхинтеллект будет мыслить и вести себя. Их исследования показывают: конфликт продвинутых ИИ-моделей и человечества неизбежен. И нам не победить. Как и зачем искусственный сверхинтеллект уничтожит наш вид? Что необходимо сделать человечеству для выживания? Авторы подробно, доступно и образно обосновывают свою точку зрения и призывают каждого человека на планете задуматься об экзистенциальной угрозе, нависшей над всеми нами. Пока еще не поздно.

УДК 004.8
ББК 32.813

ISBN 978-5-17-182679-6

Содержание

ЧАСТЬ III
Лицом к лицу с проблемой

Всем людям, умершим в ходе
долгой истории нашего вида,
всем, кто еще жив,
и всем детям, которые когда-нибудь
могут появиться

ВВЕДЕНИЕ

Сложные прогнозы и простые прогнозы

“Снижение риска вымирания человечества из-за искусственного интеллекта должно стать глобальным приоритетом — наравне с пандемиями, ядерной войной и другими угрозами глобального масштаба”.

В начале 2023 года сотни ученых, работающих в сфере искусственного интеллекта, подписали открытое письмо, содержавшее одну эту фразу. Среди тех, кто поставил подпись, были весьма титулованные специалисты в этой области, в частности нобелевский лауреат Джеффри Хинтон и Йошуа Бенджио, получившие премию Тьюринга за создание глубокого обучения.

Мы, Элиезер Юдковский и Нейт Соарес, также подписали это письмо, хотя и сочли его формулировку слишком сдержанной.

Ни нас, ни других подписантов не беспокоят ИИ-модели образца 2023 года. Нас не беспокоят и те модели, что существуют на момент написания этих строк — в начале 2025 года. Современные ИИ-модели все еще кажутся поверхностными в каком-то глубинном смысле, который сложно описать. У них есть ограничения: например, неспособность формировать новые долговременные воспоминания. Эти недостатки пока что не позволяют им проводить серьезные на-

учные исследования или заменять людей на рабочих местах в сколько-нибудь значимом количестве*.

Наше беспокойство вызывает то, что появится дальше: машинный интеллект, который будет по-настоящему разумным — разумнее любого человека и даже всего человечества в целом. Нас тревожит ИИ, который превзойдет человеческую способность мыслить, обобщать опыт, решать научные задачи, изобретать новые технологии, планировать, вырабатывать стратегии, просчитывать ходы и совершенствовать себя. Когда он превзойдет человека почти во всех интеллектуальных задачах, такой ИИ можно будет назвать искусственным сверхинтеллектом.

Пока еще ИИ-модели не достигли этого уровня. Но сегодня они умнее, чем в 2023 году, и значительно умнее, чем в 2019-м. Исследования в области искусственного интеллекта приводили к одному скачку за другим: в 2012, 2016, 2020, 2022 и 2024 годах**. Мы не знаем, выдохнется ли этот прогресс, взяв паузу на какое-то время, пока не появятся новые методы и технологии. Мы не знаем, сколько таких скачков отделяет нас от момента, когда ИИ станет угрозой уровня вымирания, о которой предупреждали авторы открытого письма. Однако история раз за разом показывает: исследователи в области ИИ изобретают новые методы и преодолевают старые препятствия. Зачастую прогресс идет удивительно быстрыми темпами. В 2015 году большинство специалистов по компьютерным наукам сказали бы вам, что собеседник уровня *ChatGPT* появится только лет через тридцать или даже пятьдесят.

* Авторы благоразумно указали, что пишут эти строки в начале 2025 года. Сегодня, буквально через год, оба процесса уже вовсю начинают происходить, хотя проблема долгосрочной памяти у языковых моделей еще не вполне решена. *(Прим. науч. ред.)*

** В 2012 году нейронная сеть *AlexNet* начала справляться с проблемой распознавания изображений. В 2016-м программа *AlphaGo* победила лучшего игрока в го. В 2020-м появилась (чисто предиктивная) языковая модель GPT-3. В 2022-м появился (широко используемый) чат-бот *ChatGPT*. В 2024-м рассуждающие модели начали хорошо писать код и гораздо лучше решать математические задачи и визуальные головоломки.

Мы не знаем, когда появится искусственный сверхинтеллект, но согласны, что эта проблема должна иметь глобальный приоритет. Более того, мы считаем, что вышеуказанное открытое письмо сильно недооценивает серьезность проблемы.

Нас пригласили подписать то письмо из одной фразы, поскольку мы руководим Институтом исследований машинного интеллекта (MIRI). Эта некоммерческая организация работает над вопросами, связанными с машинным сверхинтеллектом, с 2001 года, то есть она начала задаваться этими вопросами задолго до того, как они стали широко обсуждаться и привлекать финансирование. Если говорить упрощенно: среди тех немногих, кто следит за этой темой десятилетиями, принято считать, что именно MIRI работает над ней дольше всех. Один из нас, Юдковский, — основатель MIRI; другой, Соарес, — его нынешний президент.

MIRI — первая организованная группа, которая заявила: "В какой-то момент появится искусственный сверхинтеллект, и это событие представляется чрезвычайно важным. Возможно, будет технически сложно направить работу сверхинтеллекта в такое русло, чтобы он помогал человечеству, а не вредил ему. Не лучше ли приступить к решению этой задачи прямо сейчас, а не ждать, пока ситуация превратится в масштабный кризис?"

Но начинали мы не с этого. В 2000 году Юдковский попытался создать машинный сверхинтеллект. В 2001-м он осознал, что тот не обязательно окажется дружественным. А в 2003-м он понял, что эта проблема сложна.

Первые два десятилетия своего существования MIRI был институтом технических исследований, практически не участвовавшим в политике. Организация в основном проводила семинары для заинтересованных ученых и давала пристанище нескольким многообещающим исследователям. Мы пытались разработать математический аппарат для понимания и формирования сверхчеловеческого машинного интеллекта, а также для предсказания, что может пойти не так.

Деятельность MIRI имела и непрямые последствия, к которым мы теперь относимся неоднозначно или с сожалением. На одной из организованных нами конференций мы познакомили Демиса Хассабиса и Шейна Легга, основателей компании, которая сейчас называется *Google DeepMind*, с первым их крупным инвестором. А Сэм Альтман, генеральный директор *OpenAI*, однажды заявил, что Юдковский "привил многим из нас интерес к ОИИ*" и "сыграл ключевую роль в решении запустить *OpenAI*"**.

История MIRI сложна, но наши взаимоотношения со всей этой сферой можно описать так: за несколько лет до появления нынешних ИИ-компаний любой, кто хотел, невзирая на риск вымирания, создавать по-настоящему умную модель ИИ, должен был сперва отмахнуться от наших предупреждений.

Позже, когда сфера ИИ начала набирать обороты, мы с тревогой наблюдали, как некоторые люди, основывающие новые компании, заговорили об искусственном сверхинтеллекте как об источнике огромной чудесной силы. Силы, которую они, как им казалось, сумеют контролировать. По мнению многих из них, главная опасность заключалась в том, что искусственный сверхинтеллект может "оказаться" не у тех людей. Они рассуждали о необходимости выиграть "гонку вооружений в области ИИ". А вот о том, что искусственный сверхинтеллект "окажется" у самого же искусственного сверхинтеллекта — то есть единственным победителем в гонке вооружений в области ИИ будет сам искусственный сверхинтеллект, — основатели компаний не говорили.

Мы видели, что возможности ИИ росли крайне быстро.

* Термин ОИИ (общий искусственный интеллект) предназначен для того, чтобы отличать ИИ, который интуитивно кажется "действительно умным", от узкоспециализированных ИИ-моделей прошлого. В нашей книге мы избегаем этого термина, поскольку специалисты расходятся во мнениях, что же именно он подразумевает в свете появления ИИ-моделей, подобных *ChatGPT*.

** Если это правда, то произошло это вопреки мнению Юдковского: он считал, что *OpenAI* — ужасная идея.

Мы видели, что та область исследований, где работали мы сами, — ставящая своей целью понять ИИ и, возможно, не дать ситуации пойти по плохому сценарию — развивалась намного, *намного* медленнее.

Безоглядное рвение ИИ-компаний к созданию сверхчеловеческого искусственного интеллекта — их стремление создать его как можно быстрее, опередив конкурентов, — стало выглядеть для нас как гонка по нисходящей. Индустрия неслась под откос: ситуация могла бы войти в учебники в качестве примера, как *не* надо вести разработки (правда, написать такой анализ было бы уже некому).

Мы уже не верили, что человечество сумеет найти выход из катастрофической ситуации с помощью инженерных решений и исследований. Не в таких условиях. Не к нужному сроку.

Мы сочли наши предыдущие усилия провальными, свернули бóльшую часть исследований MIRI и направили ресурсы института на донесение одной-единственной мысли — предупреждения, лежащего в основе этой книги:

> Если какая-либо компания или группа
> на планете создаст искусственный
> сверхинтеллект, используя что-либо,
> даже отдаленно похожее
> на современные технологии,
> и основываясь на понимании ИИ,
> даже отдаленно похожем на нынешнее,
> то погибнут все люди на всей Земле.

Это не гипербола. Мы не преувеличиваем ради эффекта. Мы считаем, что это самая прямая экстраполяция современных знаний, данных и институционального поведения в области искусственного интеллекта.

В этой книге мы излагаем наши доводы в надежде сплотить достаточное число как лиц, принимающих важные ре-

шения, так и обычных людей — чтобы они отнеслись к ИИ серьезно. Исход по умолчанию летален, однако ситуация пока еще не безнадежна: машинного сверхинтеллекта пока не существует, и его создание все еще можно предотвратить.

Как можно быть в чем-то уверенным, когда дело касается искусственного интеллекта? Известный афоризм гласит: "Предсказывать очень трудно, особенно будущее". Бóльшая часть того, что мы хотели бы знать о будущем, на самом деле не поддается прогнозированию. Например, мы не можем назвать вам выигрышные номера лотереи на следующей неделе. Все наборы чисел выглядят одинаково вероятными.

Но некоторые факты о будущем действительно предсказуемы. Предположим, завтра вы купите лотерейный билет. Мы не знаем, какими изощренными теориями или уловками вы воспользуетесь при выборе чисел, и не знаем, какие номера выпадут. Но вся эта неопределенность сводится к весьма надежному прогнозу: скорее всего, вы не выиграете. Аналогично, если бросить кубик льда в стакан с горячей водой, невозможно предсказать, где окажется каждая конкретная молекула десять минут спустя. Однако вся эта неопределенность складывается в прогноз, близкий к достоверному: кубик льда растает. По такому принципу работает существенная часть физики: мы знаем, куда ведут почти все пути, даже если не можем просчитать, какой именно путь будет выбран.

Одни аспекты будущего возможно предсказать, имея нужные знания и прикладывая усилия; другие же предсказать почти невозможно. Компетентная футурология строится на понимании этой разницы.

История учит нас, что относительно легко прогнозировать будущее следующим образом: вы осознаете, что нечто выглядит теоретически возможным в соответствии с законами физики, и предсказываете, что со временем кто-нибудь это сделает. Полеты аппаратов тяжелее воздуха, оружие,

высвобождающее ядерную энергию, ракеты, летящие на Луну с человеком на борту, — все это было предсказано заранее, и на то имелись разумные причины, несмотря на возражения скептиков, которые с умным видом изрекали, что раз этого еще не произошло, то, вероятно, не произойдет никогда. Люди, привязывавшие крылья к рукам и прыгавшие с холмов, выглядели крайне глупо, они калечились и терпели неудачи, их высмеивали современники — однако это не остановило братьев Райт, которые поняли, как нужно летать.

Но куда более сложная задача — предсказать, *когда* именно появится та или иная технология. Люди могут утверждать, что до появления технологии осталось года два, тогда как на самом деле — все пятьдесят, или заявляют про пятьдесят лет, когда на самом деле — всего два года, причем они сами же эту технологию и создадут. "Человек не будет летать еще тысячу лет", — бросил Уилбур Райт Орвиллу в 1901 году, когда братьев измучили неудачи при испытаниях безмоторного планера. Два года спустя, в 1903-м, аэроплан братьев Райт поднялся в воздух.

Успешное прогнозирование заключается не в том, чтобы проявить достаточно ума и предсказать те детали, которые обычно предсказанию не поддаются. И не в том, чтобы придумать полную историю того, что произойдет в будущем, а затем чудесным образом оказаться правым. Дело в поиске тех аспектов будущего, прогнозировать которые просто, если смотреть под правильным углом.

Мы не знаем, когда наступит конец света, если люди и страны ничего не изменят в своем отношении к искусственному интеллекту. Мы не знаем, какие заголовки об ИИ появятся через два года или десять лет, и даже не знаем, есть ли у нас в запасе эти десять лет. Мы не претендуем на звание настолько умных, что якобы можем предсказывать по-настоящему труднопредсказуемые вещи. Нам просто кажется, что один конкретный аспект будущего — что произойдет со всеми людьми и всем, что им дорого, если сверхинтеллект будет создан в ближайшее время, — спрогнозировать вполне