Отзиви за „Ако някой го развие, всички ще умрем“

„Ако някой го развие, всички ще умрем“ представя убедителна аргументация, че свръхчовешкият ИИ почти със сигурност ще доведе до глобално изличаване на хората. Правителствата по света трябва да осъзнаят рисковете и да предприемат колективно и ефективно действие.

Джон Уолфстал, бивш специален помощник на президента
по въпросите на националната сигурност

Юдковски и Соарес обясняват с прости и лесни за разбиране думи защо настоящият ни път към все по-мощни ИИ е изключително опасен.

Емет Шиър, бивш временен изпълнителен директор на *ОупънЕйАй*

Задължително четиво за политици, журналисти, изследователи и за широката общественост. Майсторски написан и новаторски текст, „Ако някой го развие, всички ще умрем“ предоставя важна отправна точка за дискусия по темата за ИИ на всички нива.

Барт Селман, професор по компютърни науки,
университет „Корнел“

Макар да съм скептичен, че настоящата траектория на развитие на изкуствения интелект ще доведе до изчезване на човечеството, признавам, че това мнение може да отразява липса на въображение от моя страна. Предвид експоненциалното темпо на промяна на изкуствения интелект, няма по-добро време да се вземат предпазни мерки срещу най-лошите възможни резултати. Авторите предлагат важни идеи за глобални предпазни мерки и намаляване на риска, които заслужават сериозно внимание.

Генерал-лейтенант Джон Н. Т. Шанахан – Джак
(ВВС на САЩ, пенсиониран), първи директор на Съвместния
център за изкуствен интелект на Министерството на отбраната

„Ако някой го развие, всички ще умрем“ не е просто предупреждение; това е пожарна аларма, която звучи ясно и изисква спешни действия. Юдковски и Соарес не се шегуват: неконтролираният свръхчовешки изкуствен интелект е екзистенциална заплаха. Тя е трезво напомняне, че бъдещето на човечеството зависи от това, което правим в момента.

Марк Руфало

Сериозна книга във всяко отношение. В зловещия анализ на Юдковски и Соарес свръхмощният изкуствен интелект няма да има нужда от човечеството и ще разполага с достатъчно капацитет, за да ни унищожи. „Ако някой го развие, всички ще умрем“ е красноречив и спешен призив да се отдръпнем от ръба на самоунищожението.

Фиона Хил, бивш старши директор, Съвет за национална сигурност на Белия дом

Ясно написан и убедителен разказ за екзистенциалните рискове, които високоразвитият изкуствен интелект може да стане за човечеството. Препоръчвам я.

Бен Бернанке, Нобелов лауреат и бивш председател на Федералния резерв

Вероятно ще затворите тази книга напълно убедени, че правителствата трябва незабавно да преминат към по-предпазлив подход към изкуствения интелект; подход, който в по-голяма степен уважава огромното значение на създаваното в момента за промяната на цивилизацията. Бих искал всеки на Земята, който се интересува от бъдещето, да прочете тази книга и да обсъди идеите в нея.

Скот Аръндсън, ръководител на катедрата по компютърна наука, Тексаски университет, Остин

Невероятно сериозен проблем, който заслужава – всъщност изисква – нашето внимание. Не е необходимо да сте съгласни с прогнозите или предписанията в тази книга, нито да сте запознати с технологиите или изкуствения интелект, за да я възприемете като интересна, достъпна и провокираща мисълта.

Сюзан Сполдинг, бивш заместник-министър, Министерство на вътрешната сигурност

Най-важната книга, която съм чел от години: искам да я дам на всеки политически и корпоративен лидер в света и да стоя над тях, докато я прочетат. Юдковски и Соарес отправят силен призив към човечеството, за да ни събудят, докато ние сънливо вървим към катастрофата.

Стивън Фрай

Най-важната книга на десетилетието.

Макс Тегмарк, професор по физика, МТИ

Твърденията за рисковете от изкуствения интелект често се отхвърлят като реклама, целяща да продаде повече устройства. Щеше да е утешително, ако беше вярно, но тази книга опровергава подобна теория. Юдковски и Соарес не са от индустрията за изкуствен интелект и пишат за тези рискове още отпреди изкуственият интелект да съществува в сегашната си форма. Прочетете тяхната тревожна книга и ни кажете в какво грешат.

Хю Прайс, почетен професор по философия „Бъртранд Ръсел“,
Тринити Колидж, Кеймбридж, Великобритания

Всеки трябва да прочете тази книга. Седемдесет процента съм сигурен, че вие – да, вие, които четете това в момента – един ден неохотно ще признаете, че всички ние трябваше да послушаме Юдковски и Соарес, когато все още имахме възможност.

Даниъл Кокотайло, вътрешен информатор на *Оупън Ей Ай*
и изпълнителен директор на „Проект ИИ бъдеще“

„Ако някой го развие, всички ще умрем“ може да се окаже най-важната книга на нашето време. Юдковски и Соарес вярват, че сме далече от това да направим безопасно прехода към свръхинтелект, което ни оставя на бързата писта към изчезването. Като използват притчи и кристално ясни обяснения, те предават своите аргументи в спешен призив да се спасим – докато все още можем.

Тим Ърбан, създател, *Wait But Why*

Това провокативно произведение е сурово и спешно предупреждение, предадено с достоверност, яснота и убеденост, което предизвиква технолози, политици и граждани да се изправят пред екзистенциалните рискове на изкуствения интелект, преди да е станало твърде късно. Задължително четиво за всеки, който се интересува от бъдещето.

Ема Скай, старши научен сътрудник, Висше училище
за глобални въпроси „Джаксън“, Йейл

Тази книга предлага блестящи прозрения за най-значимото противопоставяне в историята между технологичната утопия и антиутопията. Тя показва как можем и трябва да предотвратим свръхчовешкия изкуствен интелект да убие всички ни.

Джордж Чърч, основател и главен преподавател,
институт „Уайс“, Харвард

Трезва, но изключително четивна книга за реалните рискове от изкуствения интелект. И скептиците, и вярващите трябва да разберат аргументите на авторите и да работят за това бъдещето на изкуствения интелект да е по-скоро полезно, отколкото вредно.

Брус Шнайер, автор на *A Hacker's Mind*

Това е нашето предупреждение. Прочетете го днес; утре го разпространете. Изискайте предпазни мерки. Аз ще продължа да залагам на човечеството, но първо трябва да се събудим.

Р. П. Еди, бивш директор на Съвета
за национална сигурност на Белия дом

Убедително въведение в най-важната тема в света. Свръхчовешкият изкуствен интелект може да се появи след няколко години. Тази книга разглежда сериозно последствията и обяснява без заобикалки какво ни очаква.

Скот Александър, създател на *Astral Codex Ten*

Ще изпитате истински емоции, когато прочетете тази книга. В момента живеем в последния период от историята, в който сме доминиращият вид. Хората са щастливи, че имат Юдковски и Соарес на своя страна, които ни напомнят да не пропиляваме краткия прозорец, в който трябва да вземем решения за бъдещето си.

„Граймс“

Най-доброто, ясно и просто обяснение на проблема с риска от изкуствения интелект, което съм чел.

Ишан Уонг, бивш главен изпълнителен
директор на „Редит“

Елиезер Юдковски
Нейт Соарес

АКО НЯКОЙ ГО РАЗВИЕ, ВСИЧКИ ЩЕ УМРЕМ

Защо свръхчовешкият изкуствен интелект
може да изличи всички ни

София, 2025

Преводът е направен по изданието:
Eliezer Yudkowsky and Nate Soares
If Anyone Builds It, Everyone Dies
Why Superhuman AI Would Kill Us All

ISBN 978-619-01-1696-7

Елиезер Юдковски
Нейт Соарес

АКО НЯКОЙ ГО РАЗВИЕ, ВСИЧКИ ЩЕ УМРЕМ

ЗАЩО СВРЪХЧОВЕШКИЯТ ИЗКУСТВЕН ИНТЕЛЕКТ МОЖЕ ДА ИЗЛИЧИ ВСИЧКИ НИ

Превод от английски
Людмила Андреева

КАТЕХОН

КАТЕХОН

...

№28 Артур Кордон — *Перспективата изкуствен интелект*
№29 Боян Чуков — *Матрицата 2*
№30 Тома Пикети — *Кратка история на равенството*
№31 Кирил Фурсов — *Държавата търговец*
№32 Майкъл Хъдсън — *Съдбата на цивилизацията*
№33 Халфорд Макиндър — *Географската ос на историята*
№34 Сергей Савелиев — *Нищетата на мозъка*
№35 Нийл Фъргюсън — *Големият упадък*
№36 Гунар Кайзер — *Култът*
№37 Янис Варуфакис — *Технофеодализъм*
№38 Игор Шнуренко — *Да убиеш Левиатан*
№39 Джон Мийършаймър — *Голямата заблуда*
№40 Жак Елюл — *Пропагандата*
№41 Съст. В. Аверянов — *Голямото нулиране*
№42 Олдъс Хъксли — *Обратно към Прекрасния нов свят*
№43 Еманюел Тод — *Поражението на Запада*
№44 Уолтър Липман — *Общественото мнение*
№45 Сергей Савелиев — *Еволюцията на мозъка и социалният подбор*
№46 Робърт Ф. Кенеди-мл. — *Ухан: какво остана покрито*
№47 Шеймъс Брунър — *Лицата на контролигархията*
№48 Питър Турчин — *Последни времена*
№49 Нейоми Клайн — *Двойникът*
№50 Иво Христов — *Отвъд точката на пречупване*
№51 Каръл Куигли — *Трагедия и надежда. Том 1*
№52 Жак Елюл — *Технологията. Залогът на века*
№53 Петър Волгин — *Европа след края на евроатлантизма*
№54 Филип дьо Вилие — *Дръпнах нишката на лъжата и всичко излезе наяве*
№55 Джулио Тремонти — *Криминогенната държава*
№56 Джовани Ариги — *Дългият двайсети век*
№57 Ерик Фьогелин — *Новата наука за политиката*
№58 Каръл Куигли — *Трагедия и надежда. Том 2*
№59 Е. Юдковски, Н. Ноарес — *Ако някой го развие, всички ще умрем*

Съдържание

ЧАСТ III
Изправяне пред предизвикателството

На всички хора, които са умрели в процеса
на развитието на нашия биологичен вид,
на всички, които все още са сред живите,
и на всички деца, които някой ден биха
могли да съществуват

Въведение

Трудни и лесни задачи

Намаляването на риска от изчезване в резултат на изкуствения интелект трябва да бъде глобален приоритет наред с други рискове от обществен мащаб, като пандемии и ядрена война.

В началото на 2023 г. стотици учени в областта на изкуствения интелект подписаха отворено писмо, състоящо се от това изречение. Сред тях бяха някои от най-уважаваните изследователи в тази област – Нобеловият лауреат Джефри Хинтън и Йошуа Бенджио, които споделиха наградата „Тюринг“ за изобретяването на дълбокото обучение.

Ние – Елиезер Юдковски и Нейт Соарес – също подписахме писмото, въпреки че го сметнахме за сериозно подценяващо.

Не изкуственият интелект от 2023 г. беше това, което тревожеше всички нас, подписалите. Нито пък се тревожим за изкуствения интелект, който съществува в момента, докато пишем този текст в началото на 2025 г. Днешните ИИ все още изглеждат плитки в някакъв дълбок смисъл, който е труден за описване. Те имат ограничения, например неспособност да формират нови дълговременни спомени. Тези недостатъци са достатъчни, за да попречат на днешните ИИ да извършват съществени научни изследвания или да заменят толкова много човешки работни места.

Нашето притеснение е за онова, което ще дойде после: машинен интелект, който е наистина умен, по-умен от всеки жив човек, по-умен от човечеството като цяло. Ние сме загрижени за изкуствения интелект, който надминава възможностите на човека да мисли, да прави обобщения въз основа на опита, да решава научни загадки и да изобретява нови технологии, да пла-

нира, да разработва стратегии и да очертава, да размишлява и да се усъвършенства. Можем да наречем такъв ИИ „изкуствен свръхинтелект“ (ИСИ), след като надмине всеки човек в почти всяка умствена задача.

ИИ все още не е стигнал дотам. ИИ обаче днес е по-умен, отколкото беше през 2023 г., и много по-умен, отколкото беше през 2019 г. Изследванията в областта на ИИ доведоха до скок след скок в способностите на ИИ през 2012[A], 2016[B], 2020[C], 2022[D] и 2024 г.[E]

Не знаем дали напредъкът ще спре, което ще доведе до прекратяване на тези скокове за известно време, докато не бъдат изобретени нови методи и технологии. Не знаем колко скока остават, преди изкуственият интелект да стане заплаха за изчезването ни, за която предупреждават подписалите писмото. Историята обаче е показала многократно, че изследователите в областта на изкуствения интелект откриват нови методи и преодоляват стари препятствия. Прогресът често е изненадващо бърз. Повечето компютърни учени през 2015 г. биха ви казали, че изкуственият разговор на нивото на *ЧатДжиПиТи* няма да бъде постижим още тридесет или петдесет години.

Не знаехме кога ще се появи изкуственият свръхинтелект, но бяхме съгласни, че той трябва да бъде глобален приоритет. Всъщност смятаме, че отвореното писмо драстично подценява проблема.

• • •

Бяхме поканени да подпишем това отворено писмо от едно изречение в качеството ни на съпредседатели на *Machine*

[A] През 2012 г. „АлексНет“ реши проблема с разпознаването на обекти в изображения. [Всички бележки под линия, без изрично отбелязаните, са на авторите.]

[B] „АлфаГо“ победи най-добрия играч на го през 2016 г.

[C] (Изцяло прогнозиращият) езиков модел GPT-3 беше пуснат през 2020 г.

[D] (Широкоприложимият) *ЧатДжиПиТи* се появи през 2022 г.

[E] През 2024 г. разсъждаващите модели започнаха да решават математически, програмни и зрителни пъзели.

Intelligence Research Institute (MIRI, Институт за изследване на машинния интелект) – институт с нестопанска цел. MIRI работи по въпроси, свързани с машинния свръхинтелект, от 2001 г. – много преди тези проблеми да получат голяма публичност или финансиране. За да опростим нещата: сред малкото, които следят този въпрос от десетилетия, MIRI е признат за институцията, която работи по него най-дълго. Единият от нас – Юдковски, е основател на MIRI, а другият – Соарес, е настоящият му президент.

MIRI беше първата организирана група, която заяви: „Предвидимо е, че свръхинтелигентен ИИ ще бъде разработен в някакъв момент, и това изглежда изключително важно събитие. Може да е технически трудно да се композира свръхинтелектът така, че да помага на човечеството, вместо да ни вреди. Не трябва ли някой веднага да започне да работи по това предизвикателство, вместо да чакаме всичко да се превърне в мащабна извънредна ситуация по-късно?“

Не започнахме с изричането на това. Юдковски започна с опити да създаде машинен свръхинтелект през 2000 г. През 2001 г. обаче той осъзна, че машинният свръхинтелект няма непременно да се окаже приятелски. А през 2003 г. разбра, че проблемът ще бъде сериозен.

През първите две десетилетия от съществуването си MIRI беше технически изследователски институт без особено участие в политиката. Организацията приютяваше предимно семинари за заинтересовани учени и привличаше няколко обещаващи изследователи. Опитвахме се да се ориентираме в математиката, необходима за разбирането и формирането на свръхчовешкия машинен интелект, както и за прогнозирането как това може да се обърка.

MIRI имаше и някои ефекти надолу по веригата, към които сега се отнасяме противоречиво или със съжаление. Организирахме конференция, на която представихме Демис Хасабис и Шейн Лег – основателите на това, което по-късно щеше да стане *Дийп Майнд* на „Гугъл“, на първата им голяма финансираща организация. А Сам Олтман – главен изпълнителен директор на *Оупън Ей Ай*, веднъж заяви, че Юдковски „е накарал много от

нас да се заинтересуват от изкуствения общ интелект (AGI)[A]“ и „е бил от критично важно значение за решението да стартираме *Оупън Ей Ай*“[B].

Историята на MIRI е сложна, но един от начините да обобщим нашата връзка с по-широката област може би е следният: години преди да съществуват сегашните компании за изкуствен интелект, беше смятано, че предупрежденията на MIRI трябва да се отхвърлят, ако искаш да работиш по създаването на истински умен изкуствен интелект въпреки рисковете от изчезване.

Неотдавна, с началото на разцвета на ИИ, наблюдавахме с тревога как някои от по-новите хора, които създаваха ИИ компании, започнаха да говорят за изкуствен свръхинтелект като източник на огромни, изключителни сили. Сили, които те приеха, че ще контролират. Според много от тези основатели главната опасност беше, че неподходящи хора може да „притежават“ ИСИ. Те говореха за необходимостта да се спечели „надпревара във въоръжаването в ИИ“. Що се отнася до възможността не човек да „има“ ИСИ, а ИСИ да притежава ИСИ – т.е. единственият победител в надпреварата във въоръжаването с ИИ да бъде самият ИСИ: е, тези основатели не говореха за това.

Видяхме, че възможностите на ИИ растяха много бързо.

Видяхме, че областта на изследванията, в която бяхме ангажирани, насочена към разбирането на ИИ и към това да не тръгне по лош път, – напредваше много, *много* по-бавно.

Главоломната надпревара на компаниите за ИИ към свръхчовешки ИИ – усилията им да го създадат възможно най-бързо, преди конкурентите им да го направят – започна да ни изглежда като надпревара към дъното. Индустрията се носеше към катастрофа, и то катастрофа, която би влязла в учебниците като при-

---

[A] AGI е съкращение от *Artificial General Intelligence* (изкуствен общ интелект) – термин, който се използва, за да се разграничи изкуственият интелект, който интуитивно е наистина умен, от едноцелевите видове изкуствен интелект от миналото. В тази книга избягваме да използваме този термин, защото хората имат много различни мнения за значението му в контекста на изкуствен интелект като *ЧатДжиПиТи*.

[B] Ако е вярно, това е въпреки възражението на Юдковски, че *Оупън Ей Ай* е ужасна, ужасна идея.

мер за това как *не* трябва да се занимаваш с инженерство. Само че никой нямаше да остане жив, за да напише анализа.

Вече не ни се струваше реалистично човечеството да може да се измъкне от катастрофата чрез проектиране и изследвания. Не при такива условия. Не навреме.

Отписахме предишните си усилия като неуспешни, прекратихме по-голямата част от изследванията на MIRI и пренасочихме фокуса на института към това да предадем предупреждението, което е в ядрото на тази книга:

Ако някоя компания или група, където и да е на планетата, създаде изкуствен свръхинтелект, използвайки нещо, което дори далечно прилича на сегашните техники, основано на нещо, което дори далечно прилича на сегашното разбиране за ИИ, тогава всички навсякъде по Земята ще умрат.

Не го казваме като хипербола. Не преувеличаваме за ефект. Смятаме, че това е най-пряката екстраполация от знанията, доказателствата и институционалното поведение около изкуствения интелект днес.

В тази книга излагаме аргументите си с надеждата да мобилизираме достатъчно ключови лица, вземащи решения, и обикновени хора, за да вземат ИИ на сериозно. Резултатът по подразбиране е смъртоносен, но ситуацията не е безнадеждна; машинен свръхинтелект все още не съществува и създаването му все още може да бъде предотвратено.

• • •

Как може някой да бъде сигурен какво ще се случи по отношение на изкуствения интелект? „Прогнозирането е много трудно, особено на бъдещето“ – гласи афоризмът. Повечето от нещата, които бихме искали да знаем за бъдещето, всъщност не са предсказуеми. Например не можем да ви кажем печелившите числа от лотарията за следващата седмица. Една комбинация от числа изглежда толкова вероятна, колкото и всяка друга.

Някои факти за бъдещето обаче *са* предсказуеми. Ако утре купите фиш от тотото, не знаем какви сложни теории или хрумвания ще използвате, за да изберете числата си, не знаем и кои числа ще бъдат изтеглени, но цялата тази неопределеност води

до много категорична прогноза, че няма да спечелите от тото-то. Същото е и ако пуснете кубче лед в чаша с гореща вода – неимоверно сложно е да се прогнозира къде ще се озове всяка молекула десет минути по-късно, но цялата тази несигурност води до почти сигурна прогноза, че кубчето лед ще се стопи. Половината от физиката е такава: не можем да изчислим коя точно пътека ще бъде избрана, но знаем къде водят почти всички пътеки.

Някои аспекти на бъдещето са предсказуеми с подходящите знания и усилия, а други са невъзможни за прогнозиране. Компетентният футуризъм се гради върху познаването на разликата.

Историята ни учи, че една от относително лесните прогнози за бъдещето е да осъзнаем, че нещо изглежда теоретично възможно според законите на физиката, и да предвидим, че в крайна сметка някой ще го направи. Полет с по-тежки от въздуха летателни апарати, оръжия, които освобождават ядрена енергия, ракети, които летят до Луната с човек на борда: тези събития са били прогнозирани предварително и с основателни причини въпреки съпротивата на скептиците, които мъдро са отбелязвали, че тези неща все още не са се случили и следователно вероятно никога няма да се случат. Хората, които са завързвали криле към ръцете си и са скачали от хълмове, са били смятани за безумци, били са осмивани от съвременниците си и наистина са се наранявали и са се проваляли... но това не е попречило на братята Райт да измислят как да летят.

От друга страна, прогнозирането точно *кога* една технология ще бъде разработена се е оказало много по-трудно. Хората казват, че дадена технология ще се реализира след две години, а всъщност са нужни петдесет години, или казват, че са необходими петдесет години, а в действителност само след две години те самите ще създадат тази технология. „Човекът няма да лети още хиляда години“ – казал Уилбър Райт на Орвил Райт през 1901 г., ядосан от безмоторния планер, който изпитвали по онова време. Две години по-късно, през 1903 г., братята Райт полетяват.

Успешното прогнозиране не означава да си достатъчно умен, за да предскажеш подробности, които обикновено не могат да

бъдат предсказани. Не става въпрос да измислиш цяла история за това, което ще се случи, и после по магичен начин да се окажеш прав. По-скоро става въпрос да намериш аспекти на бъдещето, които стават лесни за прогнозиране, когато се погледнат от правилния ъгъл.

Не знаем кога ще свърши светът, ако хората и страните не променят нищо в начина, по който се отнасят към изкуствения интелект. Не знаем как ще изглеждат заглавията за ИИ след две или десет години, нито дори дали ни остават десет години. Не твърдим, че сме толкова умни, че да можем да предвидим неща, които са трудни за предвиждане. По-скоро ни се струва, че един конкретен аспект на бъдещето: „Какво ще се случи с всички и всичко, за което ни е грижа, ако в близко бъдеще бъде създаден свръхинтелект?“, с достатъчно базисни познания и внимателно разсъждение може да бъде лесна задача.

На пръв поглед изчезването на човечеството от свръхчовешки ИИ може да не изглежда лесно предсказуемо. За това обаче е останалата част от тази книга. Точно както са необходими някои аритметични изчисления, за да се изчисли шансът за печалба от тотото, точно както са необходими някои идеи от термодинамиката, за да се обясни защо леденото кубче предсказуемо се топи, така е необходима и базисна подготовка, за да се разбере защо изкуственият интелект е непосредствена заплаха за изчезване на човечеството. След като тези основи са поставени обаче, прогнозирането на резултата от настоящата ни траектория на движение започва да изглежда мрачно и ужасно елементарно.

• • •

Дори пред лицето на свръхчовешкия машинен интелект може да е изкушаващо да си представим, че светът ще продължи да изглежда така, както през последните няколко десетилетия от нашия относително кратък живот. Вярно е, но е трудно да си спомним, че е имало време, толкова реално, колкото нашето, само преди няколко кратки века, когато цивилизацията е била коренно различна. Или преди хилядолетия, когато не е имало цивилизация, за която да се говори. Или преди милион години,

когато не е имало хора. Или преди милиард години, когато многоклетъчните колонии не са имали специализирани клетки.

Приемането на историческа перспектива може да ни помогне да оценим това, което е толкова трудно да видим от гледна точка на нашия кратък живот: природата допуска смущения. Природата допуска бедствия. Природата допуска светът никога да не е повече същият.

Преди 2,5 млрд. години се случило събитие, което биолозите наричат „кислородната катастрофа": нова форма на живот се е научила да използва енергията на слънчевата светлина, за да извлича ценен въглерод от въздуха. Тази форма на живот издишва опасно токсично и реактивно химическо вещество като отпадък, отровен за повечето съществуващи форми на живот: химическо вещество, което днес наричаме „кислород". Той започва да се натрупва в атмосферата. Повечето форми на живот – включително повечето бактерии, издишващи този кислород – не могат да се справят с неговата реактивност и умират. Няколко щастливи клетки са се адаптирали и в крайна сметка са еволюирали в организми, които използват кислород като гориво. Нещата обаче никога не са се върнали към старото нормално състояние. Светът никога повече не е бил същият.

Някога континентите са били безплодни скали. После, за един миг от еволюционна гледна точка, те са се покрили с растителност. Скоро след това горите са закипели от живот. Светът никога повече не е бил същият.

Някога някои хора са култивирали пшеницата и ечемика. За една съвсем малка частица от еволюционното мигване на окото те са започнали да изграждат цивилизации. Светът никога повече не е бил същият.

Някога, през 30-те години на XX в., е имало предупредителни знаци, че някои семейства вече няма да са в безопасност в Германия. Малко от тях са напуснали страната рано, но повечето са останали. След това нацисткото правителство е отнело гражданството и паспортите им, което е направило бъдещото бягство много по-трудно. Няколко години по-късно германските евреи, роми и други са били арестувани и изпращани в лагери на смъртта. Според разказите на оцелелите много от тези семейства са

останали не защото не са видели предупредителните знаци, а защото са вярвали, че животът ще се върне към нормалното[1], преди нещата да стигнат твърде далеч.

Някога човечеството е било на прага на създаването на изкуствен свръхинтелект...

Нормалността винаги свършва. Това не означава, че тя неизбежно се замества от нещо по-лошо; понякога е така, друг път – не, а понякога зависи от това как действаме. Да се осланяме обаче на надеждата, че нищо лошо няма да бъде позволено да се случи, обикновено не помага.

Хората имат способността да направляват бъдещето с помощта на интелекта си. Тази способност обаче действа само *ако я използваме* – ако правим нещата, които трябва да направим, когато трябва да ги направим. Интелектът няма сила извън това. Той работи, като променя действията ни – или изобщо не работи.

• • •

Предстоящите месеци и години ще бъдат тест на живот и смърт за цялото човечество. С тази книга се надяваме да вдъхновим отделни хора и страни да се изправят пред предизвикателството.

В следващите глави ще очертаем научната основа на нашата тревога, ще обсъдим изкривените стимули, които действат в днешната индустрия за изкуствен интелект, и ще обясним защо ситуацията е дори по-тежка, отколкото изглежда. Ще критикуваме съвременното машинно обучение с прости думи и ще опишем как и защо настоящите методи са напълно неподходящи за създаването на ИИ, които подобряват света, а не го унищожават.

В **част I** на тази книга излагаме проблема, отговаряйки на въпроси като: Що е интелект? Как се произвеждат съвременните ИИ и защо са толкова трудни за разбиране? Може ли изкуственият интелект да има желания? Ще има ли? Ако да, *какво* ще иска и защо ще иска да ни убие? Как ще ни убие? В крайна сметка прогнозираме, че изкуственият интелект няма да ни мрази, но ще има странни, необичайни, чужди предпочитания, които ще преследва, докато човечеството не изчезне.

В **част II** събираме всички тези въпроси, за да разкажем историята за изкуствен интелект, който унищожава свят, много

подобен на нашия. Тази история не е предсказание, защото точният път, по който ще поеме бъдещето, е труден за прогнозиране. Единствената част от историята, която е прогноза, е нейният край – и тази прогноза важи само ако на подобна история бъде позволено да започне.

В **част III** оценяваме трудността на предизвикателството, пред което е изправено човечеството, и преглеждаме отговорите до момента. Колко добре се справят компаниите за изкуствен интелект с проблема? Защо светът не обръща повече внимание? Какво би могло да направи обществото по различен начин, ако достатъчно много от нас решат да *не* умрат? Какво ще е необходимо, за да *не* се изгради машинен свръхинтелект?

Онлайн приложение към тази книга е достъпно на уебсайта IfAnyoneBuildsIt.com. В края на всяка глава ще намерите URL адрес и QR код, който ви препраща към приложението за съответната глава. То ще изглежда така:

IfAnyoneBuildsIt.com/intro

Хората имат всякакви противоречиви интуиции за изкуствения интелект и през годините сме чували най-различни въпроси и възражения, произтичащи от широк спектър от предпоставки и гледни точки. В нашите допълнителни материали разглеждаме повече предупреждения, тънкости и често задавани въпроси, заедно с някои от принципните теоретични основи и разширени аргументи, които щяха да направят тази книга няколко пъти по-дебела и много по-малко достъпна. Ако в края на някоя глава ви хрумне възражение, ви насърчаваме да продължите да четете онлайн.

Много от главите започват с притчи: истории, които, надяваме се, ще помогнат да предадем някои идеи по-просто, отколкото е възможно по друг начин. Те вероятно ще добавят и малко лекота към една иначе тежка тема. Това е в духа на най-древната традиция, може би по-стара от човешкия вид в сегашния му вид, да се смеем в лицето на смъртта.

• • •

Признаваме, че тази книга не е пълна с добри новини. С това обаче не искаме да ви кажем, че сте обречени. Изкуственият свръхинтелект все още не съществува. Човечеството все още може да реши да не го построи.

През 50-те години на миналия век много хора очакваха да избухне ядрена война между големите световни сили. Предвид историята на човешките конфликти до този момент имаше основания човек да е песимист. Въпреки това и до днес не е избухнала ядрена война. Това не е, защото ядрените бомби са се оказали чиста научна фантастика, която никога не би могла да се случи в реалния живот, а защото хората са работили усилено, за да изградят устойчиви системи, които да предотвратят избухването на ядрени войни. Те са направили всичко това, защото световните лидери са знаели, че в случай на ядрена война както те, така и хората в техните страни ще имат лош ден.

Те биха имали лош ден и ако някой някъде по Земята създаде машинен свръхинтелект. Не е в *ничий* интерес да умре заедно със семейството и приятелите си, със страната си и нейните деца.

Спирането на продължаващата ескалация на технологията на изкуствения интелект, ограничаването на хардуера, използван за създаване на все по-мощни модели за изкуствен интелект няма да е *лесно* да се направи в днешния свят. Ще са нужни обаче много по-малко усилия, за да се спре по-нататъшното ескалиране на възможностите на ИИ, отколкото например за воденето на Втората световна война. За да се събуди волята за живот, е необходимо само някои страни, лидери и избиратели да осъзнаят, че се намират на трудна за преценяване, вероятно много кратка дистанция от ръба на смъртта.

Работата няма да е лесна, но все още не сме мъртви. Човешкото достойнство и достойнството на човечеството изискват да се борим.

Където има живот, има надежда.

ЧАСТ I

Нечовешки умове

1

Специалната сила на човечеството

Представете си – макар че, разбира се, нищо подобно не се е случвало, тъй като това е само притча, – че биологичният живот на Земята е резултат от игра между богове. Че е имало бог тигър, който е създал тигрите, и бог секвоя, който е създал секвоите. Представете си, че е имало богове за видовете риби и видовете бактерии. Представете си, че тези играчи са се състезавали за господство на семейството от видове, на което са били попечители, докато формите на живот са бродили по планетата под тях.

Представете си, че преди около два милиона години неизвестен бог маймуна погледнал огромната си игрална дъска с размерите на планетата.

– Ще ми трябват още няколко хода – казал богът хоминид, – но мисля, че вече съм спечелил тази игра.

Настъпило объркване и тишина, докато много богове погледнали игралната дъска, опитвайки се да разберат какво са пропуснали.

Богът скорпион казал:

– Как? Твоето хоминидно семейство няма броня, няма нокти, няма отрова.

– Но има мозък – казал богът хоминид.

– Аз ги заразявам и те умират – казал богът едра шарка[A].

[A] Не е известно едрата шарка да е съществувала повече от няколко хиляди години. Заради художествената свобода богът на едрата шарка символизира факта, че древните хора са умирали от вируси, а съвременните хора имат силата да елиминират ужасните вируси, когато решат.

– Засега – отговорил богът хоминид. – Краят ти ще дойде бързо, едра шарко, щом мозъкът им се научи да се бори с теб.

– Те дори нямат най-големите мозъци! – възкликнал богът кит.

– Не е само въпрос на размер – отговорил богът хоминид. – *Строежът* на мозъка им също има значение. Дай им два милиона години и те ще стъпят на луната на своята планета.

– *Наистина* не виждам къде се произвежда ракетното гориво в метаболизма на това същество – казал богът секвоя. – Не можеш да излезеш в орбита, като просто *мислиш*. В даден момент твоят вид трябва да развие метаболизъм, който пречиства ракетното гориво. Освен това трябва да стане доста голям, идеално висок и тесен – с твърда външна обвивка, за да не се подуе и да умре във вакуума на космоса. Независимо колко усърдно мисли твоята майруна, тя просто ще остане прикована към земята, макар мислейки много усилено.

– Някои от нас играят тази игра от милиарди години – казал богът на бактериите, поглеждайки косо хоминидния бог. – Досега мозъкът не е бил толкова голямо предимство.

– И все пак – казал хоминидният бог.

• • •

Ето историята на нашия вид такава, каквато е била в действителност: хората са придобили мозъци, които са били необичайно големи за животни с техния размер. Те са открили огъня, построили са ферми и са извличали желязо. За изненадващо кратко време по-късно – според стандартите на биологичната еволюция – хората са кацнали на Луната, въпреки че нашият метаболизъм не може да преработва ракетно гориво и кожата ни не оцелява във вакуум.

Другите видове на Земята се раждат със специализирани умения: пчелите строят кошери, бобрите – язовири. Човекът гледа язовира на бобъра и разгадава как е направен; ние сме се научили да строим язовири, без да се нуждаем от знания, които са заложени в гените ни. Сега преграждаме реки за язовири като бобрите; строим къщи за себе си като пчелите; плетем нишки

в мрежи като паяците. Строим електроцентрали и космически ракети, които никой друг вид не строи.

Хората могат да правят неща, които нашите предци никога не са правили и които другите животни не могат да правят, благодарение на качество, което понякога наричаме „интелигентност". Повечето от нашите способности не са заложени в гените ни, но ние сме наблюдавали, опитвали, запомняли, обобщавали и след това сме постигали.

Способността да се учи не е уникална за хората. Мозъкът на мишката може да се научи как да се ориентира в лабиринт. Ние обаче можем да изпълним по-добре всичко, което може мозъкът на мишката. Можем да научим чрез химията как да се ориентираме към по-евтини торове. Можем да разработваме сложни експерименти, да разберем физиката и да изобретяваме сателити. Можем дори да поставяме мишки в лабиринти и да се учим как *те* учат.

Човешкият мозък може да се научи да се ориентира в по-широкообхватни пътища през по-широк напречен срез на реалността, отколкото всяко друго животно. *Това* е нашата специална сила.

Как функционира тази специална сила? Какво прави и как?

• • •

Според нас[A] интелигентността се състои от два основни вида работа: работата по *прогнозиране* на света и работата по *насочването* му.

Прогнозиране е да предположиш какво ще видиш (или чуеш, или докоснеш), преди да го усетиш. Ако шофираш към летището, мозъкът ти успява в задачата по прогнозиране винаги, когато предвиждаш, че светофарът ще светне в жълто или че шофьорът пред теб ще натисне спирачките.

[A] Тази гледна точка е подкрепена от някои теории, които обсъждаме в онлайн ресурсите. В крайна сметка няма да се вторачваме твърде много в дефинициите. Ако удар от мълния подпали гората около вас, не можете да се спасите, като умело дефинирате „пожар" така, че да включва само изкуствени пожари; просто трябва да бягате.

Насочването е откриването на действия, които ви водят до избран резултат. Когато шофирате към летището, мозъкът ви успява в насочването, когато намира модел на завои по улиците, така че да стигнете до летището, или намира правилните нервни сигнали, за да свие мускулите ви, така че да завъртите волана.

Повечето възможни модели на нервни импулси, които мозъкът ви би могъл да изпрати до пръстите ви, не биха завъртели волана правилно. По-голямата част от възможните модели на нервни импулси биха довели до неконтролирани потрепвания и конвулсии – би изглеждало така, сякаш имате припадък. И все пак всеки ден мозъкът ви успява да намери нервните импулси, които управляват колата или ви държат в изправено положение, избирайки правилните възможности вместо множество грешни. Когато шофирате към летището, мозъкът ви не просто изчислява точна поредица от завои, които да ви отведат до целта, а избира един от милиарди модели на нервни импулси, които съкращават мускулите ви по правилния начин, за да завъртите волана.

Прогнозирането и насочването са взаимносвързани. Управлението на колата до летището може да включва прогнозиране кои улици водят до булевард „Летище“. Прогнозирането кои улици водят до булевард „Летище“ може да включва насочване на пръстите ви към използване на карта на телефона ви.

Бихме казали, че все пак има фундаментална разлика между прогнозиране и насочване – разлика, за която ще се окаже, че има огромно значение.

Успехът в прогнозирането е лесно измерим. Ако някой очаква да види булевард „Летище“ пред себе си, но вместо това вижда Втора улица, тогава прогнозата му е била неправилна.

Обратното, за да измерим дали някой се е насочвал успешно, трябва да имаме представа *къде е искал да отиде*.

Ако колата на човека се озове пред супермаркет, това е чудесна новина, ако той е искал да купи хранителни стоки; но е провал, ако е искал да стигне до спешното отделение на болницата.

Когато двама жители на един и същ град стават по-умни, бихте очаквали те да се съгласяват все повече по въпроси, свързани с прогнозите – например дали по Втора улица има трафик

в 17:00 ч. в делничните дни. Не бихте очаквали обаче те да започнат да се насочват към едни и същи места; единият може да предпочита да ходи в парка, а другият – на театър.

Или, с други думи, *интелигентните умове могат да се насочват към различни крайни цели,* без това да е недостатък на тяхната интелигентност.

• • •

Прогнозирането и насочването не са уникални функции на биологичния ум; машините също могат да ги изпълняват. Към момента обаче хората все още са най-добрите на планетата в...

В какво точно? Хората вече не са световни шампиони по шах. Хората вече не са единствените потребители на език на планетата.

Хората вече не са уникални в способността си да четат медицински картон или да диагностицират тумори.

Хората все още са шампиони в нещо по-дълбоко – но това специално нещо сега изисква повече усилия, за да бъде описано, отколкото преди.

Струва ни се, че хората все още имат предимство в нещо, което бихме могли да наречем „генерализиране". Какво точно означава това? Бихме казали: интелигентността е *по-генерализирана,* когато може да прогнозира и да насочва в по-широк спектър от области. Хората не са непременно най-добрите във всичко; може би мозъкът на октопода е по-добър в контролирането на осемте му пипала. В по-широк смисъл обаче изглежда очевидно, че хората имат по-общо мислене от октоподите. Ние имаме по-широки области, в които можем успешно да предвиждаме и да се насочваме.

Някои ИИ са по-умни от нас в тесни области. През 1997 г. суперкомпютърът *Дийп Блу* на Ай Би Ем стана първата машина, която победи човек световен шампион в шахматна партия. *Дийп Блу* беше много умел в прогнозирането и насочването, когато ставаше дума за шах. *Дийп Блу* обаче не можеше да прогнозира как да стигне до магазина и да купи мляко, да не говорим за това да шофира дотам. Умовете могат да бъдат повече или по-малко умели в различни области на прогнозирането и насочването.