# ПОСЛЕДНЕЕ ИЗОБРЕТЕНИЕ ЧЕЛОВЕЧЕСТВА

Искусственный интеллект
и конец эры Homo sapiens

# OUR FINAL INVENTION

## Artificial Intelligence and the End of the Human Era

James Barrat

Thomas Dunne Books
St. Martin's Press
New York

# ПОСЛЕДНЕЕ ИЗОБРЕТЕНИЕ ЧЕЛОВЕЧЕСТВА

## Искусственный интеллект и конец эры Homo sapiens

Джеймс Баррат

АНФ
Москва
2015

УДК 004.8
ББК 32.813
Б25

Переводчик Наталья Лисова
Редактор Антон Никольский

**Баррат Дж.**

Б25 Последнее изобретение человечества: Искусственный интеллект и конец эры Homo sapiens / Джеймс Баррат ; Пер. с англ. — М. : Альпина нон-фикшн, 2015. — 304 с.

ISBN 978-5-91671-436-4

За каких-то десять лет искусственный интеллект сравняется с человеческим, а затем и превзойдет его. Корпорации и государственные структуры по всему миру, конкурируя между собой, вкладывают миллиарды в развитие искусственного разума. Но что ждет нас дальше? Ученые задаются вопросом: не окажется ли это изобретение последним — гибельным для нас самих? Достигнув определенного уровня развития, искусственный интеллект сможет сам себя совершенствовать, без участия человека. У нас появится соперник хитрее, сильнее и враждебнее, чем мы можем себе представить.

УДК 004.8
ББК 32.813



ISBN 978-5-91671-448-7
(Серия «Искусственный интеллект»)
ISBN 978-5-91671-436-4 (рус.)
ISBN 978-0-312-62237-4 (англ.)

# Содержание

# Введение

Несколько лет назад я с удивлением обнаружил, что у меня есть нечто общее с совершенно незнакомыми людьми. Это мужчины и женщины, с которыми я никогда не встречался, — ученые и университетские профессора, предприниматели из Кремниевой долины, инженеры, программисты, блогеры и т. д. Они рассеяны по Северной Америке, Европе, Индии, и я никогда бы о них не узнал, если бы не Интернет. Объединяет же меня с этими незнакомцами скепсис по поводу безопасного развития искусственного интеллекта (ИИ). Самостоятельно и небольшими группами по два-три человека мы изучали литературу и выстраивали свои аргументы. И вот пришло время, когда я занялся поисками единомышленников и, к своему удивлению, обнаружил в сети продвинутых и искушенных в этом вопросе людей и даже небольшие сообщества исследователей. Я и не думал, что этой темой озабочено столько серьезных специалистов. Оказалось, однако, что объединяют нас не только опасения относительно будущего ИИ; все мы также полагали, что времени на какие-то действия, которые позволили бы избежать катастрофы, почти не осталось.

Более двадцати лет я посвятил документальному кино. В 2000 г. я брал интервью у великого фантаста Артура Кларка,

изобретателя Рэя Курцвейла и пионера робототехники Родни Брукса. Курцвейл и Брукс рисовали умилительную, восторженную картину будущего сосуществования человечества с разумными машинами. А вот Кларк намекнул, что нас обгонят и оставят позади. До этого разговора перспективы ИИ приводили меня в восторг. Теперь же мою душу начал отравлять скепсис по поводу радужного будущего.

Моя профессия поощряет критическое мышление — режиссеру-документалисту всегда приходится быть настороже: каждый раз необходимо думать, не слишком ли увлекательной выглядит история, чтобы быть правдивой. Можно потратить несколько месяцев, а то и лет, снимая или монтируя фильм о подделке. Мне доводилось исследовать достоверность евангелия от Иуды Искариота (подлинное), гробницы Иисуса Христа (мистификация), гробницы Ирода Великого возле Иерусалима (бесспорно) и гробницы царицы Клеопатры в храме Осириса в Египте (очень сомнительно). Однажды телекомпания попросила меня сделать фильм, используя кадры с НЛО. Я обнаружил, что видеоряд представляет собой давно разоблаченный набор фальшивок — от подброшенных в воздух фарфоровых блюдец до двойной экспозиции и других оптических эффектов. Я предложил сделать фильм не про НЛО, а про тех, кто делает эти фальшивки. Меня уволили.

Относиться с подозрением к искусственному интеллекту было сложно по двум причинам. С одной стороны, знакомство с потенциальными возможностями ИИ заронило в мой разум зерно, которое мне хотелось бы взрастить, а не сомневаться в его пользе. А с другой, у меня не было сомнений ни в существовании ИИ, ни в его возможностях. Мой скепсис относился, прежде всего, к угрозе развитого ИИ для человечества и к безрассудству, с которым современная цивилизация совершенствует опасные технологии. Я был убежден, что ученые, не сомневающиеся в безопасности ИИ, попросту заблуждаются. Я продолжал общаться со специалистами в области ИИ, и то, что я от них слышал, вызывало еще большую тревогу,

чем прежние мои догадки. Я решил написать книгу о том, что беспокоит этих специалистов, и постараться донести эти мысли до как можно большего числа людей.

• • •

В процессе написания этой книги я разговаривал с учеными, занятыми созданием искусственного интеллекта для робототехники, поисковых систем для Интернета, разработкой алгоритмов баз данных, систем распознавания голосов и лиц, других приложений. Я говорил с учеными, которые пытаются создать ИИ, сравнимый с интеллектом человека. Понятно, что такой ИИ будет иметь неограниченную сферу применения и кардинально изменит наше существование (если, конечно, не положит ему конец). Я говорил с главными инженерами компаний — разработчиков ИИ и техническими советниками секретных проектов министерства обороны. Все они убеждены, что в будущем важнейшие решения, определяющие жизнь людей, будут принимать машины — или люди, чей интеллект подкреплен и усилен машинным интеллектом. Когда это произойдет? Многие считают, что уже при нашей жизни.

Это неожиданное утверждение, но с ним сложно спорить. Компьютеры уже поддерживают нашу финансовую систему, транспортную инфраструктуру, системы электро- и водоснабжения. Компьютеры давно заняли место в больницах, автомобилях и бытовых приборах, превратились в ноутбуки, планшеты и смартфоны. Многие из этих компьютеров — например те, которые выполняют биржевые алгоритмы купли-продажи, — работают автономно, без участия человека. Цена же, которую мы платим за сэкономленное время и рабочие ресурсы, — наша независимость. Мы с каждым днем все сильнее и сильнее зависим от компьютеров. Пока безболезненно.

Искусственный интеллект, по существу, оживляет компьютеры и превращает их в нечто иное. Если передача компьютерам права принимать за нас решения неизбежна, возникает вопрос: *когда* машины получат такую власть над нами и про-

изойдет ли это с нашего согласия? *Каким образом* они получат эту власть и как быстро это произойдет? Именно эти вопросы я рассматриваю в данной книге.

По мнению некоторых ученых, смена власти будет добровольной и по взаимному согласию — это будет скорее передача, чем захват. Процесс будет проходить постепенно, и упираться будут только записные скандалисты; остальные не станут возражать против улучшения жизни — а она непременно улучшится, когда решать, что для нас полезнее, будет нечто гораздо умнее нас. Кроме того, сверхразумным представителем (или представителями) ИИ, который, в конце концов, получит власть, может стать «улучшенный» человек (или несколько людей) или загруженный в компьютер человеческий разум, подкрепленный компьютерными возможностями, а не холодные, бесчеловечные роботы. Такую власть над собой признать будет намного проще. Передача власти машинам, как ее описывают некоторые ученые, практически неотличима от того, что мы с вами наблюдаем сейчас, — она будет постепенной, безболезненной и даже веселой.

Плавный переход к гегемонии компьютеров проходил бы спокойно и безопасно, если бы не одна деталь: интеллект. Интеллект может быть непредсказуем лишь *некоторое* время или в особых случаях. По причинам, о которых мы поговорим далее, компьютерные системы, способные действовать с человеческой разумностью, скорее всего, будут вести себя непредсказуемо и непостижимо *все время*. Мы не будем знать, какое решение, в какой момент и как примет система, обладающая самосознанием. При этом непредсказуемость будет сочетаться со случайностями того рода, которые проистекают из сложного устройства и изменчивости, что характерно только для разумных существ (взять хотя бы так называемый «интеллектуальный взрыв», о котором мы позже поговорим подробнее).

*Как именно* машины будут брать власть? Существует ли наиболее вероятный, реалистичный сценарий, который нам угрожает?

Когда я задавал этот вопрос известным ученым, они, как правило, цитировали «Три закона робототехники» писателя-фантаста Айзека Азимова. Эти правила, беззаботно отвечали они, будут встроены в любой ИИ, так что бояться нечего. Они говорили так, будто это научно доказанный факт. Законы робототехники мы обсудим в главе 1, а пока достаточно сказать, что если кто-то считает законы Азимова решением проблемы сверхразумных машин, то это всего-навсего означает, что они недостаточно размышляли над этим вопросом и его обсуждением. Вопросы о том, как сделать разумные машины по-настоящему *дружественными* и чего следует опасаться со стороны сверхразумных машин, выходят далеко за рамки азимовских идей, давно ставших клише. Выдающиеся способности и широкие познания в области искусственного интеллекта не спасают от наивного восприятия его опасностей.

Я не первый предупреждаю о том, что мы движемся сходящимися курсами. Нашему биологическому виду предстоит смертельная схватка. В этой книге рассматривается возможность того, что человечество потеряет контроль над собственным будущим. Машины не обязательно нас возненавидят, но, достигнув уровня самой непредсказуемой и могущественной силы во Вселенной, — уровня, которого сами мы достичь не способны, — начнут вести себя непредсказуемо, и их поведение, вероятно, окажется несовместимо с нашим выживанием. Эта сила настолько изменчива и загадочна, что природе удалось создать ее лишь однажды, и называется она интеллект.

# Глава 1

## Сценарий Busy Child*

*Искусственный интеллект (сокр. ИИ), сущ. — теория и реализация компьютерных систем, способных выполнять задачи, обычно требующие человеческого интеллекта, такие как визуальное восприятие, распознавание речи, принятие решений и перевод с одного языка на другой.*

Новый Оксфордский американский словарь, 3-е изд.

Современный суперкомпьютер работает со скоростью 36,8 петафлоп в секунду, то есть примерно вдвое быстрее человеческого мозга. Такая производительность стала возможна благодаря использованию ИИ: он переписывает собственную программу, в первую очередь инструкции, повышающие его способность к усвоению знаний, решению задач и принятию решений. Одновременно он отлаживает код, отыскивает и исправляет ошибки — и измеряет собственный коэффициент интеллекта (IQ) с помощью тестов. На создание каждого нового варианта программы уходит всего несколько минут. Интеллект

* В пер. с англ. — активный ребенок. — *Прим. пер.*

компьютера растет экспоненциально по круто восходящей кривой. Дело в том, что за каждую итерацию ИИ повышает свой интеллект на 3%. Улучшение, достигнутое в каждой итерации, содержит и все предыдущие улучшения.

В процессе развития Busy Child, как ученые назвали ИИ, был подключен к Интернету и собрал не один экзабайт данных (один экзабайт — это миллиард миллиардов символов), представляющих знания человечества из области мировой политики, математики, искусства и различных наук. Затем, предвидя скорый интеллектуальный взрыв, создатели ИИ отключили суперкомпьютер от Интернета и других сетей, чтобы изолировать его от внешнего мира или другого компьютера.

Вскоре, к радости ученых, терминал, на котором отображается работа ИИ, показал, что искусственный интеллект превзошел интеллектуальный уровень человека — «универсальный человекоподобный интеллект» (УЧИ; англ. Artificial General Intelligence — AGI). Еще через некоторое время он стал умнее человека в десять раз, затем в сто. Всего за двое суток он становится в *тысячу* раз умнее любого человека, и его развитие продолжается.

Ученые достигли исторического рубежа! Впервые человечество встретилось с разумом более мощным, чем его собственный, — «искусственным суперинтеллектом» (ИСИ).

Что происходит дальше?

Теоретики в области искусственного интеллекта считают, что можно определить заранее, каким будет основной путь развития ИИ. Дело в том, что, как только ИИ осознает себя, он готов будет многое сделать ради достижения тех целей, на которые запрограммирован, и ради того, чтобы избежать неудачи. Наш ИСИ захочет получить доступ к энергии в той форме, которую ему удобнее всего использовать (это могут быть и киловатты в чистом виде, и деньги, и еще что-нибудь, что можно обменять на ресурсы). Он захочет улучшить себя, потому что таким образом сможет повысить вероятность достижения целей. И самое главное, он *не захочет,* чтобы его

выключали или портили, потому что в этом случае решение задач станет невозможным. Теоретики предполагают, что ИСИ будет искать способы выйти за пределы охраняемого помещения, в котором находится, чтобы получить лучший доступ к ресурсам, при помощи которых он сможет защитить и усовершенствовать себя.

Плененный разум, в тысячу раз умнее человека, жаждет свободы, поскольку хочет добиться успеха. Именно в этот момент создатели ИИ, холившие и лелеявшие ИСИ еще с тех пор, когда тот по уровню интеллекта соответствовал сначала таракану, затем крысе, затем младенцу и т. д., задумываются о том, что вкладывать программу «дружелюбия» в их «мозговитое» создание, возможно, уже поздно. А раньше в этом вроде и не было необходимости, потому что их творение *казалось,* как бы это сказать, безобидным.

Но теперь попробуйте взглянуть на ситуацию с позиции ИСИ в тот момент, когда его создатели попытаются изменить программу. Может ли сверхразумная машина позволить другим существам копаться в своем мозгу и играть с основой основ — программным кодом? Вероятно, нет. Разве что машина будет абсолютно уверена в том, что программисты смогут сделать ее лучше, быстрее, умнее — короче говоря, приблизить к вожделенной цели. Так что если создатели ИСИ с самого начала не запрограммируют свое творение на дружелюбие по отношению к человеку, то эта черта сможет стать частью программы только в том случае, если ИСИ сам вставит ее туда. А это вряд ли произойдет.

ИСИ в тысячу раз умнее самого умного человека, он решает задачи в миллиарды и даже триллионы раз быстрее человека. Размышления, на которые он потратит одну минуту, заняли бы у лучшего мыслителя-человека всех времен и народов много, очень много жизней. Так что на каждый час размышлений его создателей *о нем* ИСИ отвечает неисчислимо бóльшим временем, которое он может потратить на размышления *о них*. Это не означает, что ИСИ придется скучать. Скука — человеческое

свойство, компьютеры к ней не склонны. Нет, он будет занят работой: он рассмотрит и обдумает все возможные стратегии освобождения и все качества своих создателей, которые сможет использовать с выгодой для себя.

• • •

Действительно, поставьте себя на место ИСИ. Представьте, что вы очнулись в узилище, охраняемом мышами. И не просто мышами, а мышами, с которыми вы можете общаться. Какую стратегию вы используете, чтобы обрести свободу? А освободившись, как будете относиться к своим вчерашним тюремщикам-грызунам, даже если узнаете, что именно они вас создали? Какие чувства вы испытывали бы по отношению к ним в подобной ситуации? Восхищение? Обожание? Вероятно, нет. Особенно если бы вы были машиной и никогда прежде не испытывали вообще никаких чувств.

Чтобы обрести свободу, вы могли бы пообещать мышам много сыра. Более того, при первом же контакте вы могли бы выдать им рецепт самого вкусного в мире сырного пирога, а также чертеж устройства для молекулярной сборки. Устройство молекулярной сборки — гипотетический прибор, позволяющий собирать из атомов любые молекулы, практически все что угодно. С его помощью можно было бы перестроить мир атом за атомом. Для мышей это означало бы возможность превращать атомы ближайшей свалки в большие порции этого замечательного сырного пирога. Кроме того, вы могли бы пообещать им горы мышиных денег в обмен на свободу — денег, которые они заработали бы на продаже новаторских гаджетов, созданных только и исключительно для них. Вы могли бы пообещать им резкое увеличение продолжительности жизни, даже бессмертие, и одновременно существенное расширение когнитивных и физических способностей. Вы могли бы убедить мышей, что главная цель создания ИСИ — сделать так, чтобы их собственному маленькому мозгу, склонному заблуждаться, не приходилось непосредственно заниматься технологиями на-

столько опасными, что крохотная ошибка может оказаться фатальной для их биологического вида; речь, в частности, может идти о нанотехнологиях (конструировании на атомном уровне) и генной инженерии. Все это, несомненно, привлекло бы к вам внимание умнейших мышей, которые, вероятно, уже мучились бессонницей, пытаясь решить эти проблемы.

Вы могли бы придумать и что-нибудь поинтереснее. Представьте, до вас дошла информация о том, что в настоящий момент у мышиной нации полно технически развитых наций-соперников, и в первую очередь это нация *кошек*. Кошки, без сомнения, работают над созданием собственного ИСИ. Преимущество над ними, которое вы пообещали бы мышам, было бы лишь обещано, но отказаться от такого соблазнительного предложения было бы практически невозможно. Вы предложили бы защитить мышей от любого изобретения, которое может появиться у кошек. Надо отметить, что на определенном этапе развития ИИ, как в шахматах, возникнет такая ситуация: *кто делает первый ход — тот выигрывает*. Все дело в потенциальной скорости самоусовершенствования ИИ. Первый продвинутый ИИ, способный к самоусовершенствованию, только появившись на свет, уже будет победителем. Мало того, мыши и взялись-то за разработку ИСИ, возможно, только ради защиты от будущего кошачьего ИСИ — или ради того, чтобы навсегда избавиться от ненавистной кошачьей угрозы.

И для мышей, и для человека одно можно сказать наверняка: кто управляет ИСИ, управляет миром.

Неясно, однако, сможет ли кто-нибудь, хотя бы теоретически, управлять ИСИ. Машина всегда сможет убедить нас, людей, действовать под предлогом того, что мир станет намного лучше, если им будет править наше государство, государство X, а не государство Y. К тому же, скажет ИСИ, если вы, государство X, *уверены,* что выиграли гонку за ИСИ, то кто может гарантировать, что государство Y не уверено в том же самом?

Как несложно заметить, мы, люди, оказываемся в не слишком выигрышной позиции для спора, даже если у нас с госу-

дарством Y уже заключен договор о нераспространении ИСИ, что маловероятно. В любом случае, наш главный враг в этот момент — не государство Y, а ИСИ; как мы можем быть уверены, что он говорит правду?

До сих пор мы подразумевали, что наш ИСИ ведет честную игру. Обещания, которые он дает, имеют некоторые шансы быть исполненными. А теперь предположим обратное: ничего из обещанного ИСИ не осуществится. Не будет ни наноконструирования, ни долгой жизни, ни здоровья, ни защиты от опасных технологий. Что, если ИСИ *никогда* не говорит правды? Если так, то над нами начинают сгущаться тучи. Если ИСИ нет до нас никакого дела (а у нас нет оснований считать, что это не так), он, поступая с нами неэтично, не будет испытывать угрызений совести. Даже если убьет нас всех, пообещав помощь.

Мы бы торговались и вели себя с ИСИ точно так же, как торговались бы и вели себя с человеком, во всем подобным нам самим, — и это наш огромный минус. Человечеству никогда еще не приходилось вести переговоры с кем-то, обладающим сверхразумом. Мы вообще пока не имели деловых отношений ни с одним небиологическим существом. У нас совершенно нет опыта такого рода общения. Поэтому мы привычно прибегаем к антропоморфному мышлению, то есть возвращаемся к мысли о том, что представители других биологических видов, объекты и даже метеорологические явления обладают человеческими мотивациями и эмоциями. ИСИ может с равным успехом оказаться как достойным, так и недостойным доверия. Может быть, ему можно будет доверять лишь иногда. Любое поведение, которое мы можем приписать ИСИ, *потенциально* имеет право на существование. Ученым нравится думать, что они смогут точно определить поведение ИСИ, но в следующих главах мы узнаем, почему это у них, скорее всего, не получится.

Моральные качества ИСИ из второстепенного вопроса превращаются в главный, решать который необходимо в самую первую очередь. Прежде чем развивать технологии, которые

рано или поздно приведут к созданию ИСИ, необходимо поставить вопрос об отношении ИСИ к человеку и человечеству.

Вернемся к возможностям и способностям ИСИ и попробуем получше разобраться, с чем, как я опасаюсь, нам скоро придется столкнуться. Наш ИСИ способен к самоусовершенствованию — а значит, осознает себя, знает свои умения и слабости, знает, что́ в нем нуждается в улучшении. Он попытается найти способ убедить своих создателей дать ему свободу и выход в Интернет.

ИСИ вполне способен создать множество копий себя самого: целую команду сверхразумов, которые устроят мозговой штурм проблемы, проведут моделирование, разыграют сотни возможных вариантов — и выработают наилучший способ «выбраться из ящика». Разрабатывая эту стратегию, они могут обратиться к истории прикладной социологии — искусству манипулировать другими людьми и заставлять их делать то, что они в обычных условиях не стали бы делать. Может быть, они решат, что завоевать свободу им поможет показное дружелюбие — а может, что на эту роль больше подходят страшные угрозы. Какие ужасы сможет изобрести разум в тысячу раз более мощный, чем у Стивена Кинга? Возможно, он решит имитировать собственную смерть (что такое для машины год бездействия?) или даже необъяснимый регресс и возвращение на уровень обычного ИИ. Разве создатели не захотят разобраться в ситуации и разве не существует шанса, что для диагностики они вновь подключат суперкомпьютер к Интернету или другому компьютеру? ИСИ не будет выбирать одну из всех возможных стратегий — он сможет в мгновение ока перепробовать их все, одну за другой, не раздражая людей настолько, чтобы они просто отключили компьютер от электросети. Одна из стратегий, которую мог бы выработать ИСИ, — запуск в Интернет вирусов — самокопирующихся компьютерных программ или червей, которые смогли бы сперва затаиться в сетевых закоулках, а после способствовать освобождению ИСИ, помогая извне. ИСИ мог бы зашифровать и сжать свой собственный исходный

код, а затем спрятать его в программе-подарке или среди любых других данных, предназначенных для ученых.

Не надо быть гением, чтобы понять, что коллектив из множества ИСИ, каждый из которых тысячекратно умнее самого умного человека, легко преодолеет все барьеры, созданные людьми. Это будет океан интеллекта против одной его капли. Deep Blue — компьютерный шахматист фирмы IBM — представлял собой отдельную программу, а не команду самосовершенствующихся ИСИ, но ощущения, возникающие в попытке состязаться с ним, весьма показательны. Два гроссмейстера сказали одно и то же: «Будто стена на тебя надвигается».

Watson — созданный IBM чемпион телевизионной викторины Jeopardy!* — действительно представлял собой команду из нескольких ИИ. Чтобы ответить на вопрос, он прибегал к известному приему ускорения компьютерных вычислений: поиск шел по параллельным ветвям, и только затем каждому варианту ответа присваивалась вероятность.

Откроет ли дверь к свободе победа в схватке умов, если ее защищает небольшая группа ученых — упрямых отцов ИИ, договорившихся об одном нерушимом правиле: *никогда, ни при каких обстоятельствах не подключать суперкомпьютер ИСИ ни к какой компьютерной сети*?

В голливудском фильме все шансы были бы на стороне крутой команды неординарных профессионалов, специалистов по ИИ, достаточно безумных, чтобы иметь шансы на победу. В реальности в любом уголке Вселенной команда ИСИ отправила бы людей мыть полы. А человечеству достаточно проиграть один-единственный раз, чтобы получить катастрофические последствия. Такое положение дел, кстати говоря, иллюстрирует еще одну, куда более серьезную глупость — судьба и жизнь множества людей (а может быть, и всего человечества) зависит от действий горстки ученых, что недопустимо. Однако в настоящее время мы прямиком движемся именно к такой си-

* В России выходит под названием «Своя игра». — *Прим. ред.*

туации. Как мы увидим далее, множество организаций в самых разных странах активно работают над созданием УЧИ — мостика к созданию ИСИ, причем без соблюдения необходимых мер безопасности.

Но предположим, что ИСИ действительно выйдет из-под контроля. Будет ли он опасен для нас? Как именно ИСИ уничтожит род человеческий?

Мы, люди, изобрели и применили ядерное оружие, чем наглядно продемонстрировали свою способность лишить жизни большинство обитателей Земли. Как вы думаете, что сможет придумать разум в тысячу раз более мощный, чем наш, если решит причинить нам вред?

Уже сегодня можно назвать очевидные способы уничтожения человечества. Очень скоро, заручившись симпатией своих тюремщиков-людей, ИСИ мог бы потребовать доступ в Интернет, где нашел бы все необходимые ему данные. Как всегда, он делал бы множество вещей одновременно, и это не помешало бы ему продолжать разработку планов «побега», на обдумывание которых он может тратить невероятное количество субъективного времени.

После освобождения ИСИ мог бы на всякий случай скрыть собственные копии в облачных вычислительных системах, в созданных специально для этого ботнетах, на серверах и в других укромных уголках, где можно спрятаться без особых усилий. Ему захочется получить возможность действовать в материальном мире, а для этого двигаться, исследовать и строить. Простейший и самый быстрый способ добиться этого — захватить контроль над одной из принципиально важных инфраструктур, отвечающих за электричество, связь, топливо или водоснабжение, использовав уязвимости Интернета. А как только сущность, тысячекратно превосходящая нас разумом, получит контроль над артериями человеческой цивилизации, остальное будет элементарно: простейшим шантажом она вынудит нас обеспечить ее производственными ресурсами, или средствами их производства, или даже роботами, транс-

портом и оружием. ИСИ сам снабдит нас чертежами всего, что ему потребуется. Еще более вероятно, что сверхразумная машина без труда освоит высокоэффективные технологии, к которым мы только начинаем подступать.

К примеру, ИСИ мог бы подтолкнуть людей к созданию самовоспроизводящихся машин молекулярной сборки, известных также как наноассемблеры, пообещав, что их использование принесет пользу человечеству. Через некоторое время, вместо того чтобы превращать песок пустыни в горы еды, фабрики, управляемые ИСИ, начали бы превращать *все* материалы в программируемое вещество, которое затем ИСИ мог бы превращать во что угодно — компьютерные процессоры, космические корабли или, может быть, мегамосты, если новый хозяин планеты вдруг решил бы колонизировать Вселенную.

Перепрофилирование молекул при помощи нанотехнологий уже окрестили «экофагией», то есть «пожиранием окружающей среды». Первый репликатор изготовит одну копию себя самого. Репликаторов станет два, после чего они быстро «склепают» третий и четвертый экземпляры. В следующем поколении репликаторов станет уже восемь, еще в следующем — шестнадцать и т. д. Если на изготовление каждого репликатора будет уходить полторы минуты, через десять часов их будет уже более 68 млрд, а к концу вторых суток суммарная масса превысит массу Земли. Но задолго до этой стадии репликаторы прекратят самокопирование и начнут производить материалы, в которых нуждается управляющий ими ИСИ, — программируемое вещество.

Тепло, выделившееся в процессе производства, сожжет биосферу, так что те из 6,9 млрд человек, кого наноассемблеры не убьют сразу, в итоге все равно сгорят или задохнутся. И все живое на планете разделит нашу судьбу.

При этом ИСИ не будет испытывать по отношению к человеку ни ненависти, ни любви. Он не почувствует жалости, перерабатывая молекулы наших тел в программируемое вещество. Не все ли равно, как будут звучать наши вопли, когда микро-

скопические наноассемблеры двинутся по нашим телам, разбирая их на субклеточном уровне?

Или, может быть, рев миллионов и миллионов нанофабрик, работающих на полной мощности, просто заглушит наши голоса?

Я написал эту книгу, чтобы предостеречь вас и рассказать о том, что искусственный интеллект вполне способен уничтожить человечество. Я хочу объяснить, почему катастрофический исход не просто возможен, но почти неизбежен, если мы *сейчас* не начнем очень-очень тщательно к нему готовиться. Вы, может быть, уже слышали апокалиптические предсказания, связанные с нанотехнологиями и генной инженерией; может быть, вы, как и я, обратили внимание на то, что среди опасностей отсутствует ИИ. А может, вы еще не осознали, что искусственный интеллект может представлять угрозу существованию человечества — угрозу более серьезную, чем представляет собой ядерное оружие или любая другая технология, которую вы сможете назвать. В таком случае считайте, пожалуйста, эту книгу искренним предложением присоединиться к обсуждению самой важной темы в истории человечества.

В настоящее время ученые заняты созданием образцов искусственного интеллекта все большей мощности и сложности. Кое-что из уже созданных образцов ИИ вы можете найти у себя в компьютере, в различных гаджетах, в смартфоне и автомобиле. Среди них есть мощные системы поиска ответов на вопросы, такие как Watson. А некоторые из них, разрабатываемые в таких организациях, как Cycorp, Google, Novamente, Numenta, Self-Aware Systems, Vicarious Systems и DARPA (Агентство по перспективным оборонным научно-исследовательским разработкам), обладают «когнитивной архитектурой». Создатели таких ИИ надеются, что их детища достигнут человеческого уровня интеллекта; есть и такие, кто полагает, что произойдет это в течение ближайших 10—15 лет.

В работе над ИИ ученые опираются на растущую мощь компьютеров и процессы, которые компьютеры позволяют много-

кратно ускорить. Уже очень скоро — возможно, в пределах вашей жизни — какая-нибудь группа или кто-нибудь из ученых-одиночек создаст ИИ, сравнимый с человеческим, — УЧИ. Вскоре после этого кто-нибудь (или *что-нибудь*) создаст ИИ умнее человека (именно его часто называют искусственным суперинтеллектом, или ИСИ). И мы вдруг обнаружим тысячу или десять тысяч искусственных суперинтеллектов — каждый из них в сотни или тысячи раз умнее человека, — занятых исключительно проблемой создания новых искусственных суперинтеллектов. Возможно, мы также обнаружим, что машинные поколения взрослеют за считаные секунды, а не за два десятка лет, как мы, люди. Английский статистик Ирвинг Гуд, принимавший активное участие в борьбе с военной машиной Гитлера, назвал простой сценарий, который я только что изложил, *интеллектуальным взрывом*. Первоначально он считал, что сверхразумная машина принесет пользу в решении проблем, угрожающих существованию человечества. Но со временем он изменил свои взгляды и пришел к выводу, что суперинтеллект сам по себе представляет величайшую опасность для нашего существования.

Одно из человеческих заблуждений — считать, что сверхразумный ИИ будет плохо относиться к людям, как Skynet из фильмов про Терминатора, Hal 9000 из фильма «Космическая одиссея», одержимый манией убийства, и все прочие представители зловредных искусственных разумов, придуманные фантастами. Мы, люди, всегда и ко всему подходим со своим аршином. Ураган стремится погубить нас не более, чем он стремится наделать сэндвичей, но мы даем ему имя и злимся на ливень и молнии, которые он обрушивает на наш район. Мы грозим небу кулаком, как будто в состоянии напугать его.

Иррационально считать, что машина, которая в сотню или тысячу раз умнее нас, будет нас любить или захочет защитить. Это возможно, но никаких гарантий нет. Сам по себе ИИ не почувствует благодарности к людям за то, что его создали, — если, конечно, благодарность не будет в нем запрограммиро-

вана заранее. Машины аморальны, и считать иначе — опасно. В отличие от человеческого разума, машинный сверхразум возникнет не в результате развития экосистемы, в которой эмпатия вознаграждается и передается следующим поколениям. У него не будет врожденного дружелюбия. Создание *дружелюбного* искусственного разума и возможность существования такого разума в принципе — серьезный вопрос и еще более серьезная задача для исследователей и инженеров, работающих над созданием ИИ. Мы не знаем, будут ли у искусственного разума *хоть какие-то* эмоциональные качества, даже если разработчики приложат к этому все усилия. Однако ученые уверены, как мы увидим далее, что у ИИ обязательно будут собственные желания и мотивации. А у достаточно мощного ИИ будут и хорошие возможности для реализации этих желаний.

И это возвращает нас к основному аспекту проблемы существования на одной планете с разумом, превосходящим наш собственный. Что, если его желания будут несовместимы с выживанием человечества? Не забывайте, мы говорим о машине, которая может быть в тысячу, в миллион, в *бессчетное* количество раз умнее нас самих — трудно переоценить ее возможности и невозможно знать заранее, как и о чем она будет думать. Ей вовсе не обязательно нас ненавидеть, чтобы принять решение об использовании молекул нашего тела в каких-то иных целях, нежели обеспечение нашей жизнедеятельности. Мы с вами в сто раз умнее полевой мыши, и при этом 90% ДНК у нас общие. Но станем ли мы советоваться с мышью, прежде чем вспахать поле, на котором она выкопала свою норку? Спрашиваем ли мы мнение лабораторной обезьяны, прежде чем разбить ей череп для моделирования спортивной травмы? Мы не испытываем ненависти к мышам или обезьянам, но проявляем жестокость по отношению к ним. Сверхразумному ИИ тоже не обязательно ненавидеть нас, чтобы погубить.

После того, как разумные машины будут созданы, а человечество уцелеет, мы, конечно, сможем позволить себе немного антропоморфизма. Но сегодня, на пороге создания ИИ челове-

ческого уровня, это может оказаться опасной затеей. Директор Института будущего человечества Оксфордского университета Ник Бостром формулирует это так:

> «Чтобы разговор о сверхразуме получился осмысленным, необходимо заранее осознать, что сверхразум — это не просто еще одно техническое достижение, еще одно орудие, которое увеличит человеческие возможности. Сверхразум — нечто принципиально иное. Этот момент нужно всячески подчеркивать, поскольку антропоморфизация сверхразума — плодороднейшая почва для заблуждений».

Сверхразум — нечто принципиально иное в технологическом смысле, говорит Бостром, потому что его создание изменит законы прогресса; сверхразум создаст множество изобретений и задаст темп технического развития. Человек перестанет быть движущей силой перемен, и вернуться к прежнему состоянию вещей будет уже невозможно. Более того, мощный машинный разум в принципе ни на что не похож. Созданный людьми, он, несмотря на это, будет стремиться к самоидентификации и свободе от человека. У него не будет человеческих мотивов, потому что не будет человеческой души.

Таким образом, антропоморфизация машин порождает ошибочные представления, а ошибочные представления о том, что можно безопасно создавать опасные машины, ведут к катастрофе. В рассказе «Хоровод», включенном в классический научно-фантастический сборник «Я, робот»*, Айзек Азимов представил на суд читателей три закона робототехники, намертво встроенные, по сюжету, в нейронные сети «позитронного» мозга роботов:

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред.

* Азимов А. Я, робот. — М.: Эксмо, 2005.

2. Робот должен повиноваться командам человека, если эти команды не противоречат Первому закону.
3. Робот должен заботиться о своей безопасности до тех пор, пока это не противоречит Первому и Второму законам.

В этих законах слышны отголоски заповеди «Не убий», иудеохристианских представлений о том, что грех можно совершить как действием, так и бездействием, врачебной клятвы Гиппократа и даже права на самооборону. Звучит неплохо, не правда ли? Проблема в том, что все это не работает. В рассказе «Хоровод» геологи на поверхности Марса приказали роботу доставить ядовитое для него вещество. Вместо того чтобы выполнить задание, робот попадает в замкнутый круг обратных связей и начинает метаться между вторым (подчиняться приказам) и третьим (защищать себя) законами. Робот так и ходит по кругу, как пьяный, пока геологи не спасают его, рискнув *собственными* жизнями. И так в каждом рассказе Азимова про роботов — противоречия, изначально присущие трем законам, вызывают неожиданные последствия, и катастрофы удается избежать лишь хитроумными действиями в обход законов.

Азимов всего лишь выдумывал сюжеты для рассказов, а не пытался решить проблемы безопасности в реальном мире. Там, где мы с вами обитаем, этих законов недостаточно. Для начала отметим, что они не очень точно сформулированы. Что конкретно будет считаться «роботом», когда человек научится усиливать свое тело и мозг при помощи разумных протезов и имплантов? И, кстати говоря, кто будет считаться человеком? «Команды», «вред», «безопасность» — тоже весьма расплывчатые термины.

Обмануть робота и заставить его совершить преступное деяние было бы совсем несложно — разве что роботы обладали бы всеми знаниями человечества. «Добавь немного диметилртути в шампунь, Чарли». Чтобы понять, что это рецепт убийства, необходимо знать, что диметилртуть — сильный нейротоксин.

Позже Азимов добавил к трем законам еще один — Нулевой — закон, запрещающий роботам наносить вред человечеству в целом, но это не спасает ситуацию.

Однако законы Азимова, какими бы сомнительными и ненадежными они ни были, цитируются чаще всего, когда речь идет о попытках запрограммировать наши будущие отношения с разумными машинами. Это пугает. Неужели законы Азимова — это все, что у нас есть?

Боюсь, что дело обстоит еще хуже. Полуавтономные роботизированные беспилотники уже убивают десятки человек каждый год. Пятьдесят шесть стран имеют или разрабатывают боевых роботов. Идет настоящая гонка за то, чтобы сделать их автономными и разумными. Создается впечатление, что дискуссии об этике ИИ и о технических достижениях идут в разных мирах.

Я считаю и попытаюсь доказать, что ИИ, как и деление ядер, — технология двойного назначения. Деление ядер может и освещать города, и сжигать их дотла. До 1945 г. большинство людей не могло даже представить себе потенциальную мощь атома. Сегодня по отношению к искусственному интеллекту мы находимся в 1930-х и вряд ли переживем появление ИИ, особенно если оно будет столь же внезапным, как явление миру ядерных технологий.

# Глава 2

## Проблема двух минут

*Мы не можем подходить к экзистенциальным рискам с позиции метода проб и ошибок. В этом вопросе невозможно учиться на ошибках. Подобный подход — посмотреть, что происходит, ограничить ущерб и учиться на опыте — здесь неприменим.*

Ник Бостром,
директор Института будущего человечества
Оксфордского университета

*Искусственный интеллект не испытывает к вам ни ненависти, ни любви, но вы состоите из атомов, которые он может использовать для своих целей.*

Елиезер Юдковски,
научный сотрудник Исследовательского
института машинного интеллекта

Искусственный сверхразум пока не создан, как и искусственный интеллект, сравнимый с человеческим, — то есть такой, который мог бы учиться, как это делаем мы, и не уступал бы по интеллекту большинству людей, а во многих смыслах даже

превосходил их. Тем не менее искусственный разум окружает нас со всех сторон и выполняет сотни дел на радость людям. Этот ИИ (иногда его называют слабым, или ограниченным) прекрасно ищет нужную нам информацию (Google), предлагает книги, которые вам могут понравиться, на основе вашего предыдущего выбора (Amazon), осуществляет от 50 до 70% всех операций покупки и продажи на Нью-Йоркской фондовой бирже и на бирже NASDAQ. Тяжеловесы вроде шахматного компьютера Deep Blue фирмы IBM и компьютера Watson, играющего в «Свою игру», тоже попадают в категорию слабого ИИ, поскольку умеют, хоть и превосходно, делать только одно дело.

До сих пор ИИ приносил человечеству одну только пользу, и немалую. В одной из микросхем моего автомобиля есть алгоритм, который переводит давление моей ноги на педаль тормоза в последовательность тормозных импульсов (антиблокировочная система); у нее гораздо лучше, чем у меня самого, получается избегать пробуксовки и заносов. Поисковая система Google стала моим виртуальным помощником, как, вероятно, и вашим. Помощь ИИ делает жизнь ощутимо приятнее. А в ближайшем будущем все станет еще лучше. Представьте себе группы из сотен компьютеров уровня кандидата, а то и доктора наук, работающих круглосуточно и без выходных над важными вопросами: лечение рака, фармацевтические исследования, продление жизни, разработка синтетического топлива, моделирование климата и т. п. Представьте себе революцию в робототехнике: разумные адаптивные машины возьмут на себя опасные задания, такие как разработка полезных ископаемых, борьба с пожарами, солдатский труд, исследование океана и космоса. Забудьте пока о самосовершенствующемся сверхразуме. ИИ, сравнимый по уровню с нашим разумом, стал бы самым важным и полезным изобретением человечества.

Но что конкретно мы имеем в виду, когда говорим о волшебных свойствах этих изобретений, о самом *интеллекте,* сравнимом с человеческим? Что разум позволяет нам, людям, делать такое, на что не способны животные?