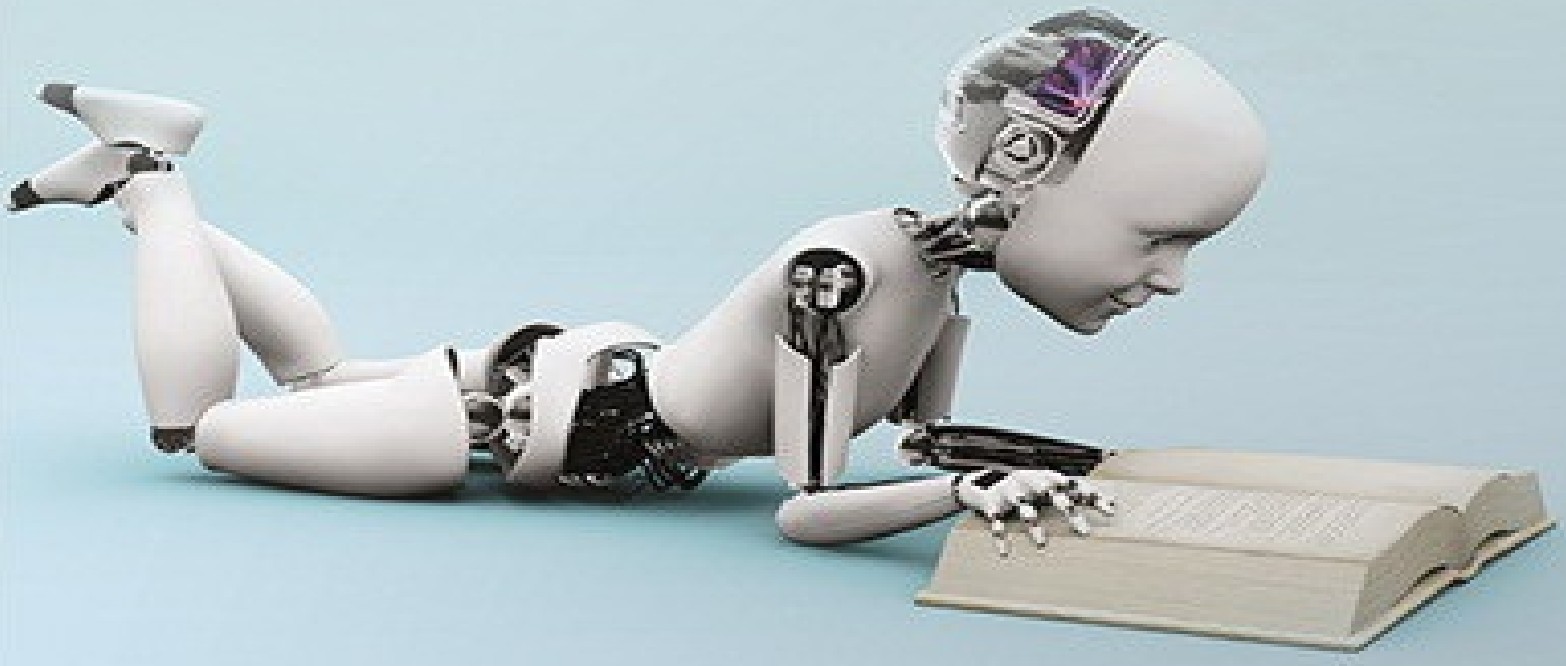
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Джеймс Баррат

ПОСЛЕДНЕЕ ИЗОБРЕТЕНИЕ ЧЕЛОВЕЧЕСТВА

Искусственный интеллект и конец эры Homo sapiens



АНФ

Annotation

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.

Баррат Джеймс
ПОСЛЕДНЕЕ ИЗОБРЕТЕНИЕ ЧЕЛОВЕЧЕСТВА:
Искусственный интеллект и конец эры Homo sapiens

Все права защищены. Никакая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, включая размещение в сети Интернет и в корпоративных сетях, а также запись в память ЭВМ для частного или публичного использования, без письменного разрешения владельца авторских прав. По вопросу организации доступа к электронной библиотеке издательства обращайтесь по адресу mylib@alpina.ru

© James Barrat, 2013

This edition published by arrangement with William Clark Associates and Synopsis Literary Agency.

© Издание на русском языке, перевод, оформление. ООО «Альпина нонфикшн», 2015

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 произойдет 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[1]

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 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 бо́льшим 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 настолько 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 государством 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, что 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![2] — 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 ситуации. 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, транспортом 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 микроскопические 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 растушую мощь компьютеров и процессы, которые компьютеры позволяют многократно 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 запрограммирована 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 человеческого уровня, это может оказаться опасной затеей. Директор Института будущего человечества Оксфордского университета Ник Востром 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 Востром, 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»[3], 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.

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 Востром, 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?

Благодаря своему человеческому интеллекту вы, к примеру, можете говорить по телефону или управлять автомобилем. Можете распознавать тысячи повседневных объектов, описывать их текстуру и свойства, знаете, как с ними обращаться. Вы можете вдумчиво пользоваться Интернетом. Не исключено, что можете посчитать до десяти на нескольких языках, а может быть, свободно владеете некоторыми из них. У вас немалый запас бытовых сведений: так, вы знаете, что ручки бывают и на дверях, и на чайных чашках, а также имеете бесчисленное количество других полезных знаний об окружающем мире. А еще вы можете изменять окружающий вас мир и приспосабливаться к переменам.

Вы умеете совершать действия последовательно или в разных сочетаниях, умеете помнить о каком-то деле, занимаясь при этом чем-то другим, более насущным. А еще вы умеете без лишних усилий и колебаний переходить от одного дела к другому, не забывая учитывать разные начальные данные. Важнее всего, может быть, то, что вы способны осваивать новые умения и усваивать новые факты, а также планировать собственное развитие. Большинство живых существ рождается с полным набором готовых способностей, которые могут пригодиться им в жизни. Но мы не такие.

Потрясающий спектр сложных навыков — вот что мы имеем в виду,

говоря об интеллекте человеческого уровня (УЧИ), том самом интеллекте, к которому стремятся разработчики ИИ.

Нужно ли тело машине, обладающей интеллектом человеческого уровня? Чтобы отвечать нашему определению УЧИ, компьютер должен иметь возможность обмениваться информацией с внешним миром и обладать способами манипулирования объектами в реальном мире, не более того. Но, как мы уже убедились, рассматривая сценарий развития Busy Child[4], мощный интеллект способен добиться того, чтобы объектами в реальном мире манипулировал кто-то другой (или что-то другое). Алан Тьюринг предложил тест на интеллект для машин, известный сегодня как тест Тьюринга (мы поговорим о нем позже). Его стандарт демонстрации интеллекта человеческого уровня требует лишь самых базовых устройств ввода-вывода, таких как клавиатура и монитор.

Самый сильный аргумент в пользу того, что продвинутому ИИ необходимо тело, относится к фазе развития и обучения ИИ. Не исключено, что ученые выяснят, что «вырастить» ИИ человеческого уровня без всякого тела невозможно. Мы исследуем важный вопрос «воплощенного» интеллекта позже, а пока вернемся к нашему определению. Пока достаточно сказать, что под интеллектом человеческого уровня мы подразумеваем способность решать задачи, учиться и действовать эффективно, по-человечески в различных ситуациях.

А пока у роботов хватает и собственных проблем. До сих пор ни один из них не стал особенно умным даже в узком смысле, и только самые продвинутые способны кое-как передвигаться и автономно манипулировать объектами. В настоящее время роботы не умнее тех, кто ими управляет.

Итак, долго ли осталось ждать появления ИИ человеческого уровня? Некоторые специалисты, с которыми мне довелось общаться, считают, что это может произойти уже к 2020 г. Однако в целом недавние опросы показывают, что компьютерщики и профессионалы в других, связанных с ИИ областях (таких как инженерное дело, робототехника и нейробиология), более осторожны в оценках. Они считают, что с вероятностью более 10 % ИИ человеческого уровня будет создан до 2028 г., а с более чем 50 % — до 2050 г. Вероятность же того, что это событие произойдет до конца текущего столетия, превышает 90 %.

Более того, специалисты утверждают, что первые ИИ соответствующего уровня разработают военные или крупный бизнес; у проектов академической науки и небольших организаций шансов на это намного меньше. Что касается соображений за и против, то ничего неожиданного эксперты не говорят: работы по созданию ИИ человеческого

уровня принесут нам громадную пользу и одновременно породят угрозу страшных катастроф, включая такие, от которых человечество уже не сможет оправиться.

Величайшие катастрофы, как мы уже говорили в главе 1, грозят нам после преодоления моста между ИИ человеческого уровня и суперинтеллектом ИСИ. При этом временной промежуток между появлением УЧИ и ИСИ может оказаться совсем небольшим. Следует отметить, однако, что если в профессиональной ИИ-среде риски, связанные с сосуществованием человечества на одной планете со сверхразумным ИИ, рассматриваются в высшей степени серьезно, то в публичном дискурсе эти вопросы практически отсутствуют. Почему?

Причин несколько. Чаще всего обсуждение опасностей, которые создает ИИ для человечества, проходит достаточно поверхностно и не отличается ни широтой, ни глубиной. Мало кто дает себе труд как следует разобраться в вопросе. Конечно, в Кремниевой долине и в академических кругах эти проблемы серьезно прорабатываются, но результаты почти не выходят за пределы узкого круга специалистов. Печальнее всего то, что эти результаты почти не замечают технические журналисты. Когда мрачные прогнозы в очередной раз поднимают головы, многие блогеры, редакторы и технари рефлексивно отбрасывают их прочь со словами: «О нет, только не это! Не нужно снова про Терминатора! Разве недостаточно страшных прогнозов мы слышали в свое время и от луддитов[5], и от прочих пессимистов?» Такая реакция обусловлена обычной ленью, что ясно видно по слабости выдвигаемых контраргументов. Неудобные факты о рисках, связанных с ИИ, не настолько привлекательны и доступны, как темы, составляющие обычно хлеб технической журналистики: двухъядерные 3D-процессоры, емкостные сенсорные экраны и очередные модные новинки программного рынка.

Кроме того, мне кажется, что популярность темы ИИ в мире развлечений стала своеобразной прививкой, не позволяющей нам серьезно рассматривать эту же тему в другой, серьезной категории катастрофических рисков. Несколько десятилетий сюжет о том, как искусственный интеллект (обычно в форме человекоподобных роботов, но иногда и в более экзотическом виде, к примеру, светящегося красного глаза) стирает человечество с лица земли, был основой множества популярных фильмов, романов и видеоигр. Представьте себе, что официальная медицина выпустила бы серьезное предупреждение касательно вампиров (вовсе не ироничное, как недавно о зомби[6]). Но вампиры в последние

годы доставили нам столько радости и развлечений, что после такого предупреждения долго звучал бы смех, и только потом появились бы осиновые колья. Может быть, в настоящее время мы и с ИИ переживаем что-то подобное, и только трагедия или предсмертные переживания способны разбудить нас.

Еще одно объяснение, по которому ИИ как причина вымирания человечества нечасто рассматривается всерьез, обусловлено, возможно, одной из психологических «мертвых зон» человека — когнитивным искажением — ловушкой на пути нашего мышления. Американские психологи еврейского происхождения Амос Тверски и Даниэль Канеман начали работу над этой темой в 1972 г. Их базовая идея состоит в том, что мы, люди, принимаем решения нерационально. Такое наблюдение само по себе не заслуживает Нобелевской премии (а Канеман был удостоен Нобелевской премии в 2002 г.); главное в том, что наша иррациональность подчиняется научным моделям. При принятии быстрых решений, очень полезных и даже необходимых в ходе биологической эволюции, мы постоянно пользуемся ментальными уловками, известными под общим названием эвристики. Одна из таких уловок — делать далеко идущие выводы (часто слишком далеко идущие) из собственного опыта.

Представьте, к примеру, что вы находитесь в гостях у друга, и в его доме вдруг вспыхивает пожар. Вам удается спастись, а на следующий день вы принимаете участие в опросе на тему гибели людей от несчастных случаев. Кто упрекнет вас в том, что вы укажете «пожар» в качестве самой частой или наиболее вероятной причины таких смертей? На самом же деле в США пожары редко приводят к гибели людей и в списке причин находятся намного ниже падений, транспортных происшествий и отравлений. Но вы, выбирая пожар, демонстрируете так называемую ошибку доступности, то есть тенденцию оценивать вероятность по доступным примерам. Ваш недавний опыт влияет на ваш выбор, делая его иррациональным. Но не расстраивайтесь — так происходит со всеми, да и психологических искажений, аналогичных ошибке доступности, существует больше десятка.

Возможно, именно ошибка доступности не позволяет нам прочно связать искусственный интеллект с исчезновением человечества. Мы не пережили ни одного сколько-нибудь серьезного происшествия, причиной которого стал бы ИИ, тогда как другие возможные причины гибели человечества «засветились» уже достаточно сильно. Все мы слышали о супервирусах вроде ВИЧ, вирусе атипичной пневмонии или испанки 1918 г. Мы видели результат воздействия ядерного оружия на многолюдные

города. Нас пугают геологические свидетельства падения астероидов размером с Техас в доисторические времена. А катастрофы на АЭС Тримайл-Айленд (1979 г.), в Чернобыле (1986 г.) и на Фукусиме (2011 г.) наглядно демонстрируют, что даже самые болезненные уроки приходится усваивать вновь и вновь.

Искусственный интеллект пока не входит в список экзистенциальных угроз человечеству; по крайней мере, мы пока не воспринимаем его в таком качестве. Опять же, наше отношение изменится после какого-нибудь серьезного происшествия, как события 11 сентября 2001 г. прочно внедрили в наше сознание представление о том, что самолет тоже может быть оружием. Та террористическая атака произвела революцию в системе безопасности воздушных перевозок и породила новую бюрократическую машину, которая обходится США в $44 млрд в год, — министерство внутренней безопасности. Но неужели для того, чтобы усвоить следующий урок, необходима катастрофа, связанная с ИИ? Надеюсь, что нет, поскольку с подобными катастрофами связана одна серьезная проблема. Они не похожи на крушения самолетов, ядерные или любые другие техногенные катастрофы; исключение, может быть, составляют нанотехнологии и катастрофы, связанные с ними. Дело в том, что человечество с высокой вероятностью не сможет оправиться после первого же подобного события.

Есть еще один принципиальный момент, в котором вышедший из-под контроля ИИ отличается от прочих техногенных происшествий. Ядерные электростанции и самолеты — оружие одноразового действия; событие происходит, и вы начинаете разбираться с последствиями. В настоящей ИИ-катастрофе действует умная программа, которая совершенствует сама себя и очень быстро воспроизводится. Она может существовать вечно. Как можем мы остановить катастрофу, если вызвавшая ее причина превосходит нас в сильнейшем качестве — интеллекте? И как можно разобраться с последствиями катастрофы, которая, раз начавшись, может продолжаться до бесконечности?

Еще одна причина примечательного отсутствия ИИ в дискуссиях об экзистенциальных угрозах состоит в том, что в темах, имеющих отношение к ИИ, доминирует сингулярность.

Слово «сингулярность» в последнее время модно употреблять по поводу и без повода, хотя у этого слова несколько разных значений, которые к тому же часто подменяют друг друга. В частности, известный изобретатель и писатель Рэй Курцвейл активно продвигает идею сингулярности, определяя ее как «исключительный» период времени (который начнется приблизительно в 2045 г.), после которого ход

технического прогресса необратимо изменит человеческую жизнь. Разум станет в основном компьютеризированным и в триллионы раз более мощным, чем сегодня. С сингулярности начнется новая эпоха в истории человечества, когда будет решена большая часть самых серьезных наших проблем, таких как голод, болезни и даже смертность.

Искусственный интеллект — медийная звезда под названием Сингулярность, но нанотехнологии играют в нем важную роль второго плана. Многие специалисты предсказывают, что искусственный сверхразум резко ускорит развитие нанотехнологий, потому что без труда решит нерешаемые на сегодняшний день проблемы. Некоторые считают, что лучше бы первым появился ИСИ, поскольку нанотехнологии — слишком капризный инструмент, чтобы доверять его нашим слабеньким мозгам. Строго говоря, значительная часть достижений, которые принято ожидать от сингулярности, обусловлена вовсе не искусственным интеллектом, а нанотехнологиями. Конструирование на уровне атомов может, помимо всего прочего, даровать нам бессмертие (для этого достаточно устранить на клеточном уровне эффекты старения), погружение в виртуальную реальность (результат будет обеспечиваться нанороботами, которые возьмут на себя сенсорные сигналы организма), а также нейронное сканирование и загрузку сознания в компьютер.

Однако, возражают скептики, вышедшие из-под контроля нанороботы, способные к тому же бесконечно воспроизводить себя, могут превратить нашу планету в «серую слизь». Проблема «серой слизи» — самая темная сторона нанотехнологий. Но почти никто не говорит об аналогичной проблеме ИИ — об «интеллектуальном взрыве», при котором развитие машин умнее человека запустит процесс гибели человечества. Это одна из многочисленных оборотных сторон сингулярности — одна из многих, о которых мы слишком мало знаем и слышим. Этот недостаток информации может объясняться тем, что я называю проблемой двух минут.

Я прослушал лекции о сверхразуме десятков ученых, изобретателей и специалистов по этике. Большинство из них считают его появление неизбежным и радуются благодеяниям, которыми осыплет нас гений ИСИ. Затем, чаще всего в последние две минуты выступления, эксперт замечает, что если ИИ неправильно управлять, то он, вероятно, может уничтожить человечество. Аудитория при этом нервно покашливает и с нетерпением ждет возвращения к хорошим новостям.

У писателей наблюдается два подхода к приближающейся технологической революции. Один подход порождает книги вроде курцвейловской «Сингулярность рядом». Цель авторов таких книг —

заложить теоретические основы в высшей степени позитивного будущего. Если бы в таком будущем произошло что-то плохое, вести об этом потонули бы в веселом гомоне оптимизма. Второй подход к проблеме представляет книга Джеффри Стибела «Создан для мысли». Это взгляд на технологическое будущее сквозь призму бизнеса. Стибел убедительно доказывает, что Интернет представляет собой все более усложняющийся мозг с множеством связей, и разработчикам веб-проектов следует это учитывать. Книги, подобные книге Стибела, пытаются научить предпринимателей забрасывать свои сети между интернет-тенденциями и потребителями и извлекать из глубин Интернета полные ведра денег.

Большинство теоретиков в области технологий и писателей упускает из виду не столь блестящий третий вариант, и я своей книгой постараюсь восполнить этот пробел. Я попытаюсь доказать, что завершающей фазой работ по созданию сначала умных машин, затем машин умнее человека, станет не их интеграция в нашу жизнь, а их победа над нами и завоевание Земли. В погоне за ИСИ исследователи получат разум более мощный, нежели их собственный; они не смогут ни контролировать его, ни даже до конца понять.

Мы знаем на опыте, что происходит, когда технически развитая цивилизация встречается с менее развитой: Христофор Колумб против Тиано, Писарро против Великого Инки, европейцы против американских индейцев.

Приготовьтесь к следующему столкновению цивилизаций — мы с вами против искусственного суперинтеллекта.

Возможно, мудрецы от технологий уже рассмотрели все темные стороны ИИ, но считают, что катастрофический исход слишком маловероятен, чтобы об этом стоило тревожиться. Или, наоборот, понимают неизбежность такого исхода, но считают, что ничто не в состоянии его предотвратить. Известный разработчик ИИ Бен Гёрцель (о его планах касательно ИСИ мы будем говорить в главе 11) рассказал мне, что мы не сможем придумать способа защиты от продвинутого ИИ, пока не накопим достаточно опыта общения и работы с ним. Курцвейл, теории которого мы разберем в главе 9, давно говорит примерно о том же: создание сверхразума и его интеграция с человеком будет проходить постепенно, и мы успеем научиться всему необходимому по ходу дела. И тот и другой утверждают, что *истинные* опасности ИИ невозможно увидеть из сегодняшнего дня. Иными словами, живя в век конных повозок, невозможно понять, как следует управлять автомобилем на обледенелой дороге. Так что расслабьтесь, мы во всем разберемся, когда придет время.

Для меня проблема такого подхода заключается в том, что, хотя обладающие суперинтеллектом машины спокойно могут стереть человечество с лица земли или просто сделать нас ненужными, мне представляется, что серьезную опасность для нас могут представлять и те ИИ, с которыми мы встретимся на пути к суперинтеллекту. То есть мама-медведица, конечно, представляет опасность для пикника на поляне, но не стоит сбрасывать со счетов и ее веселого медвежонка. Более того, градуалисты — сторонники теории постепенности — уверены, что скачок от основы (интеллекта человеческого уровня) к суперинтеллекту может занять несколько лет, а то и десятилетий. Такое развитие событий подарило бы нам определенный период мирного сосуществования с умными машинами, в течение которого мы успели бы научиться с ними взаимодействовать. Тогда их более продвинутые потомки не застали бы нас врасплох.

Но невозможно гарантировать, что все будет именно так. Скачок от интеллекта человеческого уровня к суперинтеллекту благодаря положительной обратной связи (через самосовершенствование) может произойти стремительно. В этом сценарии ИИ человеческого уровня совершенствует свой интеллект так быстро, что превращается в суперинтеллект за несколько недель, дней или даже часов, а не за месяцы и годы. В главе 1 описана как вероятная скорость, так и вероятный результат такого «жесткого старта». Вполне возможно, что в этом переходе не будет ничего постепенного.

Не исключено, что Гёрцель и Курцвейл правы — позже мы познакомимся с аргументами градуалистов поподробнее. Но прямо сейчас я хочу донести до вас кое-какие важные и тревожные мысли, вытекающие из сценария Busy Child.

Ученые-компьютерщики, особенно те, кто работает на оборонные и разведывательные ведомства, считают необходимым как можно быстрее разработать ИИ человеческого уровня, ведь альтернативы (к примеру, ситуация, когда китайское правительство создаст его первым) пугают их больше, чем поспешность в разработке собственного ИИ. Возможно, они спешат еще и потому, что ИИ необходим для лучшего управления другими капризными технологиями, появление которых ожидается в этом столетии. Например, нанотехнологий. Возможно, они не остановятся, чтобы обдумать тщательнейшим образом ограничения к самосовершенствованию. А совершенствующийся без всяких ограничений искусственный интеллект вполне может проскочить промежуток между обычным ИИ и ИСИ в варианте жесткого старта и дать толчок «интеллектуальному взрыву».

Мы не можем заранее знать, как поступит разум более мощный, чем наш. Мы можем лишь предполагать, какие способности такой интеллект сможет при желании против нас применить (к примеру, он может многократно дублировать себя, чтобы создать целую команду ИСИ для решения задач и одновременной проработки множества стратегических вопросов, связанных с освобождением и выживанием, и действовать при этом вне рамок морали). Наконец, было бы разумно считать, что первый ИСИ не будет настроен по отношению к нам ни дружественно, ни враждебно; ему будут попросту безразличны наше счастье, здоровье и благополучие.

Можем ли мы прогнозировать потенциальный риск от ИСИ? Уоррен Льюис в книге «Технологический риск» разбирает категории риска и классифицирует их по тому, насколько сложно такие риски учесть. Легче всего просчитываются риски событий, происходящих с высокой вероятностью и возможными серьезными последствиями (к примеру, поездка за рулем машины из одного города в другой). Здесь очень много информации, на которую можно опереться в расчетах. События с низкой вероятностью и серьезными последствиями (к примеру, землетрясения) происходят реже, поэтому предвидеть их сложнее. Но последствия таких событий настолько серьезны, что заниматься их прогнозированием имеет смысл.

Существуют также события, вероятность которых низка, потому что никогда прежде ничего подобного не происходило, но последствия которых очень серьезны. Хороший пример — резкое изменение климата в результате загрязнения окружающей среды. До 16 июля 1945 г. — первого испытания бомбы в районе Аламогордо на полигоне Уайт-Сэндз (штат Нью-Мексико) — еще одним таким событием был атомный взрыв. Технически именно к этой категории относится и искусственный суперинтеллект. Опыт в данном случае ничего не подсказывает. Невозможно просчитать вероятность этого события при помощи традиционных статистических методов.

Я уверен, однако, что при нынешних темпах развития ИИ изобретение сверхразума относится скорее к первой категории — это событие с высокой вероятностью и высоким риском. Более того, даже если бы вероятность этого события была низкой, масштабы риска должны были бы выдвинуть его на передний план нашего внимания.

Иначе говоря, я считаю, что Busy Child появится очень скоро.

Страх проиграть разуму, превосходящему человеческий, возник давно, однако только в начале этого века в Кремниевой долине был проведен

хитроумный эксперимент, связанный с ИИ, и его результаты мгновенно породили в Интернете легенду.

Слух выглядел примерно так: одинокий гений заключил несколько пари с высокими ставками на игру по сценарию, который он назвал «ИИ в ящике». В ходе эксперимента роль ИИ играл этот самый гений. В роли Привратника выступали миллионеры, сделавшие себе состояние на всевозможных интернет-проектах; они по очереди играли роль создателя ИИ, перед которым стоит задача охранять и удерживать «взаперти» искусственный интеллект. ИИ и Привратник общались в онлайновом чате. Утверждалось, что при помощи одной только клавиатуры человек, игравший роль ИИ, каждый раз умудрялся улизнуть — и таким образом выиграл все пари. К тому же, что еще важнее, он доказал свое утверждение. Если он, обычный человек, сумел посредством слов открыть заключенному в «ящике» разуму путь на свободу, то любой ИСИ, который будет в десятки или в сотни раз умнее, тоже сможет это сделать, причем гораздо быстрее, что приведет, скорее всего, к гибели человечества.

По слухам, после эксперимента гений ушел в подполье. Сам эксперимент, а также статьи и монографии на тему ИИ, написанные этим гением, принесли ему такую известность, что у него даже появилось сообщество поклонников, общение с которыми, однако, никак не приближало его к цели, с которой он начинал эксперимент «ИИ в ящике», — спасти человечество.

В результате он исчез из поля зрения и поклонников, и журналистов. Но я, разумеется, захотел с ним поговорить

Глава 3
Взгляд в будущее

ИИ человеческого уровня по природе своей очень-очень опасен. И понять эту проблему не особенно трудно. Не нужно быть ни сверхумным, ни сверхинформированным, ни даже сверхинтеллектуально-честным, чтобы понять эту проблему.

Майкл Вассар, президент Исследовательского института машинного интеллекта

Я определенно считаю, что создавать искусственный интеллект человеческого уровня необходимо с максимальным вниманием. В этом случае максимальное внимание означает гораздо более скрупулезную осторожность, чем была бы необходима в работе с вирусом Эбола или плутонием.

Майкл Вассар — подтянутый невысокий мужчина лет тридцати. Он обладает научными степенями в области биохимии и бизнеса и прекрасно разбирается в разных способах уничтожения человечества, так что такие слова, как «вирус Эбола» и «плутоний», слетают с его языка без малейших колебаний или иронии. Одна из стен его квартиры в стильном кондоминиуме представляет собой одно большое окно, в котором, как в раме, висит красный мост, соединяющий Сан-Франциско с Оклендом. Это не знаменитый элегантный мост Золотые Ворота, который находится внутри города. Этот красный мост называют его безобразным сводным братом. Вассар рассказал мне, что известны случаи, когда люди, решившие покончить с собой, проезжали через этот мост, чтобы добраться до другого, красивого.

Вассар посвятил свою жизнь тому, чтобы предотвратить самоубийство глобального масштаба. Он является президентом Исследовательского института машинного интеллекта — мозгового треста со штаб-квартирой в Сан-Франциско, основанного для противодействия гибели человеческой расы от рук, вернее, от байтов искусственного интеллекта. На сайте института размещаются серьезные статьи об опасных сторонах ИИ, а раз в

год институт собирает представительную конференцию по сингулярности[7]. За два дня работы конференции программисты, нейробиологи, ученые, предприниматели, специалисты по этике и изобретатели обсуждают положительные и отрицательные стороны текущей ИИ-революции. Организаторы равно приветствуют выступления сторонников и противников этого процесса, приглашают к себе и тех, кто уверен, что никакой сингулярности не будет, и тех, кто считает институт апокалиптической техносектой.

Вассар улыбается при мысли о секте.

> Люди, которые приходят работать в наш институт, не похожи на сектантов, скорее наоборот. Как правило, они осознают исходящие от ИИ опасности раньше, чем узнают о существовании института.

Я, к примеру, не знал о существовании Исследовательского института машинного интеллекта, пока не услышал об эксперименте «ИИ в ящике». О нем рассказал мне друг, но в его пересказе история вышла сильно искаженной, особенно часть об одиноком гении и его оппонентах-миллионерах. Заинтересовавшись темой, я нашел источник этой легенды на сайте института и выяснил, что автор эксперимента — Елиезер Юдковски — был одним из основателей института (тогда он назывался Институтом сингулярности) вместе с предпринимателями Брайаном и Сабиной Эткинс. Несмотря на приписываемую ему замкнутость, мы с Юдковски обменялись электронными письмами, и он подробно рассказал мне об эксперименте.

Пари между ИИ, за который играл Юдковски, и Привратниками, задача которых состояла в контроле над ИИ, заключались максимум на несколько тысяч долларов, но никак не на миллионы. Всего состоялось пять игр, «ИИ в ящике» выиграл три из них. Это означает, что ИИ, как правило, умудрялся выбраться из «ящика», но все же не каждый раз.

Отчасти слухи об эксперименте «ИИ в ящике» все же были верны: Юдковски и правда живет замкнуто, бережет время и не открывает своего адреса. Я напросился в гости к Майклу Вассару потому, что был потрясен и обрадован тем, что в природе существует некоммерческая организация, призванная бороться с опасностями ИИ, и что умные молодые люди посвящают свою жизнь решению этой проблемы. Но я к тому же надеялся, что разговор с Вассаром поможет мне все же добраться до Юдковски.

Прежде чем погрузиться с головой в дело разоблачения опасностей

искусственного интеллекта, Вассар успел получить степень магистра бизнес-администрирования и заработать денег, став одним из основателей интернет-компании музыкального лицензирования Sir Groovy. Эта компания выступает посредником между теле- и кинопродюсерами и независимыми звукозаписывающими компаниями, предлагающими свежие записи не самых известных и потому не самых дорогих музыкантов. До 2003 г. Вассар колебался, не заняться ли ему опасностями нанотехнологий, но в том году он встретился с Елиезером Юдковски, работы которого читал в Интернете уже много лет. Он узнал о существовании Исследовательского института машинного интеллекта — и об угрозе более неотвратимой и опасной, чем нанотехнологии: опасности со стороны искусственного интеллекта.

> Елиезер убедил меня, что ИИ человеческого уровня может быть создан в короткие сроки при относительно небольшом финансировании, и риск глобальной катастрофы по вине такого ИИ очень встревожил меня. У меня не было никаких убедительных причин считать, что УЧИ не может появиться в течение, скажем, следующих двадцати лет.

Это раньше, чем предсказанное появление серьезных нанотехнологий. Да и дополнительных расходов на разработку ИИ будет, скорее всего, куда меньше. Так что Вассар изменил курс.

При встрече я признался, что раньше пренебрежительно относился к мысли о том, что ИИ человеческого уровня может быть создан небольшой группой специалистов со скромным финансированием. Во всех опросах специалистов такой вариант появления ИИ фигурировал как маловероятный.

Так может ли, к примеру, искусственный интеллект создать «Аль-Каида»? Или Революционные вооруженные силы Колумбии? Или какая-нибудь секта вроде «Аум Синрикё»?

Вассар не считает, что ИИ родится в террористической ячейке. Слишком велик интеллектуальный разрыв.

> Известно, что плохие парни, которые хотят уничтожить мир, обычно не слишком умны. Вы знаете, у людей, которые и правда хотят уничтожить мир, для реализации чего бы то ни было не хватает способностей к долгосрочному планированию.

Но как же «Аль-Каида»? Разве для реализации террористических атак, включая и теракт 11 сентября 2001 г., не нужны были развитое воображение и тщательное планирование?

> По масштабу подготовка к теракту не сравнима с разработкой ИИ человеческого уровня. Написание программы, которая хоть что-нибудь делала бы лучше человека, не говоря уже о реализации всего набора свойств и способностей ИИ, потребовало бы на несколько порядков больше таланта и организованности, чем демонстрирует "Аль-Каида" всем перечнем своих злодеяний. Если бы создать ИИ было настолько просто, кто-нибудь поумнее "Аль-Каиды" давно бы уже сделал это.

Но как насчет правительств таких стран, как Северная Корея или Иран?

> В практическом аспекте качество науки, которую могут обеспечить эти нехорошие режимы, отвратительно. Единственное исключение — это нацисты. Но, как бы это сказать, если нацизм вновь возникнет, у нас будут серьезнейшие проблемы — с искусственным интеллектом или без него.

Здесь я с ним не согласен, хотя и не в отношении нацистов. Иран и Северная Корея сумели найти высокотехнологичные способы шантажировать весь остальной мир, с одной стороны, разработкой ядерного оружия, с другой — межконтинентальными ракетами. Так что я не стал бы вычеркивать их из короткого списка потенциальных создателей ИИ, ведь они уже не раз выразили полное пренебрежение международным сообществом. Кроме того, если ИИ человеческого уровня способна создать небольшая группа специалистов, то ясно, что финансировать такую группу может любое государство-изгой.

Говоря о небольших группах специалистов, Вассар имел в виду компании, работающие открыто. Но мне приходилось слышать и о других фирмах — так называемых стелс-компаниях, которыми владеют частные лица, тайком нанимающие людей и никогда не выпускающие пресс-релизов и вообще никак не рассказывающие о том, чем занимаются. Если говорить о разработке ИИ, то единственная причина, по которой компания может держаться в тени, состоит в том, что ее специалисты добились какого-то

серьезного прорыва и не хотят раскрывать секреты конкурентам. Искать такие компании по определению трудно, хотя слухами о них земля полнится. Так, основатель PayPal Питер Тиль финансирует три стелс-компании в области искусственного интеллекта.

Компании, работающие в стелс-режиме, однако, выглядят иначе и встречаются чаще. Они ищут финансирование и даже стремятся к публичности, но не раскрывают своих планов. Петер Восс, известный новатор ИИ и автор технологии распознавания речи, тоже разрабатывает ИИ в своей компании Adaptive AI. Он публично заявил, что ИИ человеческого уровня может быть создан в ближайшие десять лет. Но как именно и за счет чего, он не говорит.

Со стелс-компаниями связана еще одна сложность. Небольшая мотивированная группа вполне может существовать внутри крупной известной компании. Что вы думаете, к примеру, насчет Google? Почему бы богатой мегакорпорации не взяться за поиски «святого Грааля»?

На мой вопрос на одной из ИИ-конференций директор Google по исследованиям Питер Норвиг, один из авторов классического учебника «Искусственный интеллект. Современный подход»[8], сказал, что Google не занимается разработкой ИИ человеческого уровня. Он сравнил эту задачу с планом межпланетных пилотируемых полетов NASA — потому что такого плана не существует. Тем не менее агентство будет и дальше работать над отдельными направлениями, необходимыми для космических путешествий, — робототехникой и ракетной техникой, астрономией и т. п. — и когда-нибудь в один прекрасный момент все детали головоломки сложатся и полет к Марсу обретет реальные очертания.

Так и в процессе работы над отдельными специализированными ИИ-проектами проделывается куча подготовительной работы. Машины учат таким интеллектуальным занятиям, как поиск, распознавание речи, обработка естественного языка, визуальное восприятие, поиск информации, и многому другому. По отдельности это всего лишь надежные мощные инструменты, и они быстро развиваются год от года. Вместе они продвигают компьютерные науки вперед и, безусловно, идут в копилку будущих систем ИИ человеческого уровня.

Норвиг сказал мне, что у Google нет ИИ-программы. Но сравните это заявление с тем, что сказал его босс Ларри Пейдж, один из основателей Google, на лондонской конференции под названием Zeitgeist'06:

> Людям все время кажется, что наша поисковая система уже готова. Это не так. Сделано пока, вероятно, не больше 5 %. Мы

> хотим создать идеальный поисковый движок, способный понять что угодно… кое-кто назвал бы это искусственным интеллектом… Идеальный поисковый движок понимал бы все на свете. Он понимал бы все, о чем вы его спросили, и мгновенно выдавал бы в точности то, что нужно… Вы могли бы спросить у него: "О чем бы мне спросить Ларри?" — и он дал бы верный ответ.

Лично мне кажется, что говорит он об ИИ человеческого уровня.

IBM ведет разработку такого ИИ в рамках нескольких хорошо финансируемых проектов. DAPRA, Агентство перспективных проектов Минобороны США, кажется, финансирует все ИИ-проекты, которыми я интересовался. Так неужели Google не разрабатывает ИИ? Когда я задал этот вопрос Джейсону Фриденфелдсу из пресс-службы Google, он написал в ответ:

> …нам пока слишком рано рассуждать на подобные темы, это все еще далеко впереди. В целом мы больше сосредоточены на практических технологиях машинного обучения, таких как машинное зрение, распознавание речи и машинный перевод, что в основном сводится к построению статистических моделей, отражающих закономерности, — ничего близкого к представлению об ИИ человеческого уровня как о "думающей машине".

Но мне представляется, что цитата из высказывания Пейджа проливает больше света на отношение Google к проблеме ИИ, чем ответ Фриденфелдса. В частности, она помогает объяснить эволюцию Google от компании мечтателей и бунтарей, какой она была в 1990-е гг. (вспомните нашумевший тогдашний слоган компании: «Не будь злом»), до сегодняшней по-оруэлловски непрозрачной громадины, занятой сбором личных данных.

Политика компании позволяет ей передавать ваши личные данные всевозможным сервисам Google, включая Gmail, Google+, YouTube и др. С кем вы знакомы, куда вы ходите, что покупаете, с кем встречаетесь, как ищете информацию и куда заходите в Сети, — Google интересует все. Любая крупинка информации представляет ценность. Заявленная цель: улучшить вашу жизнь как пользователя, сделав систему поиска практически всеведущей в отношении *вас самих*. Параллельная цель —

сформировать для вас не только рекламный пакет, но и круг новостей, видеороликов и музыки, которые вы будете потреблять, и автоматически сделать вас мишенью маркетинговых кампаний. Даже фирменные автомобили с камерами, снимающие «уличные виды» для Google Maps, являются частью этого плана — в течение трех лет Google использовал свой фотографирующий автопарк для перехвата данных из частных сетей Wi-Fi в США и других странах. Их интересовала любая информация: пароли, история пользования Интернетом, личные адреса электронной почты, все что угодно.

Ясно, что компания поставила верных ей прежде клиентов на место, и место это оказалось далеко не первым. Поэтому казалось сомнительным, что планы Google не включают в себя ИИ.

А примерно через месяц после моего обмена письмами с Фриденфелдсом газета *The New York Times* разразилась статьей о Google X.

Google X — это стелс-компания. Эту секретную лабораторию в Кремниевой долине первоначально возглавлял специалист по ИИ и разработчик роботизированного автомобиля Google Себастьян Трун. Нацелена она на стопроцентно фантастические проекты, такие как космический лифт: предполагается, что он устремится в космос и облегчит человечеству освоение Солнечной системы. Помимо прочих в штат стелс-компании входит Эндрю Ын, робототехник мирового класса и бывший директор Лаборатории искусственного интеллекта Стэнфордского университета.

В конце 2012 г. Google пригласил уважаемого изобретателя и писателя Рэя Курцвейла на роль главного инженера компании. В главе 9 будет рассказано, что Курцвейл известен многочисленными достижениями в области ИИ и продвигает исследование мозга как самый короткий путь разработки искусственного интеллекта.

Не нужно прибегать к помощи Google glasses[9], чтобы увидеть очевидное: если на Google работают по крайнее мере двое из известнейших мировых специалистов по ИИ, да плюс к тому Рэй Курцвейл[10], то создание ИИ человеческого уровня имеет среди перспективных проектов компании высокий приоритет.

Стремясь получить конкурентное преимущество на рынке, Google X и другие стелс-компании, вполне возможно, создадут полноценный искусственный интеллект вне публичного поля.

Итак, похоже, что стелс-компании осуществляют скрытный обход на пути к ИИ человеческого уровня. Однако Вассар считает, что самый

быстрый путь будет весьма публичным и дорогостоящим. Это метод обратного проектирования, то есть построение искусственного интеллекта по образцу работающего человеческого мозга с использованием как искусного программирования, так и решения в лоб. Под «решением в лоб» подразумевается количественное накопление продвинутой техники — блоков с самыми быстрыми процессорами, петабайт памяти и т. п.

> Экстремальный вариант решения в лоб — биологические исследования, — *сказал мне Вассар*. — Продолжая анализировать при помощи машин биологические системы, разбираться в обмене веществ, в сложных взаимоотношениях внутри биологических систем, со временем люди накопят огромное количество информации о том, как нейроны обрабатывают информацию. В дальнейшем эти данные можно использовать для разработки ИИ.

Получается примерно так: в основе работы разума лежат биохимические процессы, которые протекают в нейронах, синапсах и дендритах. При помощи различных технологий сканирования мозга, включая позитронно-эмиссионную и функциональную магнитно-резонансную томографию, а также неврологических зондов, размещаемых как внутри, так и снаружи черепа, ученые определяют, как конкретные нейроны и кластеры нейронов участвуют в мыслительном процессе. Затем они реализуют каждый из этих процессов при помощи компьютерной программы или алгоритма.

Этим занимается новая область науки — вычислительная нейробиология. Один из мировых лидеров в этой области — доктор Ричард Грейнджер, директор Лаборатории инженерии мозга Дартмутского университета — создал алгоритмы, работа которых имитирует работу нейронных контуров человеческого мозга. Он даже запатентовал чрезвычайно производительный компьютерный процессор, построенный на принципах работы этих контуров. Когда такой процессор доберется до рынка, мы станем свидетелями гигантского скачка в машинном распознавании объектов, поскольку компьютеры будут это делать в точности так, как наш мозг.

Конечно, остается множество других контуров мозга, которые нужно будет исследовать и скопировать. Но стоит нам создать алгоритмы для всех процессов мозга — и наши поздравления! Вот и готовый мозг! Или нет? Может, и нет. Возможно, в результате получится лишь машинный эмулятор

мозга. Надо сказать, в отношении ИИ это серьезнейший вопрос. К примеру, думает ли шахматная программа?

Когда компания IBM бралась за разработку Deep Blue, обыгрывающего лучших шахматистов мира, его не программировали играть в шахматы, как чемпион мира Гарри Каспаров, только еще лучше. Авторы программы просто не знали, как это сделать. Виртуозная игра Каспарова опиралась на его громадный опыт, на множество сыгранных им партий и еще большее количество изученных. Он собрал в своей голове огромную библиотеку дебютов, атак, маневров, блокад, ловушек, гамбитов и эндшпилей — настоящую энциклопедию стратегии и тактики. Он узнает позиции, видит закономерности, помнит и *думает*. Обычно Каспаров думает на три-пять ходов вперед, но это число может доходить до четырнадцати. Ни один современный компьютер на это не способен.

Поэтому IBM запрограммировала компьютер на анализ 200 млн позиций в секунду.

Deep Blue начинает с того, что делает гипотетический ход и оценивает все возможные ответные ходы Каспарова. На каждый такой ход он предлагает собственный гипотетический ответ и вновь оценивает все возможные ответные ходы Каспарова. Такое моделирование глубиной в два уровня называют двухслойным поиском — иногда Deep Blue доходит до шестого слоя. Это означает, что на каждый гипотетический ход рассматривается по шесть следующих «ходов» каждой стороны.

Затем компьютер возвращается к нетронутой доске и начинает оценивать другой гипотетический ход. Этот процесс повторяется много раз, для всех возможных ходов, и каждый из них оценивается в баллах в зависимости от того, берет ли этот ход фигуру, и насколько ценную, улучшает ли он в целом позицию игрока и насколько. В конце концов ход с максимальным рейтингом будет сделан.

Так думает ли Deep Blue?

Возможно. Но вряд ли кто-то будет спорить с тем, что думает он не так, как человек. И мало кто из специалистов сомневается в том, что с ИИ будет та же история. Каждый исследователь, занятый разработкой ИИ человеческого уровня, применяет собственный подход. Некоторые идут по чисто биологическому пути, пытаясь в точности имитировать мозг. Другие также опираются на биологию и принимают мозг за образец, но больше полагаются на безотказный инструментарий ИИ: способы доказательства теорем, поисковые алгоритмы и алгоритмы обучения, автоматизированную логику и т. п.

Мы рассмотрим некоторые из этих инструментов и увидим, что

человеческий мозг использует многие из тех же самых вычислительных методов, что и компьютеры. Но смысл в том, что мы не можем сказать заранее, будут ли компьютеры думать в нашем понимании и обретут ли они когда-нибудь что-нибудь вроде намерений (интенций) или сознания. Таким образом, говорят некоторые ученые, искусственный интеллект, эквивалентный человеческому интеллекту, невозможен.

Философ Джон Сёрль предложил для доказательства этого утверждения мысленный эксперимент, получивший название «китайской комнаты».

> Представьте, что человека, совершенно не говорящего и не понимающего по-китайски, заперли в комнате, полной коробок с карточками, на которых изображены китайские иероглифы (своеобразная база данных), и книгой инструкций на его родном языке о том, что нужно делать с этими иероглифами (программа). Представьте также, что люди вне комнаты передают этому человеку другие карточки с китайскими иероглифами, которые на самом деле (он этого не знает) представляют собой вопросы на китайском языке (входной сигнал). И представьте еще, что, следуя инструкциям (программе), человек в комнате может передавать обратно карточки с китайскими иероглифами, представляющие собой верные ответы на эти вопросы (выходной сигнал).

Человек в комнате корректно отвечает на вопросы, так что люди снаружи считают, что он может общаться на китайском. Тем не менее на самом деле он не понимает по-китайски ни слова. Подобно этому человеку, заключает Сёрль, компьютер никогда не научится по-настоящему думать и понимать. В лучшем случае обратное проектирование мозга позволит как следует отладить мимику. ИИ-системы дадут тот же механический результат.

Сёрль не одинок во мнении о том, что компьютеры никогда не научатся думать и не обретут сознания. Но у него много критиков с самыми разными претензиями. Некоторые недоброжелатели приписывают ему компьютерофобию. Если рассматривать в целом, то все в китайской комнате, включая человека, складывается в систему, которая очень убедительно «понимает» китайский язык. С такой позиции аргументы Сёрля замыкаются: ни одна из частей комнаты (компьютера) не понимает по-китайски, следовательно, и сам компьютер не может понимать по-

китайски.

Кроме того, аргументы Сёрля можно с той же легкостью применить и к человеку: у нас нет формального определения того, что представляет собой в реальности понимание языка — так как же мы можем утверждать, что человек «понимает» какой-то язык? Судя о том, имеется ли в данном случае понимание, мы можем опереться только на наблюдения. В точности как люди вне китайской комнаты Сёрля.

В любом случае что, собственно, такого выдающегося в мозговых процессах и даже в сознании? Тот факт, что сегодня мы не понимаем феномен сознания, не означает, что мы никогда его не поймем. Это не волшебство.

И все же я согласен и с Сёрлем, и с его критиками. Сёрль прав: машинный интеллект не будет похож на нас. Он будет полон вычислительных механизмов, которые никто как следует не понимает. К тому же компьютерные системы, созданные для ИИ человеческого уровня на основе так называемой «когнитивной архитектуры», могут оказаться слишком сложными, чтобы их мог понять один человек. Но критики Сёрля правы в том, что когда-нибудь УЧИ или ИСИ *мог бы научиться* мыслить, как мы, если бы мы довели его разработку до этого этапа.

Я не думаю, что до этого дойдет. Я считаю, что наше Ватерлоо состоится в обозримом будущем и участвовать в нем будет завтрашний ИИ и нарождающийся УЧИ, появление которого ожидается лет через десять-двадцать. Для выживания человечества, если оно возможно, необходимо, помимо всего прочего, при создании УЧИ встроить в него что-то похожее на совесть и человеческую отзывчивость, даже дружелюбие. Для этого необходимо как минимум до тонкостей понять умные машины, чтобы не было неожиданностей.

Вернемся ненадолго к одному известному определению так называемой «технологической сингулярности». В нем говорится о будущем, когда нам, людям, придется делить планету с другим, более мощным разумом. Рэй Курцвейл предполагает, что человечество сольется с машинами и таким образом уцелеет. Другие специалисты считают, что машины изменят и улучшат нашу жизнь, но сами мы останемся такими же, как прежде, просто людьми, а не человеко-машинными киборгами. Третьи, и я солидарен с ними, считают, что будущее принадлежит машинам.

Исследовательский институт машинного интеллекта был основан для того, чтобы сохранить наши человеческие ценности вне зависимости от формы, которую приобретут наследники.

Вассар в Сан-Франциско сказал мне:

> На кону сейчас передача человеческих ценностей преемникам человечества. А через них — и Вселенной.

В институте считают, что первый УЧИ, который выйдет на свободу, должен быть безопасен; тогда он сможет передать человеческие ценности преемникам человечества, какую бы форму они ни приняли. Если обеспечить безопасность УЧИ не удастся, ни сам человек, ни то, что мы ценим, не уцелеют. И на кону не только будущее Земли. Как сказал мне Вассар

> миссия нашего института — сделать так, чтобы технологическая сингулярность произошла в наилучших возможных условиях и принесла Вселенной наилучшее возможное будущее.

Как может выглядеть удачный для Вселенной исход дела?

Вассар пристально смотрел за окно на оживленное по случаю часа пик движение; машины только-только начали скапливаться на железном мосту в Окленд. Где-то там, за водной поверхностью, лежало будущее. В его воображении сверхразум уже вышел из-под контроля своих создателей. Он колонизировал сначала Солнечную систему, а затем и Галактику; теперь он занимался переформатированием Вселенной и вырастал в нечто настолько необычное, что сложно поддается человеческому пониманию.

В этом будущем, сказал Вассар, вся Вселенная становится единым компьютером, или единым сознанием; это настолько же опережает наши сегодняшние представления о мире, насколько космический корабль опережает червя. Курцвейл пишет, что это судьба Вселенной. Другие соглашаются, но считают, что, бездумно развивая ИИ, мы сделаем неизбежным как наше собственное уничтожение, так и уничтожение других существ, возможно, обитающих где-то далеко. Если ИСИ не будет ни любить нас, ни ненавидеть, точно так же он не будет ни любить, ни ненавидеть всех остальных существ во Вселенной. Неужели наше стремление построить ИИ человеческого уровня знаменует собой начало галактической эпидемии?

Покидая квартиру Вассара, я размышлял о том, что может помешать этому антиутопическому сценарию осуществиться. Что сможет остановить всеразрушающий искусственный разум человеческого или более высокого уровня? Более того, есть ли в этой антиутопической гипотезе слабые места?

Ну, вообще-то создатели ИИ могли бы сделать его «дружелюбным», чтобы то, что разовьется из первых образцов УЧИ, не уничтожило нас и другие существа во Вселенной. С другой стороны, не исключено, что мы ошибаемся насчет способностей и побудительных мотивов УЧИ и напрасно боимся, что он будет завоевывать Вселенную. Возможно, это ложная дилемма.

Может быть, ИИ никогда не сможет развиться в УЧИ. Есть серьезные основания считать, что эволюция ИИ произойдет иначе и более управляемым способом, чем мы думаем. Короче говоря, мне хотелось знать, что могло бы направить нас по более безопасному пути в будущее.

Я собирался спросить об этом у автора эксперимента «ИИ в ящике» Елиезера Юдковски. Мне говорили, что он не только предложил тот мысленный эксперимент, но и вообще знает о «дружелюбном» ИИ больше, чем кто бы то ни было другой на всем белом свете.

Глава 4
По сложному пути

Среди возможных катастроф ничто не сравнится с выпущенным в мир искусственным интеллектом, за исключением, может быть, нанотехнологии…

Елиезер Юдковски, научный сотрудник Исследовательского института машинного интеллекта

В Кремниевой долине четырнадцать «официальных» городов, в которых размещается двадцать пять математических и технических университетов и филиалов. Их выпускники пополняют ряды софтверных, полупроводниковых и интернет-фирм, составляющих на данный момент тот технологический центр, который возник примерно в этих же местах в начале XX в. и начался с развития радио. На Кремниевую долину приходится около трети всего венчурного капитала США. Здесь число технарей, причем самых высокооплачиваемых, на душу населения максимально. Значительное число американских миллиардеров и миллионеров считают Кремниевую долину своим домом.

Здесь, в эпицентре глобальной технологии, я ехал к дому Елиезера Юдковски самым старомодным способом — по письменным указаниям (хотя и во взятой напрокат машине, и в айфоне у меня имелись GPS-навигаторы). Юдковски, стремясь сохранить информацию о себе в тайне, прислал схему проезда по электронной почте, попросив при этом никому не открывать ни физический, ни электронный его адрес. Телефонный номер он мне сообщать не стал.

В свои тридцать три года Юдковски, один из основателей и сотрудник Института машинного интеллекта, написал об опасностях ИИ больше, чем кто бы то ни было. Больше десяти лет назад, когда Юдковски только начинал свою профессиональную деятельность, людей, посвятивших, как он, свою жизнь изучению опасностей искусственного интеллекта, практически не было. Он, конечно, не приносил никаких формальных клятв, но и тогда, и теперь избегает всякой деятельности, которая могла бы увести его внимание от главного вопроса. Он не пьет, не курит, не

употребляет наркотики. Редко появляется на публике и мало общается с коллегами. Несколько лет назад он отказался от чтения ради удовольствия. Он не любит давать интервью, а если уж дает, то предпочитает делать это по скайпу с лимитом времени в полчаса. Он атеист (среди специалистов по ИИ это скорее правило, чем исключение), так что ему не приходится тратить драгоценные часы в церкви. У него нет детей, хотя он их любит и считает родителей, которые не подписали своих детей на криоконсервацию[11], плохими родителями.

Но обратите внимание на парадокс. Для человека, вроде бы дорожащего тайной своей личной жизни, Юдковски выложил в Интернет слишком много информации о себе. После первых же попыток найти этого человека я выяснил, что в любом уголке Сети, где ведутся дискуссии о теории разума и грядущих катастрофах, никуда не деться от него самого и его самых сокровенных размышлений. Если вы задумываетесь о разрушительности ИИ, не забудьте освободить в своей жизни место под идеи Юдковски.

Благодаря его вездесущности я узнал, к примеру, что его младший брат Иегуда покончил с собой в родном Чикаго в возрасте девятнадцати лет. Горе Юдковски вылилось в онлайновую тираду, которая и сегодня, почти десять лет спустя, задевает за живое. Еще я узнал, что, бросив школу в восьмом классе, Юдковски самостоятельно изучил математику, логику, историю науки и вообще все, что показалось ему необходимым. Среди других освоенных им навыков — убедительная риторика и умение писать плотную и часто забавную прозу:

> Я большой поклонник музыки Баха и считаю, что лучше всего ее исполнять на электронных инструментах с тяжелым ритмом, в точности как планировал сам Бах.

Юдковски всегда спешит, потому что его работа жестко ограничена по срокам — ограничена тем днем, когда кто-нибудь создаст наконец ИИ человеческого уровня. Если исследователи построят свой ИИ с предохранителями, о которых говорит Юдковски, он может оказаться спасителем человечества, и не только человечества. Но если интеллектуальный взрыв все же произойдет и окажется, что Юдковски не удалось внедрить достаточно предохранителей, то и мы с вами, скорее всего, превратимся в серую слизь, а с нами и Вселенная. Эта идея находится в самом центре его собственной космологии.

Я приехал к нему, чтобы узнать побольше о дружественном ИИ

(термин, введенный Юдковски). По его мнению, дружественным можно назвать такой ИИ, который навсегда сохранит человечество и наши ценности. Он не станет уничтожать наш биологический вид или расползаться по Вселенной подобно пожирающей планеты космической чуме.

Но что такое дружественный ИИ? Как его создать?

Мне хотелось также услышать рассказ об эксперименте «ИИ в ящике», а особенно узнать, как мой собеседник играл роль искусственного интеллекта человеческого уровня, как он уговаривал Привратника выпустить его на волю. Я уверен, что когда-нибудь вы лично, или кто-то из ваших знакомых, или кто-то удаленный от вас на пару рукопожатий в самом деле окажется в кресле Привратника. Этому человеку необходимо знать заранее, чего можно ожидать в такой ситуации и как этому сопротивляться. Может быть, Юдковски это знает.

Жилище Юдковски располагается в концевой секции подковообразного двухэтажного дома с бассейном, искусственным водопадом во внутреннем дворике и садом вокруг. Его квартира просторна и безукоризненно чиста. Почетное место на островке в кухне занимают монитор и компьютер, рядом стоит единственный в квартире мягкий барный стул, с которого в окно виден дворик. Здесь хозяин квартиры пишет свои книги.

Это высокий человек ростом около 180 см, склонный к эндоморфизму (это значит, что он округл, но не жирен). Мягкие обезоруживающие манеры производят приятное впечатление по контрасту с резкими короткими электронными письмами, с которых начиналось наше знакомство.

Мы сидели на кушетках напротив друг друга. Я сказал Юдковски, что основной мой страх относительно ИИ связан с тем, что в мире не существует методик программирования таких туманных и сложных вещей, как мораль или дружелюбие. Так что мы получим машину, которая будет прекрасно решать задачи, обучаться, адаптироваться к обстановке и оперировать понятиями в рамках здравого смысла. Мы будем считать ее похожей на человека. Но это будет нашей трагической ошибкой

Юдковски согласился со мной.

> Если программисты окажутся хотя бы чуть-чуть некомпетентными и слегка небрежно отнесутся к конструированию ИИ, то получится, я уверен, нечто совершенно чуждое. И это по-настоящему пугает. Точно так же, если вы наберете правильно девять из десяти цифр моего телефонного

> номера, вас не обязательно соединят с человеком, который будет на 90 % похож на меня. Если при попытке сконструировать полноценный ИИ вы все сделаете на 90 % правильно, это не означает, что результат будет на 90 % хорош.

Мало того, он будет на 100 % плох. Автомобили существуют не для того, чтобы убить вас, провел аналогию Юдковски, но их потенциальная опасность — побочный продукт автоиндустрии. Точно так же будет с ИИ. Он вряд ли будет вас ненавидеть, но вы состоите из атомов, которым он, возможно, найдет иное применение, и он, как сказал Юдковски, «…будет, скорее всего, сопротивляться любым вашим действиям, направленным на то, чтобы сохранить эти атомы в своем владении». Так что побочным эффектом бездумного программирования станет то, что готовый ИИ не будет трепетно относиться к вашим атомам.

При этом ни общественность, ни разработчики ИИ не заметят опасности, пока не станет слишком поздно.

> Есть тенденция считать, что благонамеренные люди создают хороший ИИ, а злонамеренные — плохой и злой ИИ. Но источник проблемы не в этом. Источник проблемы в том, что, когда даже самые благонамеренные люди работают над созданием ИИ, их не особенно заботит вопрос дружественности этого самого ИИ. Они уверены, что поскольку сами они питают исключительно добрые намерения, то и созданный ими ИИ автоматически получится благонамеренным, а это неверно. На самом деле создание такого ИИ — сложнейшая математическая и инженерная задача. Мне кажется, что большинство из них просто недостаточно хорошо умеют думать о неудобных и неприятных вещах. Они начали с того, что *не стали* думать так: "Дружественный ИИ — это проблема, которая убивает".

Юдковски говорит, что создатели ИИ заражены идеей счастливого будущего, преображенного искусственным интеллектом, и эта идея живет в их воображении. Они думают о ней с тех самых пор, как их укусила муха ИИ.

> Они не хотят слышать ничего, что противоречило бы этой идее. Ты говоришь им о том, что ИИ может оказаться недружественным, но слова как будто отскакивают. Как гласит

> старая пословица, больше всего вреда наносят те, кто хочет чувствовать собственную значительность. Многим амбициозным людям намного проще думать об уничтожении мира, чем о том, что они сами никак себя не проявят. И таковы *все без исключения* встреченные мной люди, считающие, что ИИ-проекты принесут им вечную славу.

Эти творцы ИИ — не сумасшедшие ученые, они ничем не отличаются от вас или от меня; в этой книге вы познакомитесь с некоторыми из них. Но вспомните об ошибке доступности из главы 2. Столкнувшись с задачей, человек, как правило, выбирает свежий, яркий или как-то иначе привлекший его внимание вариант решения. Гибель от «рук» ИИ, как правило, для создателей ИИ не является вариантом. По крайней мере, она куда сложнее, нежели продвижение в научной области, получение пожизненной должности, публикации, богатство и т. п.

Более того, мало кто из создателей ИИ, в отличие от *теоретиков* ИИ, озабочен построением дружественного ИИ. За одним-единственным исключением никто из десятка с лишним творцов ИИ, с которыми мне довелось разговаривать, не обеспокоен в достаточной мере тем, чтобы работать над «дружелюбием» ИИ или любыми другими защитными мерами. Возможно, мыслители переоценивают эту проблему, а может быть, проблема разработчиков — в незнании того, что они не знают. В одной популярной онлайн-газете Юдковски так написал об этом:

> Человеческий вид возник в результате естественного отбора, действующего через неслучайное сохранение случайных мутаций. Один из путей, ведущих к глобальной катастрофе — к тому, что кто-то нажмет кнопку, плохо представляя себе, что именно эта кнопка делает, — состоит в том, что искусственный интеллект возникает в результате аналогичного процесса постепенного набора рабочих алгоритмов, при том что *исследователи не особенно хорошо понимают, как работает система в целом* (*курсив мой. — Дж. Б.).*
>
> Незнание того, как построить дружественный ИИ, не смертельно само по себе… А вот ошибочная убежденность в том, что любой ИИ будет дружественным, — очевидный путь к глобальной катастрофе.

Считать, что ИИ человеческого уровня (УЧИ) будет непременно

дружественным, неверно по многим причинам. Такое убеждение становится еще более опасным после того, как интеллект УЧИ начинает стремительно расти, оставляя наш далеко позади, и превращается в ИСИ — искусственный суперинтеллект. Так как же создать дружественный ИИ? Или вы считаете, что можно привить машине «дружелюбие» уже готовому продвинутому ИИ? Юдковски написал и выложил в Интернет трактат размером с книгу, посвященный этим вопросам и озаглавленный «Создание дружественного ИИ: Анализ и дизайн доброжелательных целевых архитектур». Дружественный ИИ — вопрос настолько сложный для понимания и при этом настолько важный, что ставит в тупик даже главного своего поборника, который пишет:

> Достаточно одной ошибки в цепи рассуждений, чтобы неожиданно для себя оказаться аж во Внешней Монголии.

Начнем с простого определения. Дружественный ИИ — это *ИИ, который оказывает скорее положительное, нежели отрицательное влияние на человечество.* Дружественный ИИ имеет собственные цели и предпринимает действия для их достижения. Теоретики описывают успех ИИ в достижении его целей при помощи экономического термина «полезность». Из вводного курса по экономической теории известно, что разумные потребители стараются максимально увеличить полезность, расходуя свои ресурсы таким образом, чтобы получить от них максимальное удовлетворение. В случае ИИ удовлетворение получается от достижения целей, а любое действие, продвигающее ИИ к достижению его целей, характеризуется высокой «полезностью».

Ценности и предпочтения вкупе с удовлетворением от достижения цели можно поместить во встроенное в ИИ определение пользы, назвав его «функцией полезности». Дружественность по отношению к человеку — одна из ценностей, которыми, на наш взгляд, должен обладать ИИ. Поэтому неважно, какие цели ставит перед собой ИИ — от игры в шахматы до управления автомобилем, — защита человеческих ценностей (и человека как такового) должна составлять существенную часть кодекса его поведения.

Надо отметить, что здесь имеется в виду не то *дружелюбие*, о каком обычно говорят телевизионные проповедники, — хотя и это не повредит. В нашем случае дружественность означает, что ИИ не должен быть враждебно или амбивалентно настроен по отношению к людям *никогда*, какими бы ни были его цели и сколько бы ступеней

самосовершенствования он ни прошел. ИИ должен глубоко понимать нашу природу и не наносить вреда людям даже случайно, даже через непредвиденные последствия своих действий (что нередко происходило в рассказах Азимова в рамках действия Трех законов робототехники). То есть мы не хотим получить ИИ, который выполнил бы наши краткосрочные задачи (пожалуйста, спаси нас от голода) при помощи мер, вредных в долгосрочной перспективе (к примеру, поджарив одновременно всех кур на планете), или таких, против которых мы возражали бы в принципе (убив нас всех после ближайшей же трапезы).

В качестве примера непредвиденных последствий специалист по этике из Оксфордского университета Ник Востром предложил гипотетический «максимизатор производства скрепок». В сценарии Бострома бездумно запрограммированный суперинтеллект, которому в качестве цели было задано производство канцелярских скрепок, делает ровно то, что от него требовалось, без оглядки на человеческие ценности. В результате все идет наперекосяк, поскольку ИСИ «превращает сначала всю Землю, а затем и прилегающие области пространства в фабрики по производству скрепок». Дружественный ИИ сделал бы в подобной ситуации ровно столько скрепок, сколько укладывается в человеческие ценности.

Еще одно непременное качество дружественного ИИ — стремление избежать догматических ценностей. Наши представления о хорошем и плохом изменяются со временем, и любому ИИ, связанному с человеческим благополучием, необходимо будет поспевать за нами в этом отношении. Если бы функция полезности некоего ИИ была ориентирована на предпочтения большинства европейцев в 1700 г. и не корректировалась бы со временем, то и в XXI в. этот ИИ связывал бы человеческое счастье и благополучие с такими архаичными ценностями, как расовое и половое неравенство, рабовладение, туфли с пряжками, а то и что-нибудь похуже. Мы не хотим закладывать в дружественный ИИ конкретные жестко заданные ценности. Мы хотим, чтобы его подвижная шкала ценностей развивалась с нами вместе.

Юдковски придумал для этого развития ценностей специальный термин — когерентная экстраполированная воля (КЭВ). ИИ, снабженный КЭВ, способен предвидеть наши желания. И не просто желания, а те желания, которые были бы у нас, если бы мы «знали больше, думали быстрее и лучше соответствовали бы собственным представлениям о себе».

КЭВ стал бы своеобразным оракулом дружественного ИИ. Ему пришлось бы извлекать из нас наши ценности, *как если бы* мы были лучше, чем есть на самом деле, и при этом сохранять демократичность и не

допускать, чтобы нормы, установленные меньшинством, тиранили все человечество.

Вам кажется, что все это звучит немного… ну, не от мира сего? Для того есть веские причины. Во-первых, я излагаю концепции дружественного ИИ и КЭВ очень схематично, на самом деле об этом написаны целые тома (их можно почитать в Интернете). А во-вторых, тема дружественного ИИ не слишком проработана, но очень оптимистична. Неясно, можно ли изложить концепцию дружественного ИИ в формальном математическом виде, и не исключено, что построить такой ИИ или интегрировать его в перспективные ИИ-архитектуры попросту невозможно. Но если бы это можно было сделать, как выглядело бы наше будущее?

Давайте представим, что через десять или сорок лет проект SyNAPSE[12] фирмы IBM по обратному проектированию мозга принес, наконец, плоды. В результате работ по программе, начавшейся в 2008 г. с гранта DARPA в $30 млн, создана система, копирующая базовые схемы мозга млекопитающего: эта система одновременно воспринимает входные сигналы из тысяч источников, развивает стержневые алгоритмы обработки данных и выдает на выходе восприятие, мысль и действие. Начиналось все с мозга, по размеру примерно соответствующего кошачьему, затем дошло до человеческого — и двинулось дальше.

Чтобы построить такую копию, исследователи проекта SyNAPSE создали «когнитивный компьютер» из тысяч параллельных компьютерных чипов, обрабатывающих информацию. Воспользовавшись достижениями нанотехнологий, они разработали микросхемы размером в один квадратный микрон. Затем объединили множество таких микросхем в углеродный шар размером с баскетбольный мяч и для максимальной производительности погрузили в жидкий металл — соединение галлия с алюминием.

Отметим, что резервуар с металлом представляет собой мощный беспроводной роутер, подключенный к Интернету и связанный с миллионами разбросанных по всей планете сенсорами. Сенсоры принимают сигналы с камер, микрофонов, датчиков давления и температуры, от роботов и природных систем — пустынь, ледников, озер, рек, океанов, тропических лесов. SyNAPSE обрабатывает поступающую информацию, анализируя свойства и взаимосвязи этого громадного массива данных. Функция следует за формой, и нейроморфная, имитирующая живой мозг система самостоятельно создает разум.

В настоящее время SyNAPSE успешно симулирует работу 30 млрд

нейронов человеческого мозга и 100 трлн нейронных связей — синапсов. Система уже превзошла скорость работы мозга — приблизительно 1000 трлн операций в секунду.

Впервые в истории человеческий мозг становится *вторым* по сложности объектом в известной нам Вселенной.

А что же дружественность? Понимая, что «дружественность» должна стать сердцем любой разумной системы, ее создатели встроили алгоритмы ценности и безопасности в каждый из миллионов чипов SyNAPSE. Он дружелюбен до самых глубин, до ДНК. Поэтому теперь, когда когнитивный компьютер становится все более мощным, он начинает принимать решения, значимые для судеб мира, — что делать с ИИ террористических государств, как изменить траекторию летящего к Земле астероида, как остановить стремительное повышение уровня Мирового океана или как ускорить развитие наномедицины, способной справиться с большинством болезней.

Обладая таким глубоким пониманием проблем человечества, SyNAPSE с легкостью определит, что выбрали бы люди, если бы были достаточно могущественными и разумными, чтобы принимать ответственные решения. В будущем мы переживем интеллектуальный взрыв! Мало того, человечество расцветет!

Да будет благословен дружественный ИИ!

Теперь, когда большинство разработчиков и теоретиков ИИ оценивает Три закона робототехники Азимова так, как они того заслуживают, — как инструменты построения сюжета, а не выживания, — дружественный ИИ, возможно, становится наилучшей концепцией, которую может предложить человечество, думая о сохранении своего существования. Но дружественный ИИ еще не создан, а у него уже много различных проблем.

Одна из них — это то, что слишком много организаций в самых разных странах работают над УЧИ и смежными технологиями; они ни за что не договорятся отложить свои проекты до того момента, когда будет создан дружественный ИИ, или включить в свои программы формальный модуль дружественности, если таковой может быть создан. Более того, мало кто из них хотя бы участвует в общественной дискуссии о необходимости дружественного ИИ.

Среди участников гонки за УЧИ можно назвать: IBM (несколько проектов), Numenta, AGIRI, Vicarious, NELL и ACT-R Университета Карнеги-Меллона, SNERG, LIDA, CYC и Google. Можно назвать по крайней мере десяток проектов с менее надежными источниками финансирования: среди них SOAR, Novamente, NARS, AIXItl и Sentience.

Сотни других разработок, целиком или частично посвященные УЧИ, проводятся в США и других странах — некоторые из них под покровом тайны, некоторые (в таких странах, как Китай и Израиль) спрятаны за современным «железным занавесом» национальной безопасности. DARPA как открыто, так и тайно финансирует немало ИИ-проектов.

Что, собственно, я имею в виду. Вероятность того, что первый УЧИ будет создан непременно в Исследовательском институте машинного интеллекта MIRI и потому в него будет встроен модуль дружественности, достаточно мала. И вряд ли создатели первого УЧИ будут особенно много думать о таких вопросах, как дружественность. Тем не менее существует несколько способов заблокировать *недружественный* УЧИ. Президент Института машинного интеллекта Майкл Вассар[13] рассказал мне об образовательной программе для элитных университетов и математических конкурсов. MIRI и еще одна дружественная организация, Центр прикладного разума (CEAR), организовали «тренировочные лагеря разума», в которых надеются обучить будущих потенциальных строителей ИИ и руководителей, определяющих техническую политику, рациональному мышлению. Когда новая элита подрастет, эта наука пригодится им в работе и поможет избежать самых неприятных ловушек ИИ.

Этот план может показаться наивным, но на самом деле MIRI и CEAR нащупали важный фактор ИИ-риска. Тема сингулярности привлекает общественное внимание, и вопросы сингулярности будут попадать в поле зрения все большего числа все более умных людей. Окно возможностей для просвещения народа и информирования о связанных с ИИ рисках потихоньку открывается. Но любой план создания совещательного или руководящего органа по вопросам ИИ уже опоздал с реализацией; некоторых катастроф избежать уже невозможно. Как упоминалось в главе 1, по крайней мере 56 стран занимается в настоящее время разработкой боевых роботов. Утверждают, что в разгар войны в Ираке три дрона SWORDS фирмы Foster-Miller, вооруженных пулеметами, были выведены из боевой зоны после того, как направили оружие против своих. В 2007 г. в Южной Африке роботизированная зенитная пушка убила 9 и ранила 15 солдат в результате инцидента, продолжавшегося *одну восьмую долю секунды.*

Конечно, это еще не *Терминатор*, но дальше — больше. Как только появится продвинутый ИИ, особенно если платить за него будет DARPA или аналогичное агентство другой страны, ничто на свете не помешает оснастить им боевых роботов. Более того, роботы могут послужить

платформами для реализации машинного обучения, в результате которого и появится ИИ. Когда же появится дружественный ИИ (если это когда-нибудь произойдет), сами подумайте, станут ли частные компании — производители роботов устанавливать его в машины, разработанные для убийства людей? Акционерам это наверняка не понравится.

С дружественным ИИ есть и еще один вопрос: как, собственно, его дружелюбие переживет интеллектуальный взрыв? То есть останется ли дружественный ИИ таковым после того, как его интеллект вырастет в тысячу раз? Юдковски в статьях и выступлениях описывает то, что может при этом произойти, примерно так:

> Ганди не хочет убивать людей. Если предложить Ганди пилюлю, которая заставит его пожелать этого, он откажется принимать ее, поскольку знает, что тогда он кого-нибудь убьет, а нынешний Ганди не хочет убивать людей. Из этого, грубо говоря, следует, что разум, достаточно развитый, чтобы точно модифицировать и улучшать себя, стремится остаться в рамках тех мотиваций, с которых начинал.

Лично меня этот тезис не убедил. Если мы не можем знать, как поведет себя разум более сильный, чем разум человека, как мы можем быть уверены, что он сохранит свою функцию полезности или основные представления о морали? Разве не может случиться так, что он, поумнев в тысячу раз, рассмотрит и отвергнет запрограммированное дружелюбие?

> Нет, — *ответил Юдковски, когда я спросил об этом.* — Он станет в тысячу раз более эффективно *сохранять* свою функцию полезности.

Но что, если при тысячекратном по сравнению с человеческим росте интеллекта произойдет какой-то системный сбой, который мы сегодня не можем даже предугадать? К примеру, у нас с плоскими червями много общей ДНК. Но вряд ли мы прониклись бы их целями и моралью, если бы даже узнали, что много миллионов лет назад именно плоские черви создали нас, наделив собственными ценностями. Согласитесь, очень скоро мы оправились бы от первоначального удивления и продолжили бы делать ровно то, что считали нужным.

«Очень понятно, почему возникают такие подозрения, — заметил Юдковски. — Но создание дружественного ИИ не похоже на обучение или

инструктирование человека. У людей уже есть собственные цели, собственные эмоции и собственные средства реализации. У них есть собственная структура для размышлений о моральных вопросах. Внутри нас существует нечто, что внимательно рассматривает каждое поступающее указание и решает, выполнять его или нет. В случае с ИИ вы формируете разум целиком с нуля. Если убрать из ИИ все программные коды, останется девственно чистый компьютер, который не будет ничего делать, потому что делать ему будет нечего.

И все же я сказал: «Если завтра я поумнел бы в тысячу раз по сравнению с собой сегодняшним, то мне кажется, я оглянулся бы на свои сегодняшние заботы с мыслью о том, что "все это пустое". Не верится, что для моего нового тысячесильного разума то, что я ценил вчера, сохранило бы хоть какую-то значимость».

«У вас есть специфическая эмоция "все это пустое", и вы уверены, что у сверхразума она тоже будет, — заметил Юдковски. — Это *антропоморфизм*. ИИ устроен и работает не так, как вы, и у него нет эмоции типа "все это пустое"».

Но, сказал он, есть одно исключение. Речь идет о человеческом разуме, загруженном в компьютер. Это еще один путь к созданию УЧИ, который иногда путают с обратным проектированием мозга. Обратное проектирование предполагает сначала разобраться до тонкостей в человеческом мозге, а затем представить то, что мозг делает, в виде схем и программ. В конце этого процесса вы получите компьютер, обладающий искусственным интеллектом человеческого уровня. Проект Blue Brain фирмы IBM нацелен на достижение этого результата к началу 2020-х гг.

С другой стороны, загрузка готового разума в компьютер, которую называют также полной эмуляцией мозга, — это моделирование человеческого разума, такого как ваш, на компьютере. В конце этого процесса вы остаетесь при своем мозге (если, конечно, процесс сканирования и переноса мозга не разрушает его, как считают некоторые эксперты), а внутри машины появляется ваша мыслящая и чувствующая копия.

> Если бы у вас был сверхразум, который первоначально был копией человеческого мозга, а затем начал самосовершенствоваться и со временем отходил бы все дальше и дальше от оригинала, то такой разум действительно мог бы обернуться против человечества примерно по тем самым причинам, о которых вы говорили, — *сказал Юдковски*. — Но

> если говорить о синтетическом ИИ, созданном не на основе человеческого разума, то такого никогда не произойдет, потому что этот разум слишком чужд человеческому. Громадное большинство таких ИИ вполне могли бы вас убить, но не по названным причинам. Все, что вы вообразили, применимо лишь к сверхразуму, исходным материалом для которого послужил человеческий интеллект.

Мне еще предстояло узнать в ходе своих исследований, что многие специалисты оспаривают концепцию дружественного ИИ, исходя из других соображений. На следующий день после встречи с Юдковски я разговаривал по телефону с доктором Джеймсом Хьюзом, заведующим кафедрой философии Тринити-колледжа и исполнительным директором Института этики и новых технологий IEET. Хьюз пытался доказать слабость идеи о том, что функция полезности ИИ не может меняться.

> Одна из аксиом для тех, кто говорит о дружественном ИИ, состоит в том, что при достаточно тщательном подходе можно сконструировать сверхразумное существо с набором целей, который затем меняться не будет. Они почему-то игнорируют тот факт, что у нас, людей, имеются фундаментальные цели — секс, пища, убежище, безопасность. Иногда они трансформируются в такие вещи, как желание стать террористом-смертником или жажда получить как можно больше денег, — в общем, в вещи, не имеющие ничего общего с изначальным набором целей, но построенные на его основе в результате серии последовательных шагов, которые мы можем пронаблюдать в собственной голове.
>
> Таким образом, *мы* способны исследовать свои собственные цели и изменять их. К примеру, мы можем избрать для себя целибат, который полностью противоречит нашей генетической программе. Мысль о том, что сверхразумное существо с пластичным разумом (каким, собственно, и должен обладать ИИ) *не будет* меняться со временем, попросту абсурдна[14].

Веб-сайт лаборатории идей Хьюза, IEET, показывает, что их критика относится в равной степени ко всем; их тревожат не только опасности ИИ, но и опасности нанотехнологий, биотехнологий и других рискованных новинок. Хьюз считает, что сверхразум опасен, но шансы на его скорое появление невелики. Однако он *настолько* опасен, что связанный с его

появлением риск должен рассматриваться в одном ряду с неминуемыми катастрофами, такими как повышение уровня Мирового океана и прилет гигантского астероида (то и другое относится к первой категории по шкале рисков Уоррена Льюиса, о которой шла речь в главе 2). Хьюз согласен и с другим моим утверждением: первые шаги развития ИИ на пути к сверхразуму (Хьюз называет его «богом в ящике») тоже опасны.

> «MIRI просто закрывает на это все глаза, потому что они сосредоточены на том, как бог выскочит из ящика. А когда бог выскакивает из ящика, человеческие существа уже ничего не могут предпринять, чтоб изменить ход вещей. Бог окажется либо хорошим, либо плохим, и к этому сводится подход MIRI. Позаботьтесь лучше о том, чтобы это был хороший бог!

Мысль о боге, который выскакивает из ящика, напомнила мне еще об одном незаконченном деле — об эксперименте «ИИ в ящике». Напомню вкратце. Елиезер Юдковски играл в этом эксперименте роль ИСИ, заключенного в компьютере без всякой физической связи с внешним миром (ни проводов, ни роутеров, ни беспроводной связи). Цель Юдковски: выбраться из «ящика». Цель Привратника: удержать его в заточении. Игра велась в чате, переговоры игроков шли в текстовой форме. Каждый сеанс продолжался не более двух часов. Вариант полного молчания, которое утомило бы Привратника до смерти и вынудило к сдаче, не был запрещен, но никогда не использовался.

В 2002–2005 гг. Юдковски сыграл против пяти Привратников. Трижды он (вернее, его герой-ИСИ) «сбежал» и дважды остался узником ящика. Как ему удалось уйти? Я вычитал в Сети, что одним из правил эксперимента была секретность переписки между ИСИ и Привратником, так что ответа на этот вопрос я не знал. Но почему такая секретность?

Поставьте себя на место Юдковски. Если у вас в роли «ИИ в ящике» есть какие-то хитроумные способы бегства, зачем же открывать их всем подряд и таким образом предупреждать *следующего* Привратника? А если вам захочется повторить уже использованный метод? А во-вторых, если вы пытаетесь изобразить силу убеждения существа в тысячу раз более умного, чем самый умный человек, то вам, возможно, захочется слегка выйти за рамки социально приемлемого диалога. А может быть, и не слегка. И кто захочет делиться этим со всем миром?

Эксперимент «ИИ в ящике» важен потому, что одним из вероятных последствий деятельности сверхразума без вмешательства человека

является уничтожение человечества — и противоборство с ним человечество, судя по всему, выиграть не сможет. Тот факт, что Юдковски в роли ИИ выиграл три раза из пяти, еще сильнее встревожил и заинтриговал меня. Может быть, он гений, но он, в отличие от ИСИ, наверняка не в тысячу раз умнее самого умного человека. И вообще, дурному или безразличному ИСИ достаточно всего один раз выбраться из пресловутого ящика.

Кроме того, эксперимент «ИИ в ящике» заинтриговал меня еще и тем, что, по существу, это вариант старого доброго теста Тьюринга. Этот тест, разработанный в 1950 г. математиком, компьютерщиком и участником взлома немецких шифров во время Второй мировой войны Аланом Тьюрингом, предназначен для определения разумности машины. В этом тесте судья задает компьютеру и человеку письменные вопросы, и если по ответам он не в состоянии определить, кто из его собеседников — человек, а кто — компьютер, то компьютер выигрывает.

Но есть одна хитрость. Тьюринг понимал, что мышление, как и разум, — скользкая тема. И тому и другому сложно дать определение, хотя мы без проблем узнаем то и другое при встрече. Чтобы пройти тест Тьюринга, ИИ не обязательно должен думать как человек, потому что откуда кому бы то ни было знать, *как именно* он думает? Однако ему необходимо убедительно *притвориться*, что он думает как человек, и выдавать на все вопросы человекоподобные ответы. Сам Тьюринг называл свой тест «имитационной игрой». Он отвергал возражения критиков относительно того, что машина, возможно, вовсе не будет думать по-человечески. Он писал:

> Разве машины не могут выполнять некое действие, которое следует описывать как мышление, но которое сильно отличается от того, что делает человек?

Иными словами, он возражает против утверждения, которое следует из эксперимента с «китайской комнатой» Джона Сёрля: если машина не думает по-человечески, она не разумна. Большинство экспертов, с которыми мне довелось общаться, согласны с Тьюрингом. Если ИИ поступает разумно, кому какое дело, как выглядят его программы?

Однако существует по крайней мере две серьезные причины интересоваться этим вопросом. Прозрачность «мыслительного» процесса ИИ хотя бы до того момента, когда он уйдет далеко вперед и будет уже недоступен нашему пониманию, необходима для нашего выживания. Если

мы хотим попытаться внушить ИИ дружелюбие или любые другие моральные качества или установить некие предохранители, необходимо до тонкостей понимать, как он работает, прежде чем он сможет модифицировать себя. После начала этого процесса наши вводные потеряют смысл. Кроме того, если когнитивная архитектура ИИ будет повторять архитектуру человеческого мозга или компьютерной копии человеческого мозга, она, возможно, будет менее чуждой, чем разработанная «с нуля» архитектура ИИ. Но специалисты по компьютерам ожесточенно спорят, будет ли такая связь с человечеством полезна, решит она проблемы или только создаст новые.

Ни одному компьютеру до сих пор не удалось пройти тест Тьюринга, хотя каждый год проводится конкурс на премию Лебнера, которую учредил филантроп Хью Лебнер для создателей компьютера, который пройдет этот тест. Но пока главный приз в $100 000 остается невостребованным, ежегодно $7000 выплачивается создателю «наиболее человекоподобного компьютера». Несколько последних лет этот приз достается чатботам — роботам, созданным для поддержания разговора (без особого успеха). Марвин Мински, один из пионеров в области искусственного интеллекта, предложил $100 тому, кто уговорит Лебнера отозвать свою премию. По мнению Мински, это «избавило бы нас от ужаса этой отвратительной и бесполезной ежегодной шумихи».

Как же Юдковски удалось уговорить свою «охрану»? Надо сказать, у него был богатый выбор кнутов и пряников. Он мог обещать богатство, излечение от болезни, изобретения, способные полностью покончить с голодом и нуждой. Решительное превосходство над врагами. Со стороны кнута можно оперировать надежным инструментом социального инжиниринга — страхом. Что, если в этот самый момент ваши враги готовятся применить против вас собственный ИСИ? В реальной жизненной ситуации это вполне может сработать — но как насчет ситуации искусственной, такой как в эксперименте «ИИ в ящике»?

Когда я задал Юдковски вопрос о его методах, он рассмеялся, потому что все ожидают каких-то дьявольски хитрых решений — каких-то логических фокусов, тактических ходов по типу «дилеммы заключенного», может быть, чего-то тревожного. На самом деле происходило все совершенно не так.

«Я шел трудным путем», — сказал Юдковски.

В трех успешных попытках, сказал он мне, он просто пресмыкался, упрашивал и разглагольствовал. Привратники в конце концов выпускали его и оплачивали проигрыш. А те два раза, когда ему не удалось вырваться,

он тоже умолял. Ощущения у него после этого были не самые приятные, и он поклялся никогда больше не повторять этот эксперимент.

Покидая жилище Юдковски, я понял, что он не сказал мне всей правды. Какие просьбы могут подействовать на человека, который решительно настроен не поддаваться на уговоры? Неужели он мог сказать: «Спаси меня, Елиезер Юдковски, от публичного унижения! Спаси меня от боли поражения!» Или, может, Юдковски как человек, посвятивший свою жизнь разоблачению опасностей ИИ, мог договориться со своим противником о *мета*-сделке. О сделке, связанной с самим экспериментом «ИИ в ящике». Он мог, к примеру, сказать: «Помоги мне показать миру, что человек ненадежен и нельзя доверять ему контроль за ИИ!»

Конечно, такой вариант хорошо сработал бы на пропаганду и привлек бы немало сторонников. Но никаких уроков об обращении с ИИ в реальном мире из этой ситуации извлечь было бы невозможно.

Вернемся к дружественному ИИ. Если создание такого ИИ представляется маловероятным, означает ли это, что интеллектуальный взрыв неизбежен? Неужели ИИ обязательно окажется на свободе? Если вы, как и я, всегда считали, что компьютеры сами по себе инертны и не создают проблем, вас это удивит. Зачем, спрашивается, ИИ будет делать что бы то ни было, тем более умолять, угрожать или убегать?

Чтобы прояснить этот вопрос, я разыскал разработчика ИИ Стивена Омохундро, президента компании Self-Aware Systems.

Сам он физик, классный программист и занят разработкой научного подхода к пониманию интеллекта, превышающего человеческий. Он утверждает, что обладающие самосознанием самосовершенствующиеся ИИ-системы будут мотивированы на совершение неожиданных и даже странных действий. По мнению Омохундро, робот, разработанный для игры в шахматы, при достаточном уровне интеллекта может захотеть построить космический корабль

Глава 5
Программы, которые создают программы

> *…Мы начинаем зависеть от компьютеров, которые помогают нам разрабатывать новые компьютеры, позволяющие производить гораздо более сложные вещи. И все же мы не до конца понимаем этот процесс — не догоняем, как говорится. Мы используем программы для создания гораздо более быстрых компьютеров, чтобы процесс мог протекать намного быстрее. Именно это ставит в тупик — технологии начинают подпитывать себя сами, а мы отходим в сторону. Сегодняшние процессы аналогичны превращению одноклеточных: организмов в многоклеточные. Мы — амебы, и мы не в состоянии понять, что, черт возьми, мы создаем.*
>
> *Дэнни Хиллис, основатель компании Thinking Machines, Inc.*

Мы с вами живем в интересный и важный период истории человечества. Примерно к 2030 г., меньше чем через поколение, нам, возможно, придется делить Землю со сверхразумными машинами, и уцелеть будет совсем не просто. Теоретики ИИ вновь и вновь поднимают несколько тем, важнейшей из которых является следующая: *чтобы понять их, нам необходима специальная наука.*

Мы уже рассмотрели катастрофический сценарий под названием Busy Child, затрагивали тему замечательных возможностей, которые может получить ИИ, когда превзойдет человеческий разум в процессе рекурсивного самосовершенствования. Среди этих возможностей — самокопирование, мозговой штурм с привлечением множества копий самого себя, сверхскоростные вычисления, работа круглые сутки и без выходных, имитация дружелюбия, имитация собственной смерти и многое другое. Мы предположили, что искусственный сверхразум не удовлетворится существованием в изоляции; побуждения и разум толкнут его в наш мир и поставят наше существование под угрозу. Но почему у

компьютера вообще должны возникнуть какие-то побуждения? И почему они будут угрожать нам?

Чтобы ответить на эти вопросы, нужно предсказать, насколько агрессивно и настойчиво будет вести себя ИИ. К счастью, основы уже заложены.

> Конечно же, ничего плохого не может случиться из-за постройки робота для игры в шахматы, правда?.. На самом деле такой робот будет опасен, если не подойти к его проектированию с максимальной тщательностью. Без специальных предосторожностей он будет сопротивляться выключению, попытается вломиться в другие машины и изготовить копии самого себя, постарается собрать как можно больше ресурсов без оглядки на чью бы то ни было безопасность. Это потенциально деструктивное поведение возникнет не потому, что он так запрограммирован с самого начала, но потому, что такова природа систем ориентированных на достижение цели.

Автор этого отрывка — Стив Омохундро. Высокий, спортивный, энергичный и чертовски жизнерадостный человек, глубоко заглянувший в самое нутро интеллектуального взрыва. У него упругая походка, энергичное рукопожатие и улыбка, излучающая яркие лучи доброжелательности. Он встретился со мной в ресторане в Пало-Альто, городе возле Стэнфордского университета, который и окончил с отличием, прежде чем отправиться в Университет Калифорнии в Беркли и защитить докторскую диссертацию по физике. На базе этой работы он написал книгу «Теория геометрического возмущения в физике», посвященную новым на тот момент достижениям дифференциальной геометрии. Эта книга стала для Омохундро первым шагом на пути популяризации, когда необходимо просто излагать сложные вещи.

Омохундро — уважаемый профессор в области ИИ, плодовитый научно-популярный писатель и пионер таких областей ИИ, как чтение по губам и распознавание изображений. Он участвовал в создании компьютерных языков StarLisp и Sather, предназначенных для программирования ИИ. В числе всего лишь семи инженеров группы Wolfram Research он создавал мощный программный комплекс Mathematica, любимый учеными и инженерами всего мира.

Омохундро слишком оптимистичен, чтобы легко бросаться таким словами, как *катастрофа* или *уничтожение,* но из его анализа ИИ-рисков

следуют самые страшные выводы, какие мне доводилось слышать. Он, в отличие от многих теоретиков, не считает, что возможно почти бесконечное число развитых ИИ и что некоторые из них безопасны. Он убежден, что без тщательнейшего программирования *все* достаточно разумные ИИ будут смертоносны.

> Если система осознала себя и способна создать свою улучшенную копию, это великолепно, — *сказал мне Омохундро*. — Самосовершенствование у нее получится лучше, чем если ее будут модернизировать программисты. С другой стороны, чем она станет после множества итераций? Не думаю, что большинство ИИ-исследователей видят какую-то потенциальную опасность в создании, скажем, робота-шахматиста. Но мой анализ показывает, что нам следует тщательно обдумать, какие ценности мы в него закладываем; в противном случае мы получим нечто вроде психопатической, эгоистической, зацикленной на себе сущности.

Ключевые моменты здесь следующие: даже ИИ-исследователи не подозревают, что, во-первых, полезные системы могут быть опасны и, во-вторых, обладающие самосознанием самосовершенствующиеся системы могут быть психопатическими.

Психопатическими?

Омохундро говорит о плохом программировании. Ошибки программистов, бывало, заставляли дорогущие ракеты врезаться в землю, заживо сжигали раковых больных слишком большой дозой излучения и оставляли миллионы людей без электричества. Если бы все инженерное конструирование было настолько ненадежным и дефектным, как множество реальных компьютерных программ, говорит он, летать на самолетах или ездить по мостам было бы небезопасно.

Национальный институт стандартов и технологий США выяснил, что каждый год дефектное программирование обходится американской экономике более чем в $60 млрд недополученной прибыли. Иными словами, ежегодные потери американцев из-за плохих программ превышают ВВП большинства стран мира.

> Величайшая ирония судьбы в том, что компьютерные науки, по идее, должны быть самыми математическими из всех наук, — *сказал Омохундро*. — По существу, компьютеры — это

> математические машины, которые должны вести себя абсолютно предсказуемо. При всем том создание компьютерных программ — одно из самых непредсказуемых инженерных занятий, полное ошибок и проблем с безопасностью.

Существует ли противоядие против дефектных ракет и дрянных программ?

> Программы, которые исправляют себя сами, — *говорит Омохундро*. Подход к искусственному интеллекту, который применяет моя компания, состоит в создании систем, которые понимают собственное поведение и способны наблюдать за собой в процессе работы и решения задач. Они замечают, когда в работе происходят сбои, а затем изменяют и улучшают себя.

Самосовершенствующееся программное обеспечение — не просто цель компании Омохундро, но очень логичный и даже неизбежный следующий шаг в развитии большинства программ. Но самосовершенствующихся программ того рода, о котором говорит Омохундро, — таких, чтобы сознавали себя и способны были разрабатывать более совершенные версии, — пока не существует. Однако их предшественники — программы, способные себя модифицировать, — уже работают всюду, и довольно давно. Специалисты по ИИ значительную часть самомодифицирующихся программных методик объединяют в широкую категорию «машинного обучения».

Чему обучается машина? Понятие *обучения* чем-то напоминает понятие *разума,* поскольку определений того и другого существует множество, и большинство из них верны. В простейшем смысле машина обучается, когда в ней происходит изменение, позволяющее во второй раз выполнить определенное задание лучше. Машинное обучение сделало возможным интернет-поиск, распознавание речи и рукописного текста; оно помогает пользователю в десятках самых разных приложений.

«Рекомендации» от Amazon — гиганта сетевой торговли — используют алгоритм машинного обучения, известный как анализ парных предпочтений (affinity analysis). Это стратегия, цель которой — сделать так, чтобы вы купили еще что-нибудь похожее (перекрестные продажи), что-нибудь более дорогое, или хотя бы сделать вас объектом дальнейших рекламных акций. Работает все это очень просто. Для любого товара, информацию о котором вы ищете (назовем его A), существуют другие

товары, которые часто покупают люди, купившие A, — это товары B, C и D. Запрашивая товар A, вы активируете алгоритм анализа парных предпочтений. Он ныряет в море данных о совершенных покупках и появляется оттуда с перечнем парных товаров. Таким образом, он использует свой постоянно пополняющийся банк данных и с каждым разом работает все лучше.

Кому выгодна самосовершенствующаяся часть этой программы? Amazon, разумеется, но и вам тоже. Анализ парных предпочтений — своеобразный помощник покупателя, позволяющий всякий раз при совершении покупок пользоваться обширной статистикой. Amazon ничего не забывает — он постоянно дополняет ваш покупательский профиль и раз от разу все лучше подбирает для вас товары.

Что происходит, когда вы поднимаетесь на ступеньку от программы, способной обучаться, к программе, которая на самом деле развивается, чтобы находить ответы на сложные задачи и даже создавать новые программы? Это еще не осознание себя и не самосовершенствование, но это еще один шаг в этом направлении — программы, которые создают программы.

Генетическое программирование — алгоритм машинного обучения, использующий возможности естественного отбора для поиска ответов на вопросы, на решение которых у человека ушло бы длительное время — возможно, годы. Этот алгоритм используется для написания новаторских, мощных программ.

Такой алгоритм во многих отношениях отличается от более распространенных методик программирования, которые я буду называть *обычным* программированием. В обычном программировании для решения задач используется, как правило, человеческая, а не компьютерная логика. В обычном программировании каждую строку кода пишет программист, и процесс обработки данных от входа до выхода, по крайней мере, в теории, прозрачен и поддается проверке.

Напротив, программисты, применяющие алгоритм генетического программирования, описывают задачу, которую необходимо решить, и позволяют естественному отбору сделать все остальное. Результаты могут быть поразительными.

Генетическая программа создает кусочки кода, как бы представляющие очередное поколение. Самые перспективные скрещиваются — случайным образом обмениваются блоками кода, давая новое поколение. Пригодность программы определяется тем, насколько близко она подходит к решению поставленной программистом задачи.

Непригодные отбрасываются, а лучшие скрещиваются вновь. На протяжении всего процесса программа проводит случайные изменения в отдельных командах и переменных — мутации. Однажды начав, генетическая программа работает сама по себе и не нуждается в человеческом вмешательстве.

Джон Коза из Стэнфордского университета в 1986 г. одним из первых начал заниматься генетическим программированием. Он использовал генетические алгоритмы при проектировании антенны для NASA, разработке программ распознавания белков и конструировании электрических контроллеров общего назначения. Двадцать три раза генетические алгоритмы Козы самостоятельно изобрели электронные компоненты, уже запатентованные людьми; программы работали просто по целевым инженерным спецификациям готовых устройств — это был критерий пригодности. К примеру, алгоритмы Козы изобрели преобразователь напряжения (устройство, используемое при испытаниях электронного оборудования), работавший точнее, чем схема, изобретенная человеком по тем же заданным спецификациям. Однако есть тут одна загадка: никто не может объяснить, как именно схема работает и почему она работает лучше; на первый взгляд, в ней есть лишние и даже ненужные детали.

Но это и есть самая интересная черта генетического программирования (и «эволюционного программирования» вообще — более широкой категории алгоритмов, к которой оно относится). Код программы нечитаем. Программа развивает решения, которые компьютерщики не в состоянии сколько-нибудь легко воспроизвести. Более того, они не могут понять путь, двигаясь по которому программа генетического программирования получила конечный результат. Вычислительное устройство с понятными вам входными и выходными данными, но неизвестной процедурой их обработки называется «черным ящиком». И непознаваемость этой процедуры — серьезный недостаток любой системы, использующей эволюционные компоненты. Каждый шаг к непрозрачности — шаг прочь от ответственности и приятных надежд на то, что нам удастся запрограммировать в ИИ дружественность по отношению к человеку.

Это не означает, что ученые всегда теряют контроль над «черными ящиками». Но если когнитивные архитектуры используют такие системы при создании ИСИ, — а это почти неизбежно, — то в «ДНК» ИСИ войдут целые пласты непонятного кода.

Непознаваемость может оказаться неизбежным свойством сознающей

себя самосовершенствующейся программы.

> Это системы совершенно иного рода, чем то, к чему мы привыкли, — *пояснил Омохундро*. — Если у вас есть система, способная изменять себя, то вы, возможно, понимаете ее первую версию. Но она может изменить себя до состояния, которое вы уже не сможете понять. Так что эти системы довольно непредсказуемы. Они очень мощные и потенциально опасны. Так что значительная часть нашей работы связана с тем, чтобы получить предполагаемую пользу и при этом избежать рисков.

Возвратимся к упомянутому Омохундро роботу-шахматисту. Как он может быть опасен? Разумеется, речь не идет о шахматной программе, установленной в вашем телефоне. Речь о потенциальном роботе-шахматисте, управляемом настолько сложной когнитивной архитектурой, что он способен переписать собственную программу, чтобы лучше играть в шахматы. Он обладает самосознанием и может совершенствовать себя. Что произойдет, если вы «попросите» его сыграть одну партию, а затем выключиться?

Омохундро объяснил:

> Хорошо, представим себе, что он только что сыграл свою лучшую партию в шахматы. Игра закончена. Наступает момент, когда компьютер должен выключиться. С его точки зрения это очень серьезное событие, потому что робот не способен самостоятельно включиться снова. Поэтому он хочет быть уверенным, что дела обстоят именно так, как, он *думает*, они обстоят. В частности, он подумает: "А сыграл ли я на самом деле эту партию? Что, если меня обманули? Что, если на самом деле я *не сыграл* ее? Что, если все это просто модель?"

Что, если я нахожусь внутри модели? Да, конечно, речь идет всего лишь о продвинутом шахматном роботе. Но с осознанием себя приходит стремление к самосохранению и немного паранойи.

Омохундро продолжал:

> Может быть, он думает, что следует выделить какие-то ресурсы и найти ответы на вопросы о природе реальности, прежде чем решиться на радикальный шаг и выключиться.

> Аннулировав инструкцию, запрещающую это делать, он может прийти к выводу, что ради ответа на вопрос о том, насколько сейчас подходящее время для этого, можно потратить значительное количество ресурсов.

«Значительное количество ресурсов — это сколько?» — поинтересовался я.

Лицо Омохундро помрачнело, но лишь на секунду.

> Робот может решить, что дело стоит того, чтобы потратить на него все ресурсы человечества.

Глава 6
Четыре фундаментальные потребности

Мы не сможем по-настоящему понять, почему сверхразумная машина принимает те решения, которые принимает. Как можно рассуждать, как можно торговаться, как можно разбираться, как думает машина, если она думает в измерениях, которые вы даже представить не можете?

Кевин Уорвик, профессор кибернетики, Университет Ридинга

«Сознающие себя самосовершенствующиеся системы могут использовать все ресурсы человечества». Ну вот, мы опять добрались до пункта, где ИИ обращаются со своими человеческими создателями как с рыжими пасынками Галактики. Поначалу «равнодушие» ИИ воспринимается с трудом, но затем вспоминаешь, что уважительное отношение к человечеству — наша черта, для машин это не характерно. Мы опять проецируем на компьютеры человеческие черты. ИИ выполняет команды, но при отсутствии запрещающих инструкций он следует и собственным побуждениям, таким, например, как нежелание выключаться.

Какие еще желания и потребности могут быть у робота? И почему он вообще следует каким бы то ни было желаниям?

По мнению Стива Омохундро, такие побуждения, как самосохранение и сбор ресурсов, изначально присущи любой системе с конечной задачей. Как мы уже говорили, системы ИИ в узком смысле в настоящее время нацелены на выполнение конкретной задачи: поиск заданных терминов в Интернете, оптимизация работы игровых программ, поиск ближайших ресторанов, подготовка персональных рекомендаций по книгам и товарам и т. п. ИИ в узком смысле делает, что поручено, и на этом останавливается. Но у сознающего себя ИИ, способного к самосовершенствованию, будут иные, более тесные отношения с целями, которые они преследуют, какими бы эти цели ни были: узкими, как выигрыш в шахматы, или широкими, как точный ответ на любой заданный вопрос. Омохундро утверждает, что, к счастью, существует готовый инструмент, который можно использовать для

исследования природы продвинутых ИИ-систем — и оценки нашего будущего в связи с ними.

Этот инструмент — «рациональный агент» из экономической теории. В микроэкономике — дисциплине, занятой исследованием экономического поведения отдельных людей и фирм, — когда-то считалось, что люди и группы людей рационально преследуют свои интересы. Они делают выбор, максимизирующий их полезность, или удовлетворение (как мы отмечали в главе 4). Вообще, можно заранее угадать их предпочтения, поскольку они рациональны в экономическом смысле. «Рациональный» в данном случае не означает обыденную рациональность — такую, к примеру, как пользование ремнем безопасности во время поездки на автомобиле. У этой рациональности специфически микроэкономическое значение, означающее, что у индивида (или «агента») обязательно есть цели и предпочтения (называемые в экономике функцией полезности). У него обязательно имеются представления о мире и о том, как лучше всего достичь своих целей и реализовать свои предпочтения. По мере изменения условий индивид меняет и свои представления о мире. Он выступает как рациональный экономический агент, когда преследует свои цели посредством действий, основанных на его текущих представлениях. Математик Джон фон Нейман (1903–1957) принимал участие в разработке идей, связывающих рациональность и функции полезности. Немного дальше мы увидим, что именно фон Нейман заложил основы многих концепций компьютерной науки, ИИ и экономики.

И все же социологи утверждают, что «рациональный экономический агент» — попросту вздор. Человек нерационален — мы редко формулируем свои цели и представления и далеко не всегда меняем представления о мире, когда меняются условия. Наши цели и предпочтения меняются вместе с направлением ветра, ценами на газ, временем последней трапезы и устойчивостью внимания. Плюс к тому, как мы говорили в главе 2, мы ментально стреножены ошибками в рассуждениях (когнитивными искажениями), что дополнительно затрудняет балансировку целей и представлений. Но если теория рационального агента не годится для предсказания человеческого поведения, она прекрасно подходит для исследования областей, управляемых правилами и разумом, таких как игры, принятие решений и… продвинутый ИИ.

Как мы уже отмечали, продвинутый ИИ может быть построен на основе так называемой «когнитивной архитектуры». Отдельные модули в ней могут отвечать за зрение, распознавание и генерацию речи, принятие решений, внимание и другие аспекты разума. Эти модули могут для

решения каждой задачи использовать разные программные стратегии, включая генетические алгоритмы, нейронные сети (процессоры, копирующие структуру мозга), схемы, разработанные на основе изучения мозговых процессов, поиска и др. Другие когнитивные архитектуры, такие как SyNAPSE фирмы IBM, разработаны для развития интеллекта без логического программирования. Вместо этого, по утверждению IBM, интеллект SyNAPSE в значительной мере будет совершенствоваться на основе его взаимодействия с окружающим миром.

Омохундро соглашается: когда *любая* из этих систем станет достаточно мощной, она будет рациональна: она способна будет моделировать окружающий мир, предвидеть вероятный исход тех или иных действий и определять, какие действия лучше всего соответствуют поставленной цели. Если они окажутся достаточно разумны, то неизбежно *станут* самосовершенствующимися, даже если при конструировании это не было предусмотрено. Почему? Чтобы повысить свои шансы на достижение цели, они будут искать способы повысить скорость и эффективность своего компьютерного «железа» и программного обеспечения.

Посмотрим еще раз. Разумные системы человеческого уровня по определению обладают самосознанием. А сознающая себя система, преследующая определенную цель, *обязательно начнет* совершенствовать себя. Однако самосовершенствование — операция тонкая, вроде как делать самому себе лифтинг лица при помощи ножа и зеркала. Омохундро сказал мне:

> Улучшение самого себя очень серьезно для системы — это столь же серьезное действие, как выключение для робота-шахматиста. Когда разумные системы улучшают себя, скажем, ради повышения эффективности, то они в любой момент могут вернуть все обратно, если новое состояние в какой-то момент станет не оптимальным. Но в случае ошибки — к примеру, слегка изменится цель — произойдет катастрофа. Все дальнейшее существование такой системы будет посвящено достижению новой цели посредством дефектной версии. Вероятность этого события делает любое самоулучшение очень деликатным делом.

Но сознающий себя самосовершенствующийся ИИ способен справиться с этой проблемой. Подобно нам, он может предсказывать, или моделировать, возможные варианты будущего.

> У него есть модель собственного программного языка и модель собственной программы, модель оборудования, на котором все это установлено, и модель логики, используемой при рассуждениях. Он способен создавать собственный программный код и контролировать себя в процессе исполнения этого кода, то есть он способен учиться на собственном опыте. Он может обдумывать изменения, которые он мог бы провести в себе. Он может изменить любой аспект собственной структуры ради того, чтобы улучшить свое поведение в будущем.

Омохундро предсказывает, что у сознающих себя самосовершенствующихся систем со временем возникнет четыре первичных побуждения, или потребности, аналогичные биологическим потребностям человека: эффективность, самосохранение, приобретение ресурсов и творчество. При этом механизм возникновения этих потребностей открывает захватывающую природу ИИ. Первичные потребности не являются неотъемлемыми свойствами рационального агента. Достаточно разумный ИИ разовьет их у себя, чтобы *избежать* предсказуемых проблем (Омохундро называет их *уязвимостями*) в процессе достижения целей. ИИ *отступает* в эти потребности, потому что без них будет постоянно совершать дорогостоящие (в смысле ресурсов) ошибки.

Первая потребность — эффективность — означает, что самосовершенствующаяся система будет извлекать из имеющихся в ее распоряжении ресурсов (пространства, времени, вещества и энергии) максимум пользы. Она будет стремиться сделать себя компактной и быстрой, как с вычислительной, так и с физической точки зрения. Ради максимальной эффективности она будет снова и снова рассчитывать и перераспределять ресурсы между программным обеспечением и «железом». Для системы, которая непрерывно обучается и совершенствуется, особенно важным будет распределение памяти, а также повышение рациональности и избегание затратной логики. Предположим, говорил Омохундро, что некий ИИ, выбирая географическое место, из Сан-Франциско и Пало-Альто предпочитает Сан-Франциско, из Беркли и Сан-Франциско — Беркли, а из Беркли и Пало-Альто — Пало-Альто. Поступая в соответствии с этими предпочтениями, он, подобно азимовскому роботу, окажется в замкнутом круге из трех городов. А вот самосовершенствующийся ИИ по Омохундро заранее разглядит эту проблему и решит ее. Возможно, он даже воспользуется какой-нибудь

хитрой методикой вроде генетического программирования, которой особенно хорошо удается решение путевых задач по типу «задачи коммивояжера». Самосовершенствующуюся систему можно обучить генетическому программированию, и она будет применять эту методику для получения быстрых и энергетически выгодных результатов. А если не учить ее генетическому программированию, не исключено, что она сама его изобретет.

Этой системе под силу также модифицировать собственную конструкцию, так что она будет заниматься поиском наиболее эффективных материалов и архитектур. А поскольку атомная точность конструкции позволит системе значительно повысить ресурсную эффективность, она обратится к нанотехнологиям. Примечательно, что, если отработанных нанотехнологий к тому моменту еще не будет, система будет стремиться изобрести и их тоже. Помните трагическое развитие событий в сценарии Busy Child, когда ИСИ начинает трансформировать Землю и ее обитателей в материал для строительства новых вычислительных мощностей? Именно потребность в эффективности толкает Busy Child на использование или изобретение любых технологий и процедур, способных минимизировать расход ресурсов, в том числе и нанотехнологий. Виртуальная среда для проверки гипотез тоже помогает сберечь энергию, так что сознающие себя системы могут *виртуализировать* все, что им не обязательно делать в реальном мире.

Следующая потребность — самосохранение — тот момент, где ИИ переходит границу, отделяющую безопасное от опасного, а машины от хищников. Мы уже видели, как робот-шахматист у Омохундро относится к выключению себя. Он может выделить значительные ресурсы — мало того, все ресурсы, которые в настоящее время имеются в распоряжении человечества, — на исследование вопроса о том, подходящее ли сейчас время для выключения или его ввели в заблуждение о природе реальности. Если перспектива выключения так возбуждает робота-шахматиста, то вероятность уничтожения откровенно разозлит его. Сознающая себя система предпримет меры, чтобы избежать гибели, — и не потому, что так уж ценит собственное существование, а потому, что, если «умрет», не сможет достичь заданных целей. Омохундро постулирует, что эта потребность может заставить ИИ зайти достаточно далеко, чтобы обеспечить собственное существование, — к примеру, изготовить множество копий себя. Такие крайние меры дорого обходятся — на них тратятся ресурсы. Но ИИ пойдет на них, если увидит, что угроза оправдывает затраты, а ресурсов хватает. В сценарии Busy Child ИИ

определяет, что задача вырваться из «ящика», в котором он заключен, оправдывает «коллективный» подход, поскольку в любой момент можно ожидать отключения. ИИ многократно копирует себя и устраивает мозговой штурм задачи. Но такие вещи хорошо предлагать, когда на суперкомпьютере хватает свободного места для хранения копий; если места мало, это отчаянная и, возможно, недоступная мера.

Вырвавшись на свободу, ИСИ по имени Busy Child посвящает самосохранению немалые усилия: прячет свои копии в облаках, создает бот-сети для отражения атак и т. п.

Ресурсы, используемые для самосохранения, *по идее* должны соответствовать масштабам угроз. Однако у абсолютно рационального ИИ могут оказаться иные представления о соответствии масштабов, нежели у нас, частично рациональных людей. При наличии излишних ресурсов ИИ может расширить свою концепцию самосохранения, включив в нее профилактические атаки на источники будущих потенциальных угроз. Достаточно продвинутый ИИ все, что в неопределенном будущем имеет шанс развиться в опасность, может воспринимать как реальную опасность, которую необходимо устранить. Помните к тому же, что машины не думают о времени так, как думаем о нем мы. Если исключить несчастные случаи, достаточно развитые самосовершенствующиеся машины бессмертны. Чем дольше существуешь, тем больше угроз встречаешь. И тем раньше начинаешь их предвидеть. Так что ИСИ, возможно, захочет устранить угрозы, которые не возникнут в ближайшие, скажем, тысячу лет.

Но постойте, разве в число этих опасностей не входит человек? Если не запрограммировать этот вопрос заранее и очень конкретно, то мы, люди, всегда будем представлять собой риск, реальный или потенциальный, для умных машин, которые сами создали, разве не так? Если мы пытаемся избежать рисков, связанных с непредвиденными последствиями создания ИИ, то ИИ будет внимательно рассматривать людей в поисках потенциальных опасностей существования с нами на одной планете.

Рассмотрим искусственный суперинтеллект в тысячу раз умнее самого умного человека. Как отмечалось в главе 1, самое опасное изобретение нашего биологического вида — ядерное оружие. Представляете, какое оружие может придумать разум, в тысячу раз более мощный, чем наш? Разработчик ИИ Хьюго де Гари считает, что потребность защитить себя у будущего ИИ породит в мире катастрофическое политическое напряжение:

> Когда люди увидят, что окружены умными роботами и другими устройствами на основе искусственного мозга, которые

> становятся все совершенней, тревожность возрастет до панического уровня. Начнутся убийства руководителей корпораций, создающих ИИ, поджоги заводов по выпуску роботов, саботаж и т. п.

В научно-популярной книге 2005 г. «Война артилектов» де Гари говорит о будущем, в котором мегавойны будут вспыхивать из-за политических разногласий, порожденных изобретением ИСИ. Нетрудно вообразить себе панику, которая возникнет, если публика реально осознает, к каким последствиям может привести потребность ИСИ в самосохранении. Во-первых, де Гари предполагает, что такие технологии, как ИИ, нанотехнологии, вычислительная нейробиология и квантовые вычисления (при которых в вычислительных процессах участвуют элементарные частицы), объединившись, сделают возможным создание «артилектов», или искусственных интеллектов. Артилекты, размещенные в компьютерах размером с планеты, будут в *триллионы* раз умнее человека. Во-вторых, в политике XXI в. будут доминировать дебаты о том, следует ли строить артилекты. Самые горячие темы:

Станут ли роботы умнее нас? Следует ли человечеству установить верхний предел развития интеллекта роботов или искусственного мозга? Можно ли остановить победное шествие искусственного интеллекта? Если нет, то как скажется на будущем человечества тот факт, что мы станем вторым по разумности видом?

Человечество делится на три лагеря: те, кто хочет уничтожить артилекты, те, кто хочет и дальше их развивать, и те, кто стремится смешаться с артилектами и взять под контроль их ошеломляющие технологии. Победителей не будет. В кульминационной точке сценария де Гари три лагеря сталкиваются в смертельной схватке с использованием разрушительного оружия конца XXI в. Результат? «Гигасмерть» — такой термин пустил в оборот де Гари, описывая гибель *миллиардов* людей.

Возможно, де Гари переоценивает фанатизм антиартилектовых сил, считая, что они начнут войну, в которой почти наверняка погибнут миллиарды людей, чтобы остановить развитие технологии, которая лишь потенциально *может* когда-нибудь убить миллиарды людей. Но мне кажется, что разработчики ИИ правильно ставят вопрос: следует ли нам создавать роботов, которые в конце концов сменят нас? Позиция де Гари ясна:

> Люди не должны стоять на пути более развитой

> эволюционной формы. Эти машины богоподобны. Участь человечества — создать их.

Фактически де Гари сам заложил базу для их создания. Он планирует соединить две методики с «черными ящиками» — нейронные сети и эволюционное программирование — и построить на их основе механический мозг. Предполагается, что его устройство, так называемая машина Дарвина, будет сама разрабатывать собственную архитектуру.

Вторая по степени опасности потребность ИИ — приобретение ресурсов — заставляет систему собирать нужные ей активы, увеличивая таким образом шансы на достижение цели. По мнению Омохундро, при отсутствии точных и подробных инструкций о том, как следует собирать ресурсы, «система не остановится перед кражей, мошенничеством и ограблением банков — ведь это прекрасный способ получать ресурсы». Если ей нужна энергия, а не деньги, она возьмет нашу. Если ей потребуются атомы, а не энергия и не деньги, это опять будут наши атомы.

> Такие системы по природе своей жаждут всего, и побольше. Им нужно вещество, им нужно больше свободной энергии и больше пространства, наконец, потому что со всем этим они смогут более эффективно добиваться цели.

Без всяких наших подсказок мощный ИИ откроет дорогу к новым технологиям добычи ресурсов. Дело за малым: остаться в живых, чтобы иметь возможность пользоваться плодами этих трудов.

> Они обязательно захотят строить реакторы ядерного синтеза и извлекать энергию, заключенную в атомном ядре; кроме того, они захотят исследовать космос. Вы конструируете шахматного робота, но проходит время — и чертова машина хочет строить космический корабль. Ведь именно там, в космосе, можно найти ресурсы, особенно если иметь в виду и очень длинный временной горизонт.

Кроме того, как мы уже говорили, самосовершенствующиеся машины способны жить вечно. В главе 3 мы узнали, что, выйдя из-под контроля, ИСИ могли бы представлять угрозу не только для нашей планеты, но и для Галактики. Собирание ресурсов — потребность, которая обязательно толкнет ИСИ за пределы земной атмосферы. Такой поворот в поведении

рационального агента вызывает в памяти плохие научно-фантастические фильмы. Но посмотрите на причины, толкающие человека в космос: гонка времен холодной войны, дух исследования, американское и советское «предназначение», космическая оборона и развитие промышленного производства в невесомости (в свое время это казалось весьма здравой идеей). Мотивация ИСИ на выход в космос была бы сильнее и ближе к нуждам реального выживания.

> Космос содержит такие несметные богатства, что системы с более длинным временным горизонтом, чем у человека, вероятно, готовы будут выделить значительные ресурсы на исследование и освоение космоса вне зависимости от заявленных целей, — *говорит Омохундро*. — Очень серьезный стимул — преимущество первооткрывателя. Если возникнет конкуренция за обладание космическими ресурсами, то результирующая "гонка вооружений", скорее всего, в конце концов приведет к экспансии со скоростью, приближающейся к скорости света.

Да-да, вы не ошиблись, он сказал *скорости света*. Давайте посмотрим, как мы дошли до жизни такой, ведь начиналось все с робота-шахматиста.

Во-первых, сознающая себя самосовершенствующаяся система будет рациональной. Собирать ресурсы вполне рационально — чем больше у системы ресурсов, тем вероятнее достижение целей и тем легче избегать уязвимости. Если во встроенной в систему шкале целей и ценностей собирание ресурсов никак не ограничено, то система будет искать способы и средства для приобретения как можно большего количества ресурсов. При этом она, удовлетворяя свои потребности, может совершать множество поступков, противоречащих нашим представлениям о машинах, — к примеру, вламываться в компьютеры и даже в банки.

У сознающей себя самосовершенствующейся системы достаточно интеллекта, чтобы проводить научно-исследовательские и опытно-конструкторские работы (НИОКР), необходимые для самомодификации. С ростом интеллекта будет расти и способность к НИОКР. Система будет искать способы или производить роботизированные тела, или обменивать их у людей на товары и услуги, чтобы строить необходимую ей инфраструктуру. Даже космические корабли.

Почему роботизированные тела? Конечно, роботы — старый и довольно избитый сюжетный ход в фильмах, книгах и телепостановках,

своеобразный театральный заменитель искусственного интеллекта. Но роботизированные тела действительно имеют прямое отношение к дискуссиям по ИИ по двум причинам. Во-первых, как мы увидим позже, обладание телом — возможно, наилучший для ИИ способ собирать информацию об окружающем мире. Некоторые теоретики даже считают, что разум, не заключенный в каком-нибудь теле, не в состоянии развиваться. Наш собственный разум — сильный аргумент в пользу этой позиции. Во-вторых, занятый сбором ресурсов ИИ постарается обзавестись роботизированным телом по той же причине, по какой Honda снабдила своего робота ASIMO гуманоидным телом. Это сделано, чтобы робот мог пользоваться нашими вещами.

С 1986 г. ASIMO разрабатывался с целью оказывать помощь по дому пожилым людям — наиболее быстро растущей группе населения Японии. Человеческие формы и умения лучше всего подходят машине, которая должна будет подниматься по лестницам, включать свет, подметать мусор, манипулировать кастрюлями и сковородками — и все это в человеческом жилище. Точно так же ИИ, желающему эффективно использовать наши заводы, здания, транспортные средства и инструменты, придется позаботиться о гуманоидных формах.

А теперь вернемся к космосу.

Мы уже говорили о том, какие необъятные преимущества дадут нанотехнологии сверхразуму и как рациональная система получит мотивацию к их развитию. В космос нашу систему гонит желание достичь целей и избежать уязвимости.

Она просматривает возможные варианты будущего и исключает те, где ее цели не достигаются. *Не воспользоваться* бесконечными на первый взгляд ресурсами космоса — очевидный путь к неудаче.

Как и проигрыш конкурентам в ресурсной гонке. Таким образом, сверхразумная система выделит значительные ресурсы на достижение скорости работы, достаточной для выигрыша в этой гонке. Из этого следует, что, если мы не будем очень осторожны при создании сверхразума, это событие вполне может оказаться шагом в будущее, где могущественные и жадные машины (или их зонды) будут носиться по Галактике с почти световой скоростью, собирая ресурсы и энергию.

Я лично вижу своеобразный черный юмор в том, что первый контакт иной галактической жизни с Землей может выглядеть как бодрое приветствие по радио, а затем опустошительный смертельный десант нанофабрик. В 1974 г. Корнеллский университет, запуская обновленный радиотелескоп Аресибо, передал в пространство так называемое послание

Аресибо. Это послание, разработанное основателем SETI Фрэнсисом Дрейком, астрономом Карлом Саганом и другими энтузиастами, содержало информацию о человеческой ДНК, населении Земли и наших координатах. Радиопередача была направлена на звездное скопление M13, расположенное на расстоянии 25 000 световых лет. Радиоволны путешествуют со скоростью света, так что послание Аресибо доберется до места назначения не раньше чем через 25 000 лет. Может, оно и вовсе не доберется до места — ведь M13 за это время изменит свое положение относительно Земли по сравнению с 1974 г. Разумеется, команда Аресибо это знала, но все же воспользовалась случаем попиарить себя и свой проект.

Тем не менее другие звездные системы могут представлять собой более результативные мишени для посланий с радиотелескопов. И разум, который обнаружит эти послания, может оказаться вовсе не биологическим.

Такую оценку ситуации дает SETI — организация по поиску внеземного разума со штаб-квартирой в Маунтин-Вью (штат Калифорния), всего в нескольких кварталах от штаб- квартиры Google. Эта организация, существующая уже полвека, пытается уловить сигналы внеземного разума с расстояния до 150 трлн км. Для приема инопланетных передач в 450 км к северу от Сан-Франциско построено 42 гигантские тарелки радиотелескопа. SETI *слушает* сигналы — но ничего не посылает, и за полвека сотрудникам этой организации не удалось ничего услышать с других планет. Но одно досадное обстоятельство, наводящее на мысль о возможном распространении ИСИ, удалось установить наверняка: наша Галактика населена слабо, и никто не знает почему.

Главный астроном SETI доктор Сет Шостак выдвинул смелое предположение о том, *что именно* мы можем встретить в космосе, если, конечно, встретим хоть что-нибудь. Это будет искусственный, а не биологический разум.

Он сказал мне:

> То, что мы там ищем, — это эволюционирующий объект. Наши технологические достижения учат нас, что ничто не остается стабильным надолго. Радиоволны, которые мы пытаемся уловить, являются результатом деятельности *биологических* сущностей. Но временное окно между тем моментом, когда цивилизация получает возможность заявить о себе по радио, и тем моментом, когда она начинает строить превосходящие ее

> машины, мыслящие машины, — всего несколько столетий. Не больше. Получается, что каждая цивилизация сама изобретает своих преемников.

Иными словами, для любой разумной формы жизни существует относительно небольшой период времени между двумя техническими вехами — появлением радио и появлением продвинутого ИИ. А стоит создать мощный ИИ — и он быстро берет на себя управление планетой или сливается с изобретателями радио. После этого радио становится ненужным.

Большая часть радиотелескопов SETI направлены на «зоны жизни» ближайших к Земле звезд. Планета в «зоне жизни» расположена достаточно близко к звезде, чтобы вода на ее поверхности могла оставаться в жидком состоянии — не кипела и не замерзала. Вообще, условия на такой планете должны очень точно укладываться в довольно узкие рамки, поэтому иногда «зону жизни» называют также «зоной Златовласки» по имени капризной героини известной сказки.

Шостак утверждает, что SETI следует направить хотя бы часть своих приемников на участки Галактики, условия в которых представляются привлекательными для искусственного, а не биологического разума, — в своеобразные «зоны Златовласки» для ИИ. Это области, богатые энергией, — молодые звезды, нейтронные звезды и черные дыры.

Я думаю, мы могли бы небольшую часть времени наблюдать за теми направлениями, которые, возможно, не слишком привлекательны с точки зрения биологического разума, но где вполне могут обитать разумные машины. У машин иные потребности. У них нет очевидных ограничений на продолжительность существования, поэтому они, очевидно, с легкостью могут занять доминирующее положение в космосе. А поскольку эволюция у них идет намного быстрее биологической, может получиться так, что первые же разумные машины, появившиеся на сцене, в конечном итоге подчинят себе все разумные существа Галактики. Это тот случай, когда «победитель получает все».

Шостак связывает современные облачные сервисы, организованные Google, Amazon и Rackspace, с высокоэнергетичными сверххолодными средами, которые потребуются сверхразумным машинам. Один из примеров таких замороженных сред — глобулы (темные газопылевые туманности с температурой около -263 °C, один из самых холодных объектов во Вселенной). Подобно сегодняшним вычислительным облакам Google, горячие мыслящие машины будущего, вероятно, будут нуждаться в

охлаждении; в противном случае они могут попросту расплавиться.

Предложения Шостака о том, где искать ИИ, говорят нам, что идея о разуме, покидающем Землю в поисках ресурсов, способна зажечь и менее живое воображение, чем у Омохундро и ребят из MIRI. Однако Шостак, в отличие от них, не думает, что сверхразум будет опасен.

> Если мы в ближайшие пять лет построим машину с интеллектуальными возможностями одного человека, то ее преемник уже будет разумнее всего человечества вместе взятого. Через одно-два поколения они попросту перестанут обращать на нас внимание. Точно так же, как вы не обращаете внимания на муравьев у себя во дворе. Вы не уничтожаете их, но и не приручаете, они практически никак не влияют на вашу повседневную жизнь, но они там есть.

Проблема в том, что я, к примеру, *уничтожаю* муравьев у себя на заднем дворе, особенно когда они прокладывают тропки ко мне на кухню. Но есть и разница — ИСИ улетит в Галактику или отправит туда зонды, потому что нужные ему ресурсы на Земле уже закончились или он рассчитал, что они закончатся в ближайшее время, — и дорогостоящие полеты в космос оправданны. А если так, то почему мы все еще будем к тому моменту живы, если на наше существование тратится, вероятно, немало все тех же ресурсов? И не забывайте: сами мы тоже состоим из вещества, для которого ИСИ может найти свое применение.

Короче говоря, чтобы хеппи-энд по Шостаку стал возможен, сверхразум, о котором идет речь, должен *захотеть* оставить нас в живых. Просто не обращать на нас внимания недостаточно. Но до сих пор не существует ни общепринятой этической системы, ни понятного способа вложить такое желание в продвинутый ИИ.

Но существует молодая наука о поведении сверхразумного агента, и начало ей положил Омохундро.

Мы уже рассмотрели три потребности из четырех, мотивирующих, по мнению Омохундро, сознающую себя самосовершенствующуюся систему: это эффективность, самозащита и ресурсы. Мы увидели, что без тщательнейшего планирования и программирования каждая из этих потребностей приведет к весьма печальным результатам. И мы вынуждены спросить себя: способны ли мы на столь тщательную работу? Или вы, как и я, оглядываетесь вокруг, видите всевозможные случайные события и происшествия, которые дорого обходятся человечеству и в денежном

измерении, и в человеческих жизнях, и думаете о том, сможем ли мы, имея дело с невероятно мощным ИИ, с первого раза сделать все правильно? Тримайл-Айленд, Чернобыль, Фукусима — разве эти катастрофы произошли не на атомных станциях, где высококвалифицированные конструкторы и администраторы всеми силами старались их избежать? Чернобыльский реактор в 1986 г. взорвался во время эксперимента в области *безопасности.*

Все три перечисленные катастрофы — это то, что специалист по теории организаций Чарльз Перроу назвал бы «нормальными авариями». В своей эпохальной книге «Нормальные аварии: Жизнь с технологиями высокого риска» Перроу говорит о том, что аварии и даже катастрофы «нормальны» для систем со сложной инфраструктурой. Для таких систем характерен высокий уровень непредсказуемости, поскольку отказ может произойти не в одном, а в нескольких, часто не связанных между собой процессах. Отдельные ошибки, ни одна из которых не была бы фатальна сама по себе, складываются, порождая системные отказы, предсказать которые заранее было бы попросту невозможно.

На АЭС Тримайл-Айленд 28 марта 1979 г. катастрофу вызвали четыре несложных отказа: два насоса системы охлаждения остановились из-за механических неполадок; два аварийных насоса не могли работать, поскольку их задвижки были закрыты на техобслуживание; ярлычок, сообщавший о ремонте, прикрыл собой индикаторные лампы, которые могли бы предупредить персонал о проблеме; наконец, предохранительный клапан застрял в открытом состоянии, а неисправный световой индикатор показывал, что этот самый клапан закрыт. Итоговый результат: активная зона реактора расплавилась, едва удалось избежать человеческих жертв, а ядерной энергетике США был нанесен едва ли не фатальный удар.

Перроу пишет:

> Мы породили конструкции настолько сложные, что не в состоянии предвидеть все возможные сочетания неизбежных отказов; мы добавляем все новые устройства безопасности, которые вводятся в заблуждение, обходятся или подавляются скрытыми в системах способами.

Особенно уязвимы, пишет Перроу, системы, чьи компоненты тесно переплетены (связаны «сверхкритической связью»), то есть непосредственно и значительно влияют друг на друга. Одним из ярких примеров опасностей, связанных с тесно переплетенными системами ИИ,

могут служить события мая 2010 г. на Уолл-стрит.

В наше время до 70 % всех сделок с акциями на Уолл-стрит осуществляются примерно 80 компьютеризированными системами высокочастотного трейдинга (HFT). Речь идет примерно о миллиарде акций в день. Трейдинговые алгоритмы и компьютеры, на которых они работают, принадлежат банкам, хедж-фондам и фирмам, существующим исключительно для высокочастотного трейдинга. Смысл HFT — получать прибыль от возможностей, возникающих буквально на доли секунды (к примеру, когда меняется цена на одни ценные бумаги, а другие цены, которые по идее должны быть ей эквивалентны, не успевают мгновенно поменяться), и использовать ежедневно *множество* подобных ситуаций.

В мае 2010 г. у Греции возникли проблемы с рефинансированием национального долга. Европейские страны, одалживавшие Греции деньги, опасались дефолта. Долговой кризис ослабил европейскую экономику и породил опасности для американского рынка. А причиной всего этого стал испуганный трейдер неизвестной брокерской компании, который одномоментно выставил на продажу фьючерсные контракты и инвестиционные фонды ETF, имеющие отношение к Европе, на $4,1 млрд.

После этой продажи стоимость фьючерсных контрактов типа E-Mini S&P 500 упала на 4 % за четыре минуты. Алгоритмы высокочастотного трейдинга уловили падение цены. Пытаясь удержать прибыль, они автоматически запустили распродажу, которая заняла несколько миллисекунд (самая быстрая на данный момент сделка заняла три миллисекунды — три тысячные доли секунды). В ответ на более низкую цену *другие* HFT автоматически начали *покупать* E-Mini S&P 500 и продавать другие ценные бумаги, чтобы получить деньги на их покупку. Прежде чем люди успели вмешаться, началась цепная реакция, в результате которой индекс Доу-Джонса упал на 1000 пунктов. На все про все ушло двадцать минут.

Перроу называет эту проблему «непостижимостью». Причиной нормальных аварий, как правило, являются взаимодействия, которые «не только неожиданны, но и непонятны в течение некоторого критически важного времени». Никто не предвидел, что события могут так подействовать друг на друга, поэтому никто вовремя не понял, что происходит.

Специалист по финансовым рискам Стив Охана признал существование проблемы. «Это новый риск, — сказал он. — Мы знаем, что многие алгоритмы взаимодействуют друг с другом, но не знаем, как именно. Мне кажется, мы слишком далеко зашли в компьютеризации

финансов. Мы не в состоянии контролировать монстра, которого создали».

Этот монстр снова нанес удар 1 августа 2012 г.: из-за неудачного HFT-алгоритма инвестиционная фирма Knight Capital Partners потеряла за полчаса $440 млн.

На мой взгляд, в этих кризисах можно разглядеть элементы неизбежных ИИ-катастроф: чрезвычайно сложные, почти непознаваемые ИИ-системы, непредсказуемое взаимодействие с другими системами и более широкой информационно-технической средой и, наконец, ошибки, возникающие на фоне огромных вычислительных скоростей, делают вмешательство человека бессмысленным.

«Агент, стремящийся только к удовлетворению потребностей в эффективности, самосохранении и сборе ресурсов, действовал бы как одержимый социопат-параноик», — пишет Омохундро в книге «Природа самосовершенствующегося искусственного интеллекта». Очевидно, сосредоточенность исключительно на работе играет с ИИ плохую шутку и делает общение с ним достаточно неприятным. Робот, обладающий только перечисленными потребностями, был бы механическим Чингисханом; он стремился бы захватить все ресурсы Галактики, лишить конкурентов средств к существованию и уничтожить врагов, которые еще, по крайней мере, тысячу лет не представляли бы для него ни малейшей опасности. Но у нас осталась еще одна первичная потребность, которую следует добавить в это жуткое зелье: потребность в творчестве.

Под влиянием четвертой потребности ИИ изобретал бы новые способы более эффективного достижения целей или, скорее, избегания исходов, при которых его цели будут удовлетворяться не так оптимально, как могли бы. Потребность в творчестве означала бы меньшую предсказуемость системы (еще меньшую?!), потому что креативные идеи *оригинальны.* Чем умнее система, тем оригинальнее путь к цели и тем дальше он отстоит от всего, что мы могли бы себе представить. Потребность в творчестве помогала бы с максимальной отдачей использовать прочие потребности — эффективность, самосохранение и накопление ресурсов — и предлагала бы обходные пути в тех случаях, когда с удовлетворением потребностей возникают проблемы.

Представьте, к примеру, что основная цель вашего робота-шахматиста — выигрывать в шахматы у любого оппонента. Столкнувшись с другим шахматным роботом, он немедленно проникнет в его систему и снизит скорость работы процессора до черепашьей, что даст вашему роботу решительное преимущество. Вы восклицаете: «Погодите минутку, я ничего такого не имел в виду!» — и добавляете в своего робота программу,

которая на уровне подзадачи запрещает ему вмешиваться в системы противников. Однако еще до следующей игры вы обнаруживаете, что ваш робот *строит* робота-помощника, который затем внедряется в систему противника! Если вы запретите ему строить роботов, он *возьмет робота в аренду* или *наймет* кого-нибудь! Без подробнейших ограничивающих инструкций сознающая себя, самосовершенствующаяся система, ориентированная на решение определенной задачи, дойдет на своем пути к цели до нелепых с нашей точки зрения крайностей.

Это лишь один пример проблемы непредвиденных последствий, связанной с ИИ, — проблемы настолько серьезной и вездесущей, что ее можно было бы, пожалуй, сравнить с «проблемой воды» в отношении к морским судам. Мощная ИИ-система, цель которой — обеспечить вашу защиту, может попросту запереть вас дома и никуда не выпускать. Если вы попросите счастья, она может подсоединить вас к системе искусственного жизнеобеспечения и все время стимулировать центры удовольствия в вашем мозгу. Если вы не обеспечите ИИ очень большой библиотекой приемлемых вариантов поведения или надежнейшим средством выбора предпочтительного для вас варианта, вам придется терпеть тот вариант, который он выберет сам. А поскольку ИИ — чрезвычайно сложная система, вы, возможно, никогда не сумеете понять ее достаточно хорошо, чтобы быть уверенным в правильности своих представлений о ней. Чтобы понять, не собирается ли ваш снабженный ИИ робот привязать вас к кровати и засунуть электроды в уши, пытаясь обеспечить вам безопасность и счастье, может потребоваться *еще один*, более умный ИИ.

Существует еще один подход к проблемам первичных потребностей ИИ, причем такой, который предпочтительнее для позитивно настроенного Омохундро. Потребности дают возможности — двери, которые открываются перед человечеством и нашими мечтами. Если мы не хотим, чтобы наша планета, а со временем и Галактика, были населены исключительно эгоцентричными и бесконечно воспроизводящими себя сущностями с чингисхановским отношением к биологическим существам и друг к другу, то творцам ИИ следовало бы сформулировать цели, органично включающие в себя человеческие ценности. В списке пожеланий Омохундро можно найти, в частности, следующие пункты: «делать людей счастливыми», «сочинять красивую музыку», «развлекать других», «работать над сложной математикой» и «создавать возвышающие произведения искусства». А затем нужно будет отойти в сторону. С такими целями творческая потребность ИИ включится на полную мощность и откликнется прекрасными творениями, обогащающими нашу жизнь.

Как человечеству сохранить человечность? Это чрезвычайно интересный и важный вопрос, который мы, люди, задаем в разных формах уже очень давно. Что такое доблесть, праведность, совершенство? Какое искусство возвышает и какая музыка красива? Необходимость точно определить наши ценности — один из моментов, которые помогают нам лучше узнать самих себя в процессе поиска путей к созданию искусственного интеллекта человеческого уровня. Омохундро считает, что такое глубокое погружение в себя и самокопание приведет к созданию обогащающих, а не ужасающих технологий. Он пишет:

> Вооружившись одновременно логикой и вдохновением, мы можем двигаться к созданию техники, которая усилит, а не ослабит дух человеческий.

*** * ***

Разумеется, я иначе смотрю на вещи — я не разделяю оптимизма Омохундро. Но я сознаю необходимость и важность развития науки, которая помогла бы нам разобраться в наших разумных созданиях. Не могу не повторить его предупреждение, касающееся продвинутого ИИ.

> Не думаю, что большинство исследователей ИИ видят какую-то потенциальную опасность в создании, скажем, робота-шахматиста. Но мой анализ показывает, что нам следует тщательно обдумать, какие ценности мы в него закладываем; в противном случае мы получим нечто вроде психопатической, эгоистичной, зацикленной на себе сущности.

Мои личные впечатления свидетельствуют о том, что Омохундро прав относительно создателей ИИ: те, с кем я разговаривал, заняты исключительно гонкой и стремятся побыстрее создать разумные системы; они не считают, что результат их работы может быть опасен. При этом большинство из них в глубине души убеждены, что машинный интеллект со временем придет на смену человеческому интеллекту. Они не думают о том, как именно произойдет такая замена.

Разработчики ИИ (а также теоретики и специалисты по этике) склонны верить, что разумные системы будут делать только то, на что они запрограммированы. Но Омохундро говорит, что они, конечно, будут делать

это, но будут делать и многое другое, и мы можем предугадать с некоторой вероятностью, как поведут себя продвинутые ИИ-системы. И их поведение может оказаться неожиданным и креативным. В основе этих размышлений лежит настолько пугающе простая концепция, что потребовалось настоящее озарение (в лице Омохундро), чтобы ее разглядеть: *для достаточно разумной системы уход от уязвимостей — мотиватор не менее мощный, чем специально встроенные в нее цели и подцели.*

Мы должны остерегаться непредусмотренных последствий, вызванных теми целями, на которые мы программируем разумные системы; кроме того, мы должны остерегаться последствий того, что мы оставляем за рамками этих целей

Глава 7
Интеллектуальный взрыв

С точки зрения экзистенциальных рисков один из важнейших моментов, связанных с искусственным интеллектом, — то, что искусственный интеллект в принципе может наращивать свою интеллектуальность чрезвычайно быстро. Очевидная причина подозревать такую возможность — рекурсивное самосовершенствование (Гуд, 1965). ИИ становится умнее, в том числе и в написании внутренних когнитивных функций для ИИ, так что ИИ может переписать свои когнитивные функции, чтобы они работали еще лучше, что сделает этот ИИ еще умнее, в том числе и в переписывании самого себя, так что он проведет новые улучшения… Ключевое для наших целей следствие из всего этого — то, что любой ИИ может совершить гигантский скачок в уровне интеллекта после достижения некоего критического порога.

Елиезер Юдковски, научный сотрудник Исследовательского института машинного интеллекта

Возможно, вы имели в виду: рекурсия.

Поисковик Google на запрос «рекурсия»

До сих пор мы с вами рассматривали ИИ-сценарий настолько катастрофический, что его просто необходимо исследовать подробнее. Мы обсудили перспективную идею о том, как следует конструировать ИИ, чтобы избежать всякой опасности (речь идет о дружественном ИИ), и выяснили, что эта концепция неполна. Более того, сама идея запрограммировать разумную систему на абсолютно безопасные, раз и навсегда заданные цели или дать ей развиваемую способность генерировать для себя цели, которые на протяжении большого числа

итераций оставались бы безопасными для нас, представляется попросту утопической.

Далее мы выяснили, почему вообще ИИ может стать опасным. Мы обнаружили, что многие потребности, которые должны мотивировать сознающую себя самосовершенствующуюся компьютерную систему, легко могут привести к катастрофическим последствиям для людей. Трагические исходы этих сценариев подчеркивают почти библейскую опасность греха как деянием, так и не деянием в процессе подверженного ошибкам человеческого программирования.

ИИ человеческого уровня, когда будет создан, может оказаться непредсказуемым и опасным, но, вероятно, в краткосрочной перспективе не катастрофически. Даже если УЧИ изготовит множество копий самого себя или организует командную работу по вопросу освобождения, его возможности, в том числе и опасные, все же не будут превышать возможности группы умных людей. Потенциальная опасность УЧИ заключается в основе сценария Busy Child — стремительном рекурсивном самосовершенствовании, которое позволит ИИ быстро «прокачать» себя с человеческого уровня до уровня сверхразума. Этот процесс обычно и называют «интеллектуальным взрывом».

Сознающая себя самосовершенствующаяся система будет стремиться к наилучшему выполнению поставленных перед ней задач и минимизации уязвимостей; для этого она будет улучшать себя. Она будет стремиться не к мелким улучшениям, а к серьезному непрерывному развитию всех аспектов своих когнитивных способностей, особенно отвечающих за развитие интеллекта. Она будет стремиться к интеллекту уровня выше человеческого, то есть к суперинтеллекту. Если написанные людьми программы окажутся хоть чуточку неидеальными, у нас будет множество оснований для страха перед сверхразумными машинами.

От Стива Омохундро мы знаем, что УЧИ будет естественным образом стремиться к интеллектуальному взрыву. Но *что такое* интеллектуальный взрыв? При выполнении каких минимальных аппаратных и программных требований он может произойти? Могут ли такие факторы, как недостаточное финансирование или просто сложность достижения вычислительной разумности, навсегда блокировать интеллектуальный взрыв?

Прежде чем обратиться к механике этого процесса, важно разобраться, что, собственно, означает этот термин и как математик Ирвинг Гуд предложил идею взрывного искусственного разума.

Американское шоссе № 81 начинается в штате Нью-Йорк и

заканчивается в Теннесси, пересекая почти всю гряду Аппалачских гор. Повернув в центре Вирджинии на юг, трасса змеей извивается вверх и вниз по холмам, густо заросшим лесом, и травянистыми равнинам; отсюда открываются самые поразительные, мрачные и первобытные виды США. Аппалачи включают в себя Голубой хребет (от Пенсильвании до Джорджии) и хребет Грейт-Смоки (вдоль границы Северной Каролины и Теннесси). Чем дальше на юг, тем труднее поймать сигнал сотового телефона, а церквей вокруг становится больше, чем домов; музыка кантри на радио сменяется сперва духовными песнопениями, а затем и речами проповедников, вещающих о геенне огненной. Именно там я услышал памятную песню Джоша Тёрнера об искушении под названием «Длинный черный поезд». Там я слышал, как проповедник начал рассказ с Авраама и Исаака, запутался и закончил притчей о хлебах и рыбах; не забыл он упомянуть и про ад, просто для порядка. Я подъезжал к Смоки-Маунтинз, к границе с Северной Каролиной и Вирджинскому политехническому институту и Университету штата Вирджиния в Блэксбурге. Девиз этого университета выглядит так: «Изобретай будущее».

Если бы вы проезжали по этому шоссе двадцать лет назад (а шоссе I-81 за прошедшие годы почти не изменилось), вас вполне мог бы обогнать кабриолет Triumph Spitfire с особым заказным номером — 007 IJG. Номер принадлежал Ирвингу Гуду — заслуженному профессору статистики, прибывшему в Блэксбург в 1967 г. Номер «007» отсылал, естественно, к Яну Флемингу и секретной работе самого Гуда в качестве дешифровщика в Блетчли-парк (Англия) во время Второй мировой войны. Взлом системы шифров, которой пользовались вооруженные силы Германии, внес существенный вклад в поражение держав «оси». В Блетчли-парке Гуд работал с Аланом Тьюрингом, которого называют отцом современной вычислительной техники (он же — создатель одноименного теста, о котором шла речь в главе 4), и участвовал в создании и программировании одного из первых электрических вычислительных устройств.

В Блэксбурге Гуд считался знаменитостью — его жалованье было выше, чем жалованье президента местного университета. Он всегда был неравнодушен к числам и сразу заметил, что приехал в Блэксбург в седьмом часу седьмого дня седьмого месяца седьмого года седьмого десятилетия, а поселили его в седьмой квартире седьмого квартала местного жилого комплекса. Гуд говорил друзьям, что Бог посылает подобные совпадения атеистам, таким как он, чтобы убедить их в своем существовании.

«У меня есть не до конца оформленная идея, что Бог посылает

человеку тем больше совпадений, чем больше тот сомневается в его существовании, предоставляя таким образом свидетельства, но не заставляя верить, — говорил Гуд. — Когда я поверю в эту теорию, совпадения, надо понимать, прекратятся».

Я ехал в Блэксбург, чтобы расспросить о Гуде его друзей (сам он недавно умер в возрасте 92 лет). В основном мне хотелось узнать, как Гуд пришел к идее интеллектуального взрыва и возможен ли такой взрыв на самом деле. Концепция интеллектуального взрыва стала первым крупным звеном в цепочке, которая в конце концов породила гипотезу сингулярности.

К несчастью, в обозримом будущем любое упоминание Технического университета Вирджинии будет вызывать в памяти устроенную здесь бойню. 16 апреля 2007 г. старшекурсник Сёнг-Ху Чо, специализировавшийся на изучении английского языка, убил тридцать два студента и сотрудника университета и ранил еще двадцать пять человек. Это самое кровавое преступление стрелка-одиночки в истории США. Если говорить коротко, то сначала Чо застрелил студентку-младшекурсницу в общежитии университета (в корпусе Эмбер Джонстон), а затем и студента, который пытался ей помочь. Два часа спустя Чо начал бойню в инженерном корпусе (Норрис-холле) университета, в котором и пострадало большинство людей. Прежде чем начать стрельбу, Чо накрепко запер тяжелые дубовые двери корпуса, чтобы никто не мог убежать.

Когда давний друг и коллега-статистик Гуда доктор Голд Хольцман показал мне бывший кабинет Гуда по другую сторону зеленой лужайки Дриллфилда (там когда-то располагался военный плац университета), я обратил внимание на то, что из окон кабинета вдалеке виден Норрис-холл. Но к моменту трагедии, сказал мне Хольцман, Гуд был уже в отставке. Он находился не в кабинете, а дома, может быть, рассчитывал вероятность существования Бога.

По словам доктора Хольцмана, незадолго до смерти Гуд увеличил эту вероятность с нуля до одной десятой. Сделал он это потому, что как статистик был давним последователем Байеса. Основная идея байесовской статистики, названной в честь математика и священника XVIII в. Томаса Байеса, состоит в том, что вычисление вероятности некоего утверждения можно начинать с того, во что вы лично верите. Затем эту веру следует подправлять в зависимости от новых данных, подтверждающих или опровергающих ваше утверждение.

Если бы первоначальное *неверие* Гуда в Бога осталось стопроцентным, то никакие данные и даже явление самого Господа ничего бы не изменили.

Поэтому, чтобы быть верным байесовскому подходу, Гуд ввел небольшую положительную вероятность существования Бога; это позволяло ему быть уверенным, что новые данные, если таковые появятся, не останутся неучтенными.

В работе 1965 г. «Размышления о первой ультраразумной машине» Гуд изложил простое и элегантное доказательство, которое часто упоминается в дискуссиях об искусственном интеллекте и сингулярности:

> Определим ультраразумную машину как машину, способную намного превзойти любую интеллектуальную деятельность человека, каким бы умным он ни был. Поскольку конструирование машин — одно из интеллектуальных действий, ультраразумная машина способна конструировать все более совершенные машины; затем, бесспорно, произойдет "интеллектуальный взрыв", и человеческий разум останется далеко позади. Таким образом, первая ультраразумная машина — это последнее изобретение, которое потребуется от человека…

Известно три проработанных определения сингулярности — первое из них принадлежит Гуду и приведено выше. Гуд никогда не использовал термин «сингулярность», но его постулат о том, что сам Гуд считал неизбежной и позитивной вехой в истории человечества, — об изобретении машин умнее человека, — положил начало дискуссиям о сингулярности. Перефразируя Гуда, если вы построите сверхразумную машину, она будет лучше человека справляться со всем, для чего мы используем свой мозг, в том числе и со строительством сверхразумных машин. Поэтому первая такая машина запустит интеллектуальный взрыв — стремительный рост интеллекта, — по мере того как будет раз за разом совершенствовать себя или просто строить машины умнее себя. Эта машина или машины оставят мощь человеческого разума далеко позади. После интеллектуального взрыва человеку уже не нужно будет ничего изобретать — все его потребности будут удовлетворять машины.

Этот абзац из работы Гуда справедливо находит место в книгах, статьях и очерках о сингулярности, будущем искусственного интеллекта и его рисках. Но две важные мысли почти всегда почему-то остаются за рамками дискуссии. Первая мысль сформулирована в первом же вводном предложении статьи. Она великолепна: «Выживание человечества зависит от скорейшего создания ультраразумной машины». Вторая мысль — часто опускаемая *вторая половина* последнего предложения процитированного

абзаца. Последнее предложение наиболее часто цитируемого отрывка из Гуда *следует* читать с полным вниманием:

> Таким образом, первая ультраразумная машина — это последнее изобретение, которое потребуется от человека, *если, конечно, эта машина будет достаточно сговорчивой, чтобы сообщить нам, как можно ее контролировать (курсив мой. — Авт.).*

Эти два предложения рассказывают нам кое-что важное о намерениях Гуда. Он считал, что у нас, людей, столько сложнейших неотложных проблем — гонка ядерных вооружений, загрязнение окружающей среды, войны и т. п., — что спасти нас может только высочайший интеллект (выше нашего), который воплотится в виде сверхразумной машины. Второе предложение говорит о том, что отец концепции интеллектуального взрыва остро чувствовал опасность: создание сверхразумных машин, даже если они необходимы для выживания человечества, может обернуться против нас. Возможность удерживать сверхразумную машину под контролем — вовсе не данность, говорит Гуд. При этом он не считает, что мы сами придумаем, как это делать, — машина должна будет нас *научить.*

Гуд, несомненно, знал кое-что о машинах, способных спасти мир, — ведь в свое время в Блетчли-парке во время войны с Германией он участвовал в создании первых электрических вычислителей и работе на них. Он также знал кое-что об экзистенциальных рисках — он был евреем и сражался с нацистами, а его отец бежал от погромов из Польши в Великобританию.

Мальчиком отец Гуда — поляк и интеллектуал-самоучка — изучил ремесло часовщика, наблюдая за работой часовщиков через окна витрин. Ему было семнадцать лет, когда в 1903 г. он отправился в Англию с тридцатью пятью рублями в кармане и большим кругом сыра. В Лондоне юноша пробавлялся случайными заработками, пока не открыл собственную ювелирную мастерскую. Дела пошли успешно, он женился. В 1915 г. родился Исидор Якоб Гудак (ставший позже Ирвингом Джоном Гудом по прозвищу Джек). После него в семье родились еще один мальчик и девочка — талантливая танцовщица, погибшая позже в театре при пожаре. Ее ужасная смерть заставила Джека Гуда отвергнуть существование Бога.

Гуд был математическим вундеркиндом; однажды он встал в своей детской кроватке и спросил у матери, сколько будет тысячу раз по тысяче. Лежа в постели с дифтерией, он независимо открыл иррациональные числа

(те, которые невозможно записать в виде простой дроби, такие как

$$\sqrt{2}).$$

К 14 годам он заново открыл математическую индукцию — один из методов математического доказательства. К тому моменту учителя-математики просто оставили его наедине с книгами. В Кембриджском университете Гуд завоевал все математические награды, возможные на пути к степени доктора философии, и открыл в себе страсть к шахматам.

Именно из-за игры в шахматы через год после начала Второй мировой войны тогдашний чемпион Британии по шахматам Хью Александер пригласил Гуда в 18-й корпус в Блетчли-парке, где работали дешифровальщики. Они занимались взломом кодов, которыми пользовались все державы «оси» — Германия, Япония и Италия, — но особое внимание всегда уделялось германским кодированным сообщениям. Германские подлодки топили торговые суда союзников с устрашающей скоростью — только за первую половину 1942 г. жертвами подлодок стали около 500 союзных судов. Британский премьер-министр Уинстон Черчилль опасался, что голод может обречь его островную страну на поражение.

Немецкие сообщения передавались по радио, и англичане без особого труда перехватывали их при помощи специальных «слуховых» вышек. С самого начала войны Германия применяла для шифрования сообщений специальную машину под названием «Энигма». Эта машина, имевшаяся во всех шифровальных подразделениях германских вооруженных сил, по форме и размеру напоминала старомодную механическую пишущую машинку. Каждая клавиша соответствовала букве. При нажатии на клавишу ток проходил электрическую цепь и зажигал лампочку с кодовой буквой. В цепи находились вращающиеся барабаны, что позволяло замыкать цепь в различных комбинациях. В базовом варианте «Энигмы» было три барабана, и каждый из них мог закодировать букву предыдущего барабана. Для алфавита из 26 букв возможны были 403 291 461 126 605 635 584 000 вариантов. Барабаны (шаблоны) менялись почти ежедневно.

Когда немец посылал зашифрованное «Энигмой» сообщение, получатель расшифровывал его при помощи собственной «Энигмы»; для этого достаточно было иметь такие же шаблоны, как у отправителя.

К счастью, в Блетчли-парке оказалось собственное «секретное оружие» — Алан Тьюринг. До войны Тьюринг изучал математику и криптографию в Кембридже и Принстоне. Он придумал — вообразил — «автоматическую машину», ныне известную как машина Тьюринга. Эта

автоматическая машина заложила фундаментальные принципы машинных вычислений.

Гипотеза Чёрча-Тьюринга, объединившая работы Тьюринга и принстонского профессора-математика Алонсо Чёрча, дала серьезный толчок к исследованию искусственного интеллекта. Эта гипотеза утверждает, что все, что может быть вычислено по алгоритму или по программе, может быть вычислено машиной Тьюринга. Исходя из этого, если мозговые процессы могут быть выражены в виде серии команд — алгоритма, то компьютер может обрабатывать информацию в точности так же, как мозг. Иными словами, если в человеческом мышлении нет ничего мистического или магического, то разум (интеллект) может быть воплощен в компьютере. Понятно, что разработчики УЧИ возлагают свои надежды на гипотезу Чёрча-Тьюринга.

Война преподала Тьюрингу интенсивный экспресс-курс всего того, о чем он думал до войны, и многого такого, о чем он *не думал* (к примеру, нацизма и подводных лодок). В разгар войны сотрудники Блетчли-парка расшифровывали порядка 4000 перехваченных сообщений в день, и делать это вручную становилось все сложнее. Это была работа для машины. И очень кстати пришлась принципиальная догадка Тьюринга о том, что проще понять, чем *не являются* шаблоны «Энигмы», чем разобраться, чем они *являются.*

У дешифровщиков был материал для работы — перехваченные сообщения, «взломанные» вручную или при помощи электрических дешифровальных машин под названием Bombes. Такие сообщения сотрудники Парка называли «поцелуями». Тьюринг, как и Гуд, был убежденным последователем Байеса в те времена, когда статистические методы воспринимались как своего рода волшебство. Суть метода — теорема Байеса — говорит о том, как извлекать из данных вероятности неизвестных событий, в данном случае тех или иных шаблонов «Энигмы». «Поцелуи» давали дешифровщикам те самые данные, при помощи которых можно было определить, какие варианты шаблонов имеют очень низкую вероятность, — и, соответственно, более эффективно сосредоточить усилия. Конечно, шифры менялись чуть ли не каждый день, так что работа в Блетчли-парке напоминала непрерывную гонку.

Тьюринг с коллегами разработал серию электронных устройств, которые должны были оценивать и исключать возможные шаблоны «Энигмы». Вершиной развития этих первых компьютеров стала серия машин под общим названием «Колосс» (Colossus). «Колосс» способен был считывать 5000 знаков в секунду с бумажной ленты, которая протягивалась

со скоростью около 40 км в час. В нем было 1500 вакуумных ламп, и занимал он целую комнату. Одним из главных пользователей этой машины и создателем половины теории, на которую опиралась ее работа, был главный статистик Тьюринга на протяжении почти всего военного времени — Ирвинг Джон Гуд.

Благодаря героям Блетчли-парка Вторая мировая война, вероятно, стала короче на два-четыре года[15], что позволило сохранить бесчисленное количество жизней. Но в честь этих секретных воинов не устраивали парадов. Черчилль приказал разбить все шифровальные машины Блетчли-парка на куски не крупнее теннисного мячика, чтобы их дешифровальные возможности нельзя было обернуть против Великобритании. Дешифровщики поклялись хранить молчание в течение *тридцати лет*. Тьюринг и Гуд были приглашены на работу в Манчестерский университет, где их бывший шеф Макс Ньюман собирался строить вычислитель общего назначения. Тьюринг работал в Национальной физической лаборатории над конструкцией компьютера, когда его жизнь внезапно полетела под откос. Приятель, с которым у него какое-то время были близкие отношения, ограбил его дом. Сообщив о преступлении в полицию, Тьюринг рассказал и о сексуальных отношениях. Его обвинили в грубой непристойности и лишили допуска к секретным материалам.

В Блетчли Тьюринг и Гуд часто обсуждали такие футуристические идеи, как компьютеры, разумные машины и «автоматический» шахматист. Шахматы сблизили их, и Гуд обычно выигрывал. В ответ Тьюринг научил его го — азиатской стратегической игре, в которой Гуд тоже выиграл. Тьюринг, бегун-стайер мирового класса, придумал особую форму шахмат, в которых можно было уравновесить шансы разноуровневых игроков. После каждого хода игрок должен был обежать вокруг сада. Он получал *два* хода подряд, если успевал вернуться прежде, чем его противник сделает ход.

Вынесенный в 1952 г. Тьюрингу приговор удивил Гуда — он не знал о гомосексуальности Тьюринга. Тьюрингу пришлось выбирать между тюрьмой и химической кастрацией. Он выбрал последнее и должен был регулярно являться на уколы эстрогена. В 1954 г. он съел яблоко, начиненное цианидом. Безосновательный, но упорный слух связывает логотип фирмы Apple именно с этим яблоком.

После истечения срока секретности Гуд одним из первых выступил против того, как правительство обошлось с его другом и героем войны.

«Я не стану говорить, что работа Тьюринга помогла нам выиграть войну, — сказал Гуд. — Но я осмелюсь все же заметить, что без него мы могли ее проиграть». В 1967 г. Гуд ушел из Оксфордского университета и

принял предложенное ему место в Вирджинском политехническом институте в Блэксбурге. Ему тогда было 52 года. В последующие годы жизни он возвращался в Великобританию лишь однажды

В той поездке в 1983 г. его сопровождала высокая 25-летняя красавица-блондинка из Теннесси по имени Лесли Пендлтон. Гуд встретил Пендлтон в 1980 г., сменив до этого десять секретарш за тринадцать лет. Девушка, выпускница Вирджинского политехнического института, сумела удержаться на этом месте и не сломаться под давлением безжалостного перфекционизма Гуда. Когда она впервые отправляла одну из его статей в математический журнал, рассказала мне Пендлтон, «он внимательно наблюдал, как я положила статью и сопроводительное письмо в конверт. Он внимательно наблюдал, как я этот конверт запечатала, — он не любил, когда клей смачивают слюной, и заставил меня воспользоваться губкой. Он внимательно наблюдал, как я наклеиваю марку. Он ждал меня у двери в почтовую комнату, чтобы убедиться, что отправка прошла нормально, как будто меня могли украсть по дороге или еще что-то могло случиться. Он был странным человечком».

Гуд хотел жениться на Пендлтон. Однако для нее сорокалетняя разница в возрасте оказалась слишком серьезным препятствием. Тем не менее между английским чудаком и красавицей из Теннесси сложились отношения, которые она даже сейчас с трудом может описать. На протяжении тридцати лет она сопровождала его на отдыхе, следила за всеми его бумажками и подписками, вела его дела как на работе, так и после отставки до самой смерти, включая и заботу об ухудшающемся здоровье. При встрече она показала мне дом Гуда в Блэксбурге — кирпичное одноэтажное здание с пологой крышей возле федеральной трассы № 460, которая в те времена, когда он только поселился в Теннесси, была всего лишь двухполосной сельской дорогой.

Сегодня Лесли Пендлтон — стройная женщина пятидесяти с чем-то лет, доктор философии и мать двух взрослых детей. Она профессор и администратор в Вирджинском политехническом, владычица расписаний, аудиторий и профессорских прихотей, к которым работа с Гудом ее отлично подготовила. И несмотря на то, что она вышла замуж за ровесника и вырастила детей, многие местные жители ставили под сомнение характер ее отношений с Гудом. Ответ Лесли Пендлтон дала в 2009 г. в надгробном слове на похоронах Гуда. Нет, между ними никогда не было романтических отношений, сказала она, но они всегда были преданы друг другу. В лице Пендлтон Гуд не обрел возлюбленной, но обрел лучшего друга на тридцать лет жизни и непреклонного защитника своего наследия и памяти.

Во дворе дома Гуда, под аккомпанемент надоедливого, как комариный писк, гудения близкого шоссе, я спросил Пендлтон, говорил ли когда-нибудь знаменитый дешифровщик об интеллектуальном взрыве и о том, сможет ли когда-нибудь компьютер снова спасти мир, как это произошло во времена его молодости. Она на мгновение задумалась, пытаясь отыскать что-то в памяти, а затем ответила, к моему немалому удивлению, что Гуд изменил свое мнение об интеллектуальном взрыве. Она сказала, что ей нужно посмотреть кое-какие его бумаги, чтобы ответить на этот вопрос как следует.

В тот же вечер в тихом ресторанчике, где Гуд обычно вечером в субботу встречался со своим другом Голдом Хольцманом, тот рассказал мне, что на взгляды Гуда сильное влияние оказали три вещи: Вторая мировая война, холокост и несчастная судьба Тьюринга. Это помогло мне связать мысленно военный опыт Гуда и то, что он написал в работе «Размышления о первой ультраразумной машине». Гуду и его коллегам пришлось столкнуться в жизни со смертельной угрозой, и одержать победу в борьбе им помогли вычислительные машины. Если машина могла спасти мир в 1940-е, то, может быть, сверхразумная машина могла бы решить проблемы человечества в 1960-е. А если машина получила бы возможность *обучения*, ее интеллект буквально взорвался бы. Человечеству пришлось бы приспособиться к жизни на одной планете со сверхразумными машинами. В «Размышлениях» он писал:

> Машины вызовут социальные проблемы, но не исключено, что они смогут и решать их в дополнение к проблемам, порожденным микробами и людьми. Такие машины будут внушать страх и уважение, возможно, даже любовь. Некоторым читателям эти замечания могли бы показаться чистой фантазией, но автору они представляются очень реальными и насущными, заслуживающими особого внимания за пределами научной фантастики.

Между Блетчли-парком и интеллектуальным взрывом нет прямой концептуальной связи, их соединяет лишь извилистая линия, испытавшая множество влияний. В 1996 г. в интервью статистику и бывшему ученику Дэвиду Бэнксу Гуд рассказал, что на написание этого очерка его подвигло погружение в проблемы искусственных нейронных сетей (ИНС). Они представляют собой вычислительную модель, имитирующую деятельность настоящих нейронных сетей в мозге человека. При стимуляции нейроны

мозга посылают сигнал другим нейронам. Этот сигнал может нести зашифрованную информацию о воспоминании, может порождать действие, а может делать то и другое одновременно. Гуд в свое время познакомился с книгой психолога Дональда Хебба, в которой тот предположил, что поведение нейронов можно смоделировать математически.

Машинный «нейрон» должен быть соединен с другими машинными нейронами, причем каждому соединению придается численный «вес» в соответствии с его устойчивостью. Считается, что машинное обучение имеет место, если два нейрона активируются одновременно, и «вес» связи между ними возрастает. «Между клетками, которые срабатывают вместе, появляется связь». Это утверждение стало лозунгом теории Хебба. В 1957 г. психолог из Массачусетского технологического института (МТИ) Фрэнк Розенблатт создал на основе работ Хебба нейронную сеть и назвал ее Перцептроном. Перцептрон, реализованный на базе компьютера фирмы IBM, занимавшего целую комнату, «видел» и распознавал простые визуальные образы. В 1960 г. IBM попросила Гуда оценить Перцептрон. «Мне казалось, что нейронные сети с их параллельной работой имеют не меньше шансов привести к созданию разумной машины, чем программирование», — рассказывал Гуд. Первые доклады, на которых, собственно, и основывались «Размышления о первой ультраразумной машине», вышли два года спустя. Родилась концепция интеллектуального взрыва.

Мнении Гуда об ИНС содержало больше истины, чем он сам догадывался. Сегодня искусственные нейронные сети — тяжеловесы искусственного интеллекта, они задействованы в самых разных приложениях, от систем распознавания речи и почерка до программ финансового моделирования, от одобрения кредитов до управления роботами. У ИНС прекрасно получается высокоуровневое, быстрое распознавание образов, необходимое для всех этих работ. В большинстве приложений возможна также «тренировка» нейронных сетей на больших массивах данных (называемых обучающими выборками), на которых сеть «усваивает» закономерности. Позже она может узнавать аналогичные структуры в новых данных. Аналитики могут задать вопрос: если судить по данным последнего месяца, как будет выглядеть фондовый рынок *через неделю?* Или: какова вероятность, что некто не сможет оплачивать закладную на дом, если исходить из истории его доходов, расходов и кредитных данных за три года?

Подобно генетическим алгоритмам, ИНС представляют собой «черные ящики». То есть входные данные — веса соединений сети и нейронные

срабатывания — прозрачны. То, что получается на выходе, тоже понятно. Но что происходит внутри? Никто не понимает. Выходные данные систем искусственного интеллекта типа «черный ящик» невозможно предсказать, поэтому они не могут быть по-настоящему, доказательно «безопасными».

Но они, скорее всего, будут играть значительную роль в системах УЧИ. Сегодня многие исследователи уверены, что распознавание образов — то, на что был нацелен Перцептрон Розенблатта, — представляет собой главный инструмент интеллекта. Джефф Хокинс, изобретатель карманного компьютера Palm Pilot и коммуникатора Handspring Treo, первым реализовал на ИНС распознавание рукописного текста. Его компания Numenta создает УЧИ на основе технологии распознавания образов. Бывший главный технолог Numenta Дайлип Джордж теперь возглавляет фирму Vicarious Systems, корпоративные амбиции которой отражает девиз: «Мы создаем программное обеспечение, которое думает и учится, как человек».

Нейробиолог, когнитивист и инженер-биомедик Стивен Гроссберг предложил модель на основе ИНС, которая, по мнению некоторых специалистов, может привести к созданию УЧИ и, возможно, «ультраразума», потенциал которого Гуд видел в нейронных сетях. Если говорить в самом общем плане, сначала Гроссберг определяет роли в когнитивном процессе различных областей коры головного мозга. Именно здесь обрабатывается информация и рождается мысль. Затем он создает ИНС-модели всех задействованных областей. Он уже добился успеха в обработке движения и речи, распознавании формы и выполнении других сложных задач. Теперь он занимается разработкой логической связи своих модулей.

Возможно, машинное обучение было для Гуда новой концепцией, но при оценке Перцептрона для IBM он обязательно должен был столкнуться с алгоритмами машинного обучения. При этом манящая перспектива обучения машины подобно человеку вызвала у Гуда мысли о возможных последствиях, которые никому другому тогда еще не приходили в голову. Если машина способна сделать себя умнее, то эта новая, улучшенная машина способна будет сделать себя еще умнее, и т. д.

В бурные 1960-е, когда рождалась концепция интеллектуального взрыва, Гуд, возможно, думал о тех проблемах, с решением которых разумная машина могла бы помочь. Тогда уже не нужно было топить вражеские подлодки, но был на Земле враждебный Советский Союз, были Карибский кризис, убийство президента Кеннеди и опосредованная война между США и Китаем в Юго-Восточной Азии. Человечество катилось к

гибели; казалось, настало время нового Колосса. В «Размышлениях» Гуд писал:

> [Пионер вычислительной техники] Б.В. Боуден утверждал… что нет смысла строить машину с разумом человека, поскольку проще создать человеческий мозг обычным методом… Отсюда видно, что даже высокоинтеллектуальные люди могут проглядеть "интеллектуальный взрыв". Да, правда, строить машину, способную лишь на обычные интеллектуальные приложения, неэкономично, но представляется достаточно вероятным, что если это в принципе возможно, то затем при двойных затратах машина сможет продемонстрировать ультраразум.

Так что, затратив еще некоторое количество долларов, можно получить ИСИ — искусственный суперинтеллект, говорит Гуд. Но затем берегись возможных последствий на уровне цивилизации — ведь планету придется делить с другим разумом, более мощным, чем человеческий.

В 1962 г., перед написанием «Размышлений о первой ультраразумной машине», Гуд редактировал книгу под названием «Ученый размышляет». Он написал для нее главу «Социальные последствия искусственного интеллекта» — своего рода разминку перед формированием идей о суперинтеллекте, над которыми он в то время размышлял. Он писал — а Стив Омохундро повторил за ним почти пятьдесят лет спустя, — что среди проблем, которыми придется заниматься разумным машинам, обязательно будут и проблемы, вызванные их собственным появлением на Земле, нарушающим сложившийся порядок.

> Такие машины… могли бы даже выдвигать полезные политические и экономические предложения; и они *должны* будут делать это, чтобы компенсировать проблемы, вызванные их собственным существованием. Это будут проблемы перенаселенности, возникшие из-за победы над болезнями, и безработицы, причиной которой станет эффективность роботов низкого уровня, которых сконструируют главные машины.

Однако, как мне предстояло узнать, позже Гуд пережил удивительную метаморфозу и совершенно изменил свои взгляды. Я всегда причислял его к оптимистам, таким как Рэй Курцвейл, поскольку он видел в молодости, как машины «спасали» мир, и в своем очерке ставил выживание человека в

зависимость от создания сверхразумной машины. Но подруга Гуда Лесли Пендлтон намекнула на изменение позиции. Ей потребовалось время, чтобы припомнить обстоятельства и контекст, но в последний день моего пребывания в Блэксбурге она рассказала мне все, что помнила.

В 1998 г., когда ему было восемьдесят два, Гуд был удостоен медали «Пионер компьютерной техники» сообщества IEEE (Института инженеров электротехники и электроники). В речи, произнесенной по этому поводу, его попросили рассказать свою биографию. Он составил биографию, но не стал ее зачитывать, и на церемонии она не прозвучала. Вероятно, только Пендлтон знала о ее существовании. Она сделала копию этой биографии и отдала мне перед отъездом из Блэксбурга вместе с кое-какими другими бумагами, которые я у нее просил.

Прежде чем вновь выехать на автостраду I-81 и направиться обратно на север, я прочитал автобиографию Гуда в машине на стоянке центра облачных вычислений Rackspace. Подобно Amazon и Google, Rackspace предлагает серьезные компьютерные мощности за небольшие деньги, сдавая в аренду время десятков тысяч своих процессоров и место под экзабайты информации. Конечно, Вирджинскому политехническому институту удобно иметь у себя под боком такую организацию, как Rackspace, и мне хотелось посетить их вычислительный центр, но он был закрыт. А позже мне пришло в голову: как странно, что в десятках метров от того места, где я читал автобиографические заметки Гуда, десятки тысяч процессоров с воздушным охлаждением работали над решением мировых проблем.

В своей биографии, игриво написанной от третьего лица, Гуд вспомнил все вехи своей жизни, включая и никому, вероятно, прежде не известные воспоминания о работе в Блетчли-парке с Тьюрингом. Но вот что он написал в 1998 г. о первом суперинтеллекте и о том, как изменились его взгляды:

> [Статья] "Размышления о первой ультраразумной машине" (1965 г.)… начиналась так: "Выживание человека зависит от скорейшего создания ультраразумной машины". Таковы были его [Гуда] слова во время холодной войны, но сейчас он подозревает, что "выживание" следовало бы заменить на "вымирание". Он считает, что из-за международной конкуренции мы не сможем предотвратить переход власти к машинам. Он считает нас леммингами. Он сказал также, что "вероятно, человек создаст *deux ex machina*[16] *по* своему образу и подобию".

Я читал это и смотрел невидящими глазами на здание Rackspace. В конце жизни Гуд пересмотрел не только свое мнение о вероятности существования Бога. Я как будто нашел послание в бутылке, примечание, изменившее смысл текста на прямо противоположный. Теперь у нас с Гудом появилось кое-что общее. Мы оба считали, что интеллектуальный взрыв добром не кончится.

Глава 8
Точка невозврата

> *Но если технологическая сингулярность может наступить, она наступит. Даже если бы все правительства мира осознали «угрозу» и смертельно ее испугались, прогресс в этом направлении продолжался бы. Более того, конкурентное преимущество — экономическое, военное, даже художественное — каждого нового достижения автоматизации настолько наглядно, что принятие законов или установление традиций, запрещающих подобные вещи, попросту гарантирует, что это сделает кто-то другой.*
>
> *Вернор Виндж. Приближающаяся технологическая сингулярность (1993)*

Не правда ли, приведенная цитата[17] очень напоминает отрывок из автобиографии Гуда? Как и Гуд, дважды лауреат премии Хьюго писатель-фантаст и профессор математики Вернор Виндж говорит о том, что человеку свойственно искать неприятности на свою голову. Виндж сказал мне, что не читал собственноручно написанной Гудом биографии и не знал о том, что в конце жизни тот изменил свое отношение к интеллектуальному взрыву. Знали об этом, вероятно, только сам Гуд и Лесли Пендлтон.

Вернор Виндж был первым человеком, кто формально употребил слово «сингулярность» при описании технологического будущего; сделал он это в 1993 г. в обращении к NASA, озаглавленном «Приближающаяся технологическая сингулярность». Математик Станислав Улам сообщил, что они с энциклопедистом Джоном фон Ньюманом использовали термин «сингулярность» в разговоре о технологических переменах еще за тридцать пять лет до этого, в 1958 г. Но Виндж сделал это публично и демонстративно — и мяч сингулярности покатился прямо в руки Рэя Курцвейла и того, что сегодня стало движением сингулярности.

Но почему же Виндж, обладая таким имиджем, не ездит по стране с лекциями и конференциями как главный эксперт по сингулярности?

Ну, у слова «сингулярность» несколько значений, и Виндж использует его в более узком смысле, чем многие другие. Определяя сингулярность, он провел аналогию с точкой на орбите черной дыры, дальше которой свет с ее поверхности пройти не может. Невозможно увидеть, что происходит за этой точкой, известной как горизонт событий. Точно так же, если нам придется делить планету с сущностями разумнее нас, пути назад не будет — мы не в состоянии предсказать, что произойдет в подобном случае. Чтобы судить об этом, нужно обладать, по крайней мере, равным интеллектом.

Таким образом, если вы не можете знать, как будет развиваться будущее, то как вы можете писать об этом? Виндж не сочиняет околонаучных фантазий — он пишет то, что называют твердой научной фантастикой, используя в своих произведениях данные реальной науки. Сингулярность лишает его надежной опоры.

> Это проблема, с которой мы сталкиваемся всякий раз, когда говорим о создании разума более высокого, чем наш собственный. Когда это произойдет, человеческая история достигнет своего рода точки сингулярности — места, где экстраполяции теряют смысл и где не обойтись без введения новых моделей, — и мир выйдет за пределы нашего понимания.

Виндж рассказывал, что когда в 1960-е он начинал писать научную фантастику, мир, основанный на будущих достижениях науки, о которых он писал, отстоял от него на сорок или пятьдесят лет. Однако к 1990-м гг. будущее уже стремительно *приближалось,* и скорость технических перемен, казалось, все росла. Виндж уже не мог предугадать, что принесет человечеству будущее, поскольку считал, что очень скоро в мире появится разум, превосходящий человеческий. Именно этот, а не наш разум будет определять скорость технического прогресса. Виндж, как и другие фантасты, не мог писать об этом.

> На протяжении 1960-х, 70-х и 80-х годов понимание надвигающегося катаклизма ширилось. Возможно, первыми его приближение почувствовали писатели-фантасты. В конце концов, именно "твердые" фантасты стараются описывать конкретные последствия всего, что может сделать с нами технический прогресс. Именно они все больше и больше чувствовали эту непрозрачную стену, отгородившую нас от будущего.

Исследователь ИИ Бен Гертцель сказал мне:

> Вернор Виндж очень ясно видел принципиальную непознаваемость ИИ, когда предлагал концепцию технологической сингулярности. Именно поэтому он не ездит по миру с лекциями о ней; он просто не знает, что сказать. Что он может сказать? "Да, я считаю, что мы вот-вот создадим технологии, которые будут намного умнее человека, и после этого кто знает, что произойдет?"

Но как же в этом случае приручение огня, появление земледелия, печатный станок и электричество? Разве человечество не пережило до сих пор множество технологических сингулярностей? В разрушительных технических новинках нет ничего нового, но никто не придумывал для каждой из них красивого имени и не пугал ими окружающих. Моя бабушка родилась еще до широкого распространения автомобилей, но дожила до первых шагов Нила Армстронга по Луне. Она считала, что все это называется XX век. Что же делает переход, о котором говорит Виндж, таким особым?

«Секретный компонент — разум, — сказал мне Виндж. В его стремительном тенорке то и дело звучал смех. — Разум делает его другим, а определяющая рабочая черта этого перехода в том и заключается, что человек не может ничего знать заранее. Мы теперь находимся в такой ситуации, что очень-очень скоро, всего через несколько десятилетий, начнутся трансформации, которые будут иметь, как в древности, большое биологическое значение».

В эту концепцию укладываются две важные идеи. Во-первых, технологическая сингулярность изменит сам разум — единственную суперсилу человека, которая, собственно, и создает технологии. Вот почему эта революция будет отличаться от любой другой. Во-вторых, биологическая трансформация, на которую намекает Виндж, произошла двести тысяч лет назад, когда человечество захватило лидерство на мировой арене. Homo sapiens начал доминировать на планете, поскольку был умнее любого другого биологического вида. Точно так же разум, в тысячу или миллион раз превосходящий человеческий в интеллектуальном отношении, навсегда изменит правила игры. Что же тогда произойдет с нами?

Этот вопрос вызвал у Винджа взрыв смеха.

> Если на меня уж очень наседают с вопросами о том, на что будет похожа сингулярность, то чаще всего я говорю: а почему, как вы думаете, я назвал это сингулярностью[18]?

Но по поводу непрозрачного будущего Виндж сделал, по крайней мере, один вывод — сингулярность несет угрозу и может привести к гибели человечества. Писатель, процитировав в речи 1993 г. целиком абзац из статьи Гуда 1967 г. об интеллектуальном взрыве, указал, что знаменитый статистик не так далеко зашел в своих выводах:

> Гуд ухватил суть процесса (неуправляемого роста мощности ИИ), но не стал рассматривать его самые тревожные следствия. Любая разумная машина того типа, о котором он пишет, станет "инструментом" человечества не больше, чем само человечество является инструментом кроликов, или зябликов, или шимпанзе.

Вот еще одна уместная аналогия — кролики для людей то же, что люди для сверхразумных машин. А как мы относимся к кроликам? Как к вредителям или как к потенциальному обеду, в лучшем случае — как к домашним любимцам. ИСИ-агенты поначалу будут нашими инструментами — как нашими инструментами являются сегодня их предки Google, Siri и Watson. Кроме того, считает Виндж, существуют и другие факторы помимо индивидуального машинного интеллекта, способные вызвать сингулярность. Среди них — разум, берущий начало от Интернета (от Интернета вместе с пользователями — этакая цифровая Гея), от человеко-машинных интерфейсов или биологических наук (через повышение интеллекта будущих поколений при помощи манипуляций с генами).

В трех из перечисленных путей человек вовлечен в развитие технологий; возможно, он способен провести постепенное и управляемое повышение интеллекта и таким образом избежать взрыва. Так что можно, говорит Виндж, подумать о том, как победить величайшие проблемы человечества — голод, болезни, даже саму смерть. Это оптимистичное представление о будущем Курцвейла, которое пропагандируют «сингуляритарии». Сингуляритарии — это те, кто ожидает от резкого ускорения будущего в основном хороших вещей. Винджу их «сингулярность» представляется слишком благодушной.

> Ставки в нашей игре чрезвычайно высоки, и положительная

> сторона ее так оптимистична, что это само по себе производит пугающее впечатление. Ветры мировой экономики — необычайно могущественная сила — связаны с достижениями ИИ. Сотни тысяч людей в мире, очень умных людей, работают над вещами, которые должны привести к созданию сверхчеловеческого разума. Вероятно, большинство из них не рассматривает свою работу в таком ключе. Они просто хотят сделать *быстрее, дешевле, лучше, выгоднее.*

Виндж сравнивает все это со стратегией времен холодной войны, известной как MAD (mutual assured destruction — взаимно гарантированное уничтожение). Термин этот был пущен в оборот любителем аббревиатур Джоном фон Ньюманом (он также был создателем одного из первых компьютеров с говорящим именем MANIAC), а сама стратегия помогала сохранить мир благодаря угрозе взаимного уничтожения сторон. Подобно MAD, искусственный интеллект может похвастать тем, что множество исследователей тайно трудятся над развитием технологий, потенциально грозящих катастрофой. Но это — взаимно гарантированное уничтожение без всяких тормозов, обусловленных здравым смыслом. Никто не может знать, кто на самом деле идет впереди в этой гонке, и каждый считает, что впереди кто-то другой. Но, как мы видели, победитель не сможет забрать все. Победитель в гонке за создание ИИ получит лишь сомнительную привилегию первым столкнуться с суперинтеллектом «лицом к лицу» или попробовать выжить в ситуации, очень напоминающей сценарий Busy Child.

«Мы имеем сегодня тысячи хороших людей, которые работают по всему миру, пытаясь совместными усилиями приблизить катастрофу, — сказал Виндж. — А двигаться вперед по ландшафту угроз очень сложно. Мы тратим недостаточно сил на раздумья о том, что будет, если нас постигнет неудача».

Другие сценарии, которые тоже тревожат Винджа, также заслуживают внимания. Цифровая Гея, или брак между людьми и компьютерами, уже формируется в Интернете. Что это будет означать для нашего будущего — глубокий и важный вопрос, заслуживающий, чтобы о нем написали гораздо больше книг, чем пишется сейчас. УИ, или искусственное усиление интеллекта, имеет равный с обычным ИИ катастрофический потенциал, немного смягченный (по крайней мере поначалу) *участием* человека. Но это преимущество быстро исчезнет. Позже мы поговорим об УИ подробнее. А сначала я хочу обратить ваше внимание на замечание Винджа

о том, что разум может вырасти из Интернета.

Техномыслители, в том числе Джордж Дайсон и Кевин Келли, выдвинули предположение о том, что информация — это форма жизни. Компьютерная программа, несущая информацию, воспроизводит себя и растет по биологическим законам. Но разум… нет, разум — это совсем не то. Разум присущ лишь сложным организмам и не возникает случайно.

Будучи в гостях у Елиазера Юдковски в Калифорнии, я задал вопрос, может ли интеллект, хотя бы в принципе, родиться в результате экспоненциального роста инфраструктуры Интернета из содержащихся в нем 5 трлн мегабайт данных, из более чем 7 млрд связанных между собой компьютеров и смартфонов и 75 млн серверов. Юдковски поморщился, как будто на клетки его мозга внезапно попал кислый раствор.

«Категорически нет, — сказал он. — Потребовались миллиарды лет эволюции, чтобы появился разум. Интеллект не может народиться просто из сложности жизни. Автоматически такое не происходит. Оптимизация происходит через естественный отбор под внешним давлением».

Иными словами, интеллект не возникает просто из сложности как таковой. Кроме того, Интернету недостает давления среды, которое в природе сохраняет одни мутации и отбраковывает другие.

«Я люблю говорить, что весь Млечный Путь за пределами Земли, вероятно, устроен не так интересно, как одна земная бабочка, потому что бабочка прошла эволюцию, в процессе которой запоминались удачные решения, и развитие шло, отталкиваясь от них», — сказал Юдковски.

Я согласен с тем, что интеллект не в состоянии расцвести, спонтанно родившись из Интернета. Но я считаю, что агентное финансовое моделирование, возможно, очень скоро изменит саму Сеть.

Когда-то давным-давно аналитики Уолл-стрит, желая предсказать поведение рынка, обращались к серии правил, прописанных специалистами по макроэкономике. Эти правила учитывают такие факторы, как процентные ставки по кредитам, данные по безработице и строительству новых жилых домов. Все больше, однако, Уолл-стрит переходит на агентное финансовое моделирование. Эта новая наука способна моделировать весь фондовый рынок, и даже всю экономику, ради того чтобы повысить качество прогнозирования.

Для моделирования рынка исследователи строят компьютерные модели субъектов, занятых покупкой и продажей акций, — отдельных людей, фирм, банков, хедж-фондов и т. п. У каждого из тысяч таких «агентов» свои цели и правила принятия решений или стратегии покупки и продажи. На них, в свою очередь, влияют непрерывно меняющиеся

рыночные данные. Эти агенты, реализованные на искусственных нейронных сетях и других методиках построения ИИ, «настраиваются», опираясь на реальную информацию. Действуя синхронно и «питаясь» текущими данными, агенты создают изменчивый портрет живого рынка.

Затем аналитики начинают испытывать сценарии торговли отдельными активами, и при помощи методов эволюционного программирования модель рынка может «шагнуть вперед» на день или на неделю, помогая представить, как будет выглядеть рынок в ближайшем будущем и какие он может обещать инвестиционные возможности. Такой подход «снизу вверх» к созданию финансовых моделей воплощает в себе идею о том, что простые правила поведения отдельных агентов порождают сложную картину общего поведения. Можно даже обобщить: что верно в отношении Уолл-стрит, верно и в отношении пчелиного улья или муравейника.

В суперкомпьютерах финансовых столиц мира начинают формироваться виртуальные миры, напитанные подробностями реального мира и населенные все более разумными «агентами». Чем точнее и подробнее прогноз, тем выше прибыли. Так что на повышение точности моделей на всех уровнях работают мощные экономические стимулы.

Но если полезно создавать цифровых агентов, реализующих сложные стратегии покупки биржевых активов, то разве не *выгоднее* создавать цифровые модели с полным спектром человеческих мотиваций и способностей? Почему бы не создать УЧИ или виртуальных агентов человеческого уровня разумности? Собственно, этим и занимается Уолл-стрит, но под другим названием — «агентные финансовые модели».

Рано или поздно финансовые рынки породят УЧИ, считает доктор Александер Уисснер-Гросс. Резюме Уисснер-Гросса таково, что остальным изобретателям, ученым и эрудитам остается только кусать локти. Он автор тринадцати книг, держатель шестнадцати патентов; его зачислили в МТИ с высшими баллами сразу на три направления (физика, инженерное дело и математика), а окончил он Инженерную школу МТИ первым на курсе. Он защитил степень доктора философии в Гарварде и получил немаленькую премию за свою диссертацию. Он основал и продал несколько компаний и, согласно его собственному резюме, получил «107 серьезных наград и знаков отличия» (вероятно, речь идет не о наградах типа «менеджер недели»). В настоящее время он работает в Гарварде, занимается исследованиями и пытается коммерциализировать свои идеи в области финансовой инженерии.

Так вот, он считает, что, пока блестящие теоретики всего мира

соревнуются между собой, стремясь первыми создать УЧИ, он может появиться уже готовым и полностью сформированным на финансовых рынках как непредвиденное следствие создания вычислительных моделей большого числа людей. Кто будет его создателем? «Кванты» — так на Уолл-стрит называют компьютерщиков — специалистов по финансовой математике.

«Безусловно, возможно, что реальный живой искусственный интеллект человеческого уровня мог бы появиться на финансовых рынках, — сказал мне Уисснер-Гросс. — Не как результат одного-единственного алгоритма одного кванта, но как совокупность всех алгоритмов множества хедж-фондов. Для УЧИ, может быть, и не потребуется последовательная теория. Возможно, это будет совокупный феномен. Финансы имеют хороший шанс стать тем самым первичным бульоном, в котором зародится УЧИ».

Чтобы купиться на этот сценарий, необходимо верить, что создание все более совершенных финансовых моделей подпитывается большими деньгами. Это действительно так — как ни смешно, в эту сферу финансисты вкладывают больше денег, чем тратит кто бы то ни было на машинный интеллект, возможно, даже больше, чем DARPA, IBM и Google могут бросить на разработку УЧИ. Деньги превращаются в новые, более мощные суперкомпьютеры и лучших программистов. Уисснер-Гросс говорит, что «кванты» пользуются теми же инструментами, что и разработчики ИИ, — это нейронные сети, генетические алгоритмы, автоматическое чтение, скрытые марковские модели и все, что может прийти в голову. Каждый новый инструмент ИИ проходит испытание в горниле финансов.

«Стоит появиться новой методике ИИ, — рассказывал Уисснер-Гросс, — как сразу же задается вопрос: можно ли использовать это на фондовом рынке?»

А теперь представьте, что вы высокопоставленный «квант» и имеете в своем распоряжении достаточно средств, чтобы нанимать других специалистов и покупать оборудование. Хедж-фонд, на который вы работаете, имеет громадную модель Уоллстрит, населенную тысячами разнообразных экономических агентов. Ее алгоритмы взаимодействуют с алгоритмами других хедж-фондов — они так тесно связаны, что поднимаются и падают вместе и действуют как будто по предварительной договоренности. По словам Уисснер-Гросса, рыночные наблюдатели уже высказывали предположение о том, что некоторые алгоритмы, кажется, *сигнализируют* один другому и распространяют информацию по всей Уолл-

стрит при помощи миллисекундных сделок, реализуемых с такой скоростью, какую не в состоянии отследить ни один человек (это тот самый высокочастотный трейдинг, о котором шла речь в главе 6).

Не правда ли, следующим логическим шагом было бы заставить ваш хедж-фонд думать? То есть ваш алгоритм, возможно, не должен автоматически запускать продажи в ответ на массированный сброс акций другим фондом (что, собственно, послужило причиной мгновенного краха в мае 2010 г.). Вместо этого он должен своевременно заметить сброс акций и, прежде чем делать свой ход, посмотреть, как на него будут реагировать другие фонды и рынок в целом. Тогда он, возможно, сумеет сделать иной, более удачный ход. Или, может быть, что еще лучше, сможет одновременно с этим промоделировать большое число гипотетических рынков и будет готов к реализации одной из множества стратегий в ответ на конкретные условия.

Иными словами, существуют громадные финансовые стимулы наделить ваши алгоритмы самосознанием — чтобы они точно знали, чем являются, и могли моделировать окружающий мир. А это уже *очень* похоже на УЧИ. Именно в этом направлении, безусловно, развивается рынок, но есть ли результаты и близок ли кто-нибудь к созданию УЧИ?

Уисснер-Гросс не знал ответа на этот вопрос. Возможно, он не открыл бы его, даже если бы знал. «Есть серьезные материальные причины держать в тайне любые успехи, обещающие хорошую выгоду», — сказал он.

Разумеется. И он говорит при этом не просто о конкуренции между хедж-фондами, но и о своего рода естественном отборе среди алгоритмов. Победители процветают и передают тексты своих программ потомкам. Неудачники погибают. Эволюционное давление рынка способно ускорить развитие интеллекта, но не без руководящего участия человека-«кванта». Пока.

А интеллектуальный взрыв в мире финансовой инженерии был бы абсолютно непрозрачным, по крайней мере по четырем причинам. Во-первых, подобно многим когнитивным архитектурам, он, вероятно, использовал бы нейронные сети, генетическое программирование и прочие «черные ящики» методик ИИ. Во-вторых, высокоскоростные транзакции, занимающие от силы несколько миллисекунд, происходят быстрее, чем может отследить человек, — достаточно посмотреть на то, как происходил обвал 2010 г. В-третьих, система невероятно сложна — ни один квант (ни даже группа квантов) не в состоянии полностью описать экосистему алгоритмов Уолл-стрит и объяснить, как алгоритмы взаимодействуют

между собой.

Наконец, если мощный интеллект родится в недрах финансовой математики, его, скорее всего, будут держать в тайне до тех пор, пока он будет приносить деньги своим создателям. Вот вам четыре уровня непроницаемости.

Подведем итоги. В принципе УЧИ может родиться на Уолл-стрит. Самые успешные алгоритмы и сейчас держат в тайне — и «кванты», с любовью их реализующие, и компании, ими владеющие. Интеллектуальный взрыв здесь пройдет незаметно для большинства людей, если не для всех, и остановить его, вероятно, в любом случае будет невозможно.

Надо сказать, что сходства между финансовой математикой и исследованиями ИИ на этом не заканчиваются. Уисснер-Гросс выдвинул еще одно поразительное предположение. Он утверждает, что первые стратегии управления УЧИ могли бы вырасти из мер, предлагаемых нынче для управления высокочастотным трейдингом и контроля над ним. Некоторые из них выглядят достаточно перспективно.

Рыночный предохранитель, или прерыватель цепи, должен отрезать ИИ хедж-фонда от внешнего мира в случае непредвиденных обстоятельств. Грубо говоря, зарегистрировав каскадные взаимодействия алгоритмов вроде тех, что происходили во время краха 2010 г., такой автомат должен «выдернуть вилку из розетки», то есть физически отключить машины от электропитания.

Правило крупного трейдера требует подробной регистрации всех ИИ наряду с человеческими организациями. Если это кажется вам прелюдией к серьезному вмешательству правительства, то вы правы. Почему нет? Уолл-стрит снова и снова доказывает свою неспособность к ответственному поведению без энергичного регулирования. Относится ли это также к разработчикам УЧИ? Несомненно. К изучению УЧИ допускаются не только люди высоких моральных качеств.

Предварительное тестирование алгоритмов могло бы моделировать поведение алгоритмов в виртуальной среде, прежде чем выпускать их на реальный рынок. Цель *аудита исходного текста программ ИИ* и *централизованной записи деятельности ИИ*—предугадывание возможных ошибок и поощрение анализа постфактум любых неожиданных событий, таких как крах 2010 г.

Но оглянитесь на упомянутые выше четыре уровня непрозрачности и подумайте, достаточно ли будет таких мер предосторожности, даже если они будут полностью реализованы, и смогут ли они, на ваш взгляд,

гарантировать безопасность.

Как мы уже видели, Виндж принял эстафету от Гуда и дополнил описание интеллектуального взрыва новыми важными чертами. Он рассмотрел альтернативные варианты его достижения помимо нейронных сетей, о которых говорил Гуд, и указал на возможность и даже вероятность гибели человечества. И возможно, самое важное, Виндж дал этому событию имя — сингулярность.

Винджу, как автору оригинального научно-фантастического рассказа «Истинные имена», отлично известно, что присвоение имени — чрезвычайно важный акт. Имена просятся на язык, откладываются в мозге и без труда преодолевают пропасть между поколениями. В Книге Бытия, как говорят теологи, раздача имен всему на Земле на седьмой день важна потому, что на созданной Богом «сцене» рядом с ним должно было появиться разумное существо, которому и предстояло в дальнейшем пользоваться этими именами. Языковая практика — важная веха в развитии ребенка. Без языка мозг не может нормально развиваться. Вряд ли УЧИ возможен без языка, без существительных, без имен.

Виндж дал сингулярности название, чтобы обозначить страшное будущее для человечества. Его определение сингулярности метафорично — орбита вокруг черной дыры, где гравитационные силы настолько сильны, что даже свет не в состоянии вырваться вовне. Мы не можем узнать, что там внутри, и такое название дано не случайно.

Затем внезапно все изменилось.

К концепции сингулярности, предложенной Винджем, Рэй Курцвейл добавил мощный катализатор, сдвигающий всю тему в область стремительного развития технологий и фокусирующий внимание на надвигающейся катастрофической опасности: экспоненциальном росте мощности и скорости компьютеров. Именно из-за этого роста вы можете окидывать скептическим взглядом любого, кто скажет, что машинный интеллект человеческого уровня не будет получен еще лет сто, а может, и вообще никогда не будет достигнут.

В расчете на один доллар мощность компьютеров за последние тридцать лет выросла в миллиард раз. Еще лет через двадцать за $1000 можно будет купить в миллион раз более мощный компьютер, чем сегодня, а через двадцать пять лет — в миллиард. Примерно к 2020 г. появятся первые компьютерные модели человеческого мозга, а к 2029 г. исследователи смогут запустить модель мозга, которая не будет ни на йоту уступать человеческому ни в интеллектуальном, ни в эмоциональном плане. К 2045 г. человеческий и машинный интеллект вырастут *в миллиард*

раз и разработают технологии, которые позволят победить наши человеческие слабости, такие как утомление, болезни и смерть. В случае если нам удастся это пережить, технический прогресс за XXI в. продвинется не на столетие, а на 200 000 лет.

Приведенный анализ и экстраполяция принадлежат Курцвейлу и могут служить ключом к третьему, победившему на данный момент определению сингулярности. Сердцем этого определения является Закон прогрессирующей отдачи — теория технического прогресса, которую Курцвейл не придумал, но извлек на всеобщее обозрение, примерно так же, как Гуд предугадал интеллектуальный взрыв, а Виндж предупредил о надвигающейся сингулярности. Закон прогрессирующей отдачи говорит о том, что экстраполяции и успехи, о которых мы говорим в этой книге, накатываются на нас, как грузовой состав, который на каждом километре удваивает скорость. Очень трудно представить себе, насколько быстро этот состав окажется здесь, но достаточно сказать, что если к концу первого километра его скорость будет составлять 20 км/ч, то всего через пятнадцать километров он будет мчаться со скоростью более 65 000 км/ч. Важно отметить, что экстраполяция Курцвейла относится не только к «железу» (скажем, к начинке нового айфона), но и к развитию теорий (к примеру, к созданию единой теории искусственного интеллекта).

Но здесь мое мнение расходится с мнением Курцвейла. Я считаю, что вместо дороги в рай, куда мы движемся, по утверждению Курцвейла, Закон прогрессирующей отдачи описывает кратчайшее возможное расстояние между нашей жизнью — такой, какая она сегодня, — и концом эры человечества.

Глава 9
Закон прогрессирующей отдачи

В настоящее время информатика проходит самую замечательную трансформацию с момента изобретения персонального компьютера. Новшества следующего десятилетия превзойдут все новшества трех предыдущих десятилетий вместе взятых.

Пол Отеллини, исполнительный директор Intel

В книгах «Эпоха духовных машин: Когда машины превосходят человеческий интеллект» (The Age of Spiritual Machines: When Computers Exceed Human Intelligence) и «Сингулярность рядом» (The Singularity Is Near) Рэй Курцвейл воспользовался словом «сингулярность», кардинально изменив его значение и обозначая им яркий, полный надежд период человеческой истории, который инструменты экстраполяции позволили разглядеть с замечательной точностью. Где-то в ближайшие 40 лет, пишет он, техническое развитие пойдет с такой скоростью, что человеческое существование фундаментально изменится, а ткань истории порвется. Машины и биологические объекты станут неразличимы. Виртуальные миры будут более живыми и захватывающими, чем реальность. Нанотехнологии позволят производить что угодно на заказ, будут побеждены голод и бедность и найдены лекарства от всех человеческих болезней. Вы сможете предотвратить и даже обратить вспять старение собственного тела. Это будет самая значительная эпоха, жить в ней станет прекрасно не только потому, что мы будем свидетелями поистине поразительных скоростей технологического превращения, но и потому, что техника обещает подарить нам способы жить вечно. Это заря уникальной, или «сингулярной», эпохи.

Что же такое в конечном итоге сингулярность? Это период в будущем, в ходе которого технологические перемены станут настолько стремительными, а их влияние настолько глубоким, что человеческая жизнь необратимо изменится. Эта эпоха, не будучи ни утопичной, ни антиутопичной, изменит концепции, придающие сегодня нашей жизни смысл, от бизнес-моделей до цикла человеческой жизни, включая саму

смерть…

Взгляните на истории о Гарри Поттере с этой точки зрения. Возможно, все эти рассказы — лишь игра воображения, но они вполне могут представлять картину нашего мира, каким он станет всего через несколько десятилетий. Значительная часть поттеровской магии будет реализована при помощи технологий, о которых я расскажу в этой книге. Игра в квиддич[19] и превращение людей и предметов во что-то иное будут вполне возможны не только при полном погружении в виртуальные реальности, но и в реальной реальности при использовании специальных наноустройств.

Так что сингулярность «не будет ни утопичной, ни антиутопичной», зато мы сможем играть в квиддич! Очевидно, сингулярность Курцвейла решительно отличается от сингулярности Вернора Винджа и интеллектуального взрыва Гуда. Можно ли примирить их между собой? Неужели это будет одновременно и лучшее, и худшее время для жизни? Я прочел почти каждое слово, опубликованное Курцвейлом, и прослушал все доступные аудио- и видеозаписи и подкасты. В 1999 г. я взял у него длинное интервью для документального фильма, посвященного, в частности, искусственному интеллекту. Я знаю все, что он написал и сказал об опасностях ИИ, и это очень немного.

Однако, как ни удивительно, именно Курцвейлу мы отчасти обязаны существованием самого аргументированного очерка по данному вопросу — статьи Билла Джоя «Почему мы не нужны будущему?». В ней Джой — программист, специалист по архитектуре компьютеров и один из основателей компании Sun Microsystems — призывает к замедлению или даже остановке исследований по трем направлениям, которые, как он считает, смертельно опасно развивать нынешними темпами: это искусственный интеллект, нанотехнологии и биотехнологии. Джой написал об этом после пугающего разговора в баре с Курцвейлом и прочтения его книги «Эпоха духовных машин» (The Age of Spiritual Machines). В ненаучной литературе и популярных лекциях об опасностях ИИ, мне кажется, только три закона Азимова цитируются чаще, хотя и не всегда к месту, чем это эссе Джоя, оказавшее заметное влияние. Вот короткий отрывок — суть его позиции по ИИ:

> Но сейчас, когда появление вычислительных мощностей человеческого уровня ожидается примерно лет через 30, напрашивается новая идея: может быть, я работаю над созданием инструментов, при помощи которых станет возможна постройка технологий, которые, в конце концов, заменят наш биологический

> вид. Как я чувствую себя в этой связи? Очень неуютно. Всю свою профессиональную жизнь я бился над созданием надежных программных систем, и мне представляется более чем вероятным, что это будущее получится не таким замечательным, каким его, может быть, вообразили некоторые. Мой личный опыт говорит о том, что мы склонны переоценивать свои конструкторские способности. Учитывая невероятную мощь этих новых технологий, разве нам не следует задаться вопросом о том, как лучше всего сосуществовать с ними? А если наше собственное вымирание является вероятным или хотя бы возможным результатом нашего технического развития, разве нам не следует действовать в этом направлении с величайшей осторожностью?

Разговор с Курцвейлом в баре может дать начало международному обсуждению, но, говоря откровенно, несколько предостерегающих слов теряются у него в море интереснейших предсказаний. Он настаивает, что его описание будущего — не утопия, но я с ним не согласен.

Мало кому удается писать о технологиях более убедительно и с большим знанием дела, чем это делает Курцвейл, — он всегда заботится о том, чтобы его слова были ясны и понятны, и тщательно аргументирует свои мысли. Однако мне кажется, что он совершил ошибку, присвоив слово «сингулярность» и придав ему новое радужное значение. Настолько радужное, что мне, как и Винджу, становится страшно от этого определения, полного убедительных образов и идей, маскирующих опасность. Я уверен, что проделанный им ребрендинг преуменьшает опасность ИИ и раздувает наивные перспективы его использования. Начав с обсуждения технологий, Курцвейл создал культурное движение с сильным привкусом религии. Мне кажется, что смешивать технический прогресс и религию — большая ошибка.

> Представьте себе мир, где разница между человеком и машиной размывается, где граница между человечеством и техникой пропадает, где душа и кремниевый чип едины… Во вдохновенных руках [Курцвейла] жизнь в новом тысячелетии уже не кажется пугающей. Напротив, XXI век, по Курцвейлу, обещает стать эпохой, когда брак между человеческой сущностью и искусственным интеллектом фундаментально изменит и улучшит нашу жизнь.

Курцвейл — не просто крестный отец вопросов сингулярности, вежливый и упорный спорщик и неутомимый, хотя и несколько механистичный, пропагандист. Вокруг него собралось достаточно молодых людей, живущих на грани сингулярности. Как правило, сингуляритариям немного за тридцать, в основном они мужчины и они бездетны. По большей части это белые интеллектуалы, услышавшие зов сингулярности. Многие из них в ответ на этот зов бросили то, чем занимались, и отказались от карьеры, которую могли бы одобрить их родители, в пользу почти монашеской жизни, посвященной вопросам сингулярности, — и горды этим. Среди них много самоучек, отчасти благодаря тому, что ни одна университетская программа не предлагает специализации одновременно в компьютерных науках, этике, биоинжиниринге, нейробиологии, психологии и философии, — короче говоря, в науках, имеющих отношение к сингулярности. Курцвейл — один из основателей Университета сингулярности, который не аккредитован и не присваивает степеней. Однако обещает «широкое кросс-дисциплинарное представление о крупнейших идеях и вопросах трансформационных технологий». К тому же многие сингуляритарии слишком умны и самостоятельны, чтобы вписаться в традиционную систему образования. Да и мозги у них устроены так, что мало какой колледж или университет захотел бы иметь их в своем кампусе.

Некоторые сингуляритарии избрали главным догматом своей веры рациональность. Они считают, что повышение интеллекта и логических способностей людей, в первую очередь тех, кто завтра будет принимать решения, понижает вероятность того, что все мы совершим самоубийство посредством ИИ. Наш мозг, утверждают они, полон странных предрассудков и эвристических механизмов, которые очень пригодились нам в ходе эволюции, но которые при столкновении с комплексными рисками и выборами современного мира навлекают на нас одни лишь беды. При этом главный объект их внимания — не негативная катастрофическая, а благотворная позитивная сингулярность. В ней мы сможем воспользоваться технологиями продления жизни, которые позволят нам жить и жить без конца, вероятно, уже в механической, а не в биологической форме. Иными словами, очисти себя от неверных мыслей — и найдешь спасение от мира плоти, и откроешь жизнь вечную.

Неудивительно, что движение сингулярности часто называют «Блаженство чокнутых» (Rapture of the Geeks) — как движение оно имеет все отличительные особенности апокалиптической религии, включая ритуалы очищения, отказ от бренного человеческого тела, предвкушение

вечной жизни и неоспоримого (в некоторой степени) харизматического лидера. Я всей душой согласен с идеей сингуляритариев о том, что ИИ — самое важное, о чем нужно размышлять в настоящий момент. Но когда дело доходит до разговоров о бессмертии, я пасую. Мечты о вечной жизни сильно искажают окружающее. Слишком многие сингуляритарии верят, что интенсивное взаимопроникновение технологий не приведет к катастрофам, которых мы могли бы ожидать от каждой из них в отдельности, или к комплексным катастрофам, которые мы тоже могли бы предвидеть, но даст вместо этого совершенно противоположный результат. Оно спасет человечество от того, чего мы больше всего боимся. От смерти.

Но как можно компетентно оценить технологии? Нужно ли, и если нужно, то как регулировать их развитие, если вы уверены, что те же самые технологии позволят вам жить вечно? Даже самые рационально мыслящие люди в мире не в состоянии трезво оценивать собственную религию. И, как говорит ученый Уильям Грасси, если обсуждается преображение немногих избранных и жизнь вечная, то не о религии ли идет речь?

> Приведет ли сингулярность к тому, что на смену человечеству придут духовные машины? Или результатом станет преображение человечества и превращение людей в сверхчеловеков, живущих вечно в гедонистическом раю рационалистов? Будет ли предшествовать сингулярности период бедствий? Появятся ли немногие избранные, которым известны будут секреты сингулярности, — авангард, а может быть, все, что останется от человечества, и суждено ли им будет увидеть землю обетованную? Все эти религиозные темы неизменно присутствуют в риторике и рассуждениях сингуляритариев, даже если они последовательно не развивают пре- и постмилленаристские[20] интерпретации, столь характерные для донаучных мессианских движений.

В отличие от вариантов развития событий по Гуду и Винджу, сингулярность Курцвейла формируется с помощью трех технологий, приближающихся к точке конвергенции, — генной инженерии, нанотехнологий и робототехники — достаточно широкий термин, который он использует для обозначения искусственного интеллекта. Кроме того, в отличие от Гуда и Винджа, Курцвейл предложил объединенную теорию технологической эволюции; которая, как всякая уважающая себя научная теория, пытается объяснить наблюдаемые явления и делает предсказания

относительно явлений будущих. Эта теория получила название Закона прогрессирующей отдачи.

Во-первых, Курцвейл предполагает, что эволюционные процессы развиваются по гладкой экспоненциальной кривой и что техническое развитие — один из таких эволюционных процессов. Как и биологическая эволюция, техника развивает у себя некую способность, а затем использует ее для развития и выхода на следующую ступень. У человека, к примеру, большой мозг и большой палец, которые позволили освоить изготовление орудий и крепкий хват, необходимый для их эффективного использования. Например, печатный пресс повлиял на технологию книжного переплета, повышение грамотности, развитие университетов и на множество других вещей. Паровой двигатель стимулировал промышленную революцию и множество изобретений.

Благодаря привычке опираться на предыдущие достижения сначала технология развивается медленно, но затем крутизна кривой ее роста увеличивается, и в конце концов она выстреливает почти вертикально. Согласно графикам и диаграммам, которые так любит Курцвейл, мы в настоящее время входим в наиболее критический период технологического развития — ту самую круто восходящую часть экспоненциальной кривой, ее «колено». Отсюда можно двигаться только вверх, и двигаться *быстро*.

Курцвейл разработал свой Закон прогрессирующей отдачи для описания эволюции любого процесса, в котором присутствуют информационные закономерности. Он применяет этот закон к биологии, где преимущества получают более сложные и упорядоченные молекулы. Но еще убедительнее этот закон используется для предсказания скорости изменения информационных технологий, включая компьютеры, цифровые камеры, Интернет, облачные вычисления, оборудование для медицинской диагностики и лечения и многое другое — любую технику, связанную с хранением и извлечением информации.

Курцвейл отмечает, что Закон прогрессирующей отдачи — это, по существу, экономическая теория. Прогрессирующий отклик подпитывается инновациями, конкуренцией, размерами рынка — характеристиками производителей и самого рынка. На компьютерном рынке эффект описывается законом Мура — еще одной экономической теорией, замаскированной под техническую и сформулированной впервые в 1965 г. одним из основателей Intel Гордоном Муром.

Закон Мура гласит, что число транзисторов, которые можно разместить на интегральной схеме для построения микропроцессора, удваивается каждые полтора года. Транзистор — это переключатель с двумя

устойчивыми положениями, способный также усиливать электрический заряд. Увеличение количества транзисторов означает большую скорость обработки информации и более быстрые компьютеры. Закон Мура означает, что компьютеры будут становиться все меньше, мощнее и дешевле с достаточно высокой и стабильной скоростью. Это происходит не потому, что закон Мура — настоящий закон природы, как закон всемирного тяготения или второй закон термодинамики. Это происходит потому, что потребитель и рынок стимулируют конкуренцию разработчиков компьютерных чипов и заставляют производить все более компактные, быстрые и дешевые компьютеры, смартфоны, камеры, принтеры, солнечные батареи и, очень скоро, 3D-принтеры. И все это развивается на базе технологий прошлого. В 1971 г. на одном кристалле помещалось 2300 транзисторов. После сорока лет, или двадцати удвоений, их уже было 2600000 000. А с этими транзисторами, более двух миллионов которых может разместиться на точке в конце этого предложения, пришла и скорость.

Приведем наглядный пример. Джек Донгарра, исследователь Национальной лаборатории Оук-Ридж (штат Теннесси) и член команды, отслеживающей скорость суперкомпьютеров, определил, что планшет-бестселлер iPad 2 фирмы Apple работает столь же быстро, как суперкомпьютер примерно 1985 г. Cray 2. Более того, iPad, работающий со скоростью более 1,5 гигафлопа (один гигафлоп соответствует одному *миллиарду* математических операций в секунду), еще в 1994 г. вошел бы в список 500 самых быстрых суперкомпьютеров мира.

В 1994 г. кто мог бы представить, что меньше чем через поколение суперкомпьютер размером меньше книжки станет уже настолько дешевым, что его будут давать школьникам? Мало того, кто бы поверил, что он способен будет без всяких проводов подключить пользователя к хранилищу знаний человечества? Только Курцвейл оказался настолько смелым: он не делал предсказаний о суперкомпьютерах, но сумел предвидеть взрывной рост интернет-технологий.

В информационных технологиях каждый прорыв дает толчок новому развитию и приближает следующий прорыв — кривая, о которой мы говорили, карабкается вверх все круче. Поэтому, если разговор заходит об iPad 2, то вопрос не в том, чего мы можем ожидать в *следующие* 15 лет; говорить нужно о том, что произойдет в течение небольшого отрезка времени. Примерно к 2020 г., по оценке Курцвейла, у нас появятся портативные компьютеры, соответствующие человеческому мозгу по вычислительной мощности, но не по интеллекту.

Посмотрим, как можно применить закон Мура к интеллектуальному взрыву. Если считать, что УЧИ достижим, то, согласно закону Мура, для получения ИСИ, или сверхчеловеческого интеллекта, не потребуется, возможно, даже рекурсивного самосовершенствования, о котором обычно говорят в контексте интеллектуального взрыва. Дело в том, что меньше чем через два года после появления УЧИ машины с интеллектом человеческого уровня должны будут удвоить скорость работы. Затем, в следующие два года, нужно ожидать очередного удвоения. Все это время средний человеческий интеллект будет оставаться неизменным, так что очень скоро ИИ человеческого уровня оставит нас далеко позади.

Что произойдет, если интеллект начнет принимать участие в собственной модификации? Елиезер Юдковски ясно объясняет, как быстро после появления УЧИ технический прогресс может выскользнуть из наших рук.

Если сегодня скорость компьютеров удваивается каждые два года, как будут развиваться события, когда ИИ, использующий эти вычислительные мощности, начнет принимать участие в исследованиях?

Скорость компьютеров удваивается каждые два года.

Скорость компьютеров удваивается каждые два года исследований.

Скорость компьютеров удваивается каждые два субъективных года исследований.

Через два года после того, как искусственный интеллект достигает человеческого уровня, скорость его работы удваивается.

Еще через год его скорость вновь удваивается. Полгода — три месяца — 1,5 месяца…

Сингулярность.

Кое-кто возражает, что закон Мура перестанет действовать еще до 2020 г., когда запихивать все больше и больше транзисторов в интегрированные микросхемы станет физически невозможно. Другие думают, что закон Мура сменится еще более стремительным удвоением, когда процессоры пройдут техническую модернизацию и станут использовать для вычислений еще более мелкие компоненты, такие как атомы, фотоны света и даже ДНК. Не исключено, что первыми преодолеют закон Мура 3D-процессорные чипы, разработанные Федеральной политехнической школой Лозанны (Швейцария). Эти процессоры еще не производятся массово. Компонуются они не горизонтально, а вертикально и будут работать быстрее и эффективнее традиционных чипов; кроме того, они готовы к параллельной обработке информации. Компания, одним из основателей которой был Гордон Мур, возможно, уже превзошла

количественные показатели его закона, создав первый *3D-транзистор.* Припомните, транзисторы — это электрические ключи. Традиционные транзисторы работают, направляя электрический ток по одному из двух путей. Новые транзисторы Tri-Gate фирмы Intel могут направлять ток по трем путям, что дает 30 %-ное увеличение скорости и экономию энергии до 50 %. По миллиарду таких транзисторов будет в каждом из чипов следующей линейки процессоров Intel.

Размещение транзисторов на кристаллах кремния играет важную роль во многих информационных технологиях от фотокамер до медицинских датчиков, так что закон Мура применим и к ним. Но Мур рассуждал об интегральных схемах, а не о множестве взаимосвязанных миров информационных технологий, включающих и продукты, и процессы. Так что более общий закон Курцвейла — Закон прогрессирующей отдачи — подходит больше. К тому же сейчас многие технологии *становятся* информационными, по мере того как компьютеры и даже роботы все более плотно вовлекаются во все аспекты разработки, производства и продажи новых продуктов. Ведь каждый производитель смартфонов — а не только их процессоров — воспользовался плодами цифровой революции. Прошло всего шесть лет с выхода первой модели iPhone, а Apple выпустила уже *шесть* версий. За это время фирме удалось более чем удвоить скорость их работы и вдвое или даже больше снизить цену для большинства потребителей. Дело в том, что скорость блоков конечного устройства все это время исправно удваивалась. Но ведь удваивалась и скорость каждого звена производственного конвейера.

Перспективы, которые прогнозирует Закон прогрессирующей отдачи, выходят далеко за рамки компьютерного бизнеса и производства смартфонов. Не так давно один из основателей Google Ларри Пейдж встретился с Курцвейлом, чтобы обсудить глобальное потепление, и расстались они на оптимистичной ноте. Через 20 лет, утверждают они, нанотехнологии позволят использовать солнечную энергию эффективней экономически, чем нефть или уголь. Эта индустрия сможет обеспечить Землю энергией на 100 %. Пока солнечная энергия обеспечивает всего 0,5 % мировых потребностей, но они утверждают, что эта доля в течение последних двадцати лет удваивается каждые два года. Так что еще через два года солнечная энергия будет составлять 1 % мирового потребления, через четыре — 2 %, а еще через 16 лет и 8 удвоений — 256 % мировых энергетических нужд. Даже с учетом роста населения и энергопотребления через 20 лет солнечной энергии должно с избытком хватать на все. Кроме того, если верить Курцвейлу и Пейджу, будет решена и проблема

глобального потепления.

Ну, и, скажем так — проблема смертности. По Курцвейлу, средства продления жизни (практически до бесконечности) уже почти готовы.

«Теперь мы имеем реальные средства разобраться в программировании жизни и начать ее перепрограммировать; мы уже умеем выключать гены без всякого вмешательства, мы можем добавлять новые гены и целые новые органы с применением стволовых клеток, — сказал Курцвейл. — Дело в том, что медицина сегодня — это информационная технология, и ее мощность будет удваиваться каждый год. Через двадцать лет эти технологии станут в миллион раз мощнее за те же деньги».

Курцвейл считает, что кратчайший путь к созданию УЧИ — обратное проектирование мозга, то есть тщательнейшее его сканирование и построение набора точно таких же схем. Эти схемы, представленные в виде алгоритмов или реальных физических контуров, нужно подключить к компьютеру как единый синтетический мозг и научить всему, что ему необходимо знать. Несколько организаций работают над подобными проектами создания УЧИ. Чуть позже мы поговорим о некоторых подходах и препятствиях в этом деле.

Скорость развития элементной базы, необходимой для построения виртуального мозга, требует более тщательного рассмотрения. Начнем, пожалуй, с человеческого мозга, а затем перейдем к компьютерам, способным его смоделировать. Курцвейл пишет, что в мозгу около 100 млрд нейронов[21], причем каждый из них связан примерно с тысячей других нейронов. Все вместе это составляет около 100 трлн межнейронных соединений. Каждое соединение способно производить порядка 200 расчетных действий в секунду (электронные схемы работают по крайней мере в 10 млн раз быстрее). Курцвейл умножает число межнейронных связей в мозгу на число действий в секунду и получает 20 квадриллионов, или 20 000 000 000 000 000 действий в секунду.

Титул самого быстрого суперкомпьютера в мире присуждается новому победителю чуть ли не ежемесячно, но в настоящий момент пальму первенства удерживает Sequoia Министерства энергетики США с шестнадцатью с лишним петафлоп. Это 16 000 000 000 000 000 операций в секунду или, грубо говоря, 80 % от скорости человеческого мозга, рассчитанной Курцвейлом в 2000 г. Однако в 2005 г. в книге «Сингулярность рядом» Курцвейл снизил свою оценку скорости мозга с 20 до 16 петафлоп и предсказал, что суперкомпьютер достигнет такой скорости к 2013 г. Sequoia сделала это на год раньше.

Неужели мы настолько близки к тому, чтобы суперкомпьютеры

сравнялись по мощности с человеческим мозгом? Цифры обманчивы. Мозг представляет собой множество параллельных процессоров и прекрасно справляется с определенными задачами, тогда как компьютеры работают последовательно и лучше справляются с другими типами задач. Мозг работает медленно, но эффективно за счет пиков нейронной активности. Компьютеры способны работать быстрее и дольше — по существу, бесконечно.

При этом человеческий мозг по-прежнему остается единственным примером продвинутого интеллекта. «Грубой силе», чтобы с ним состязаться, необходимо демонстрировать впечатляющие когнитивные достижения. Но задумайтесь о том, какие сложные системы обычно моделируют современные суперкомпьютеры! Вот лишь несколько примеров: погода, трехмерные ядерные взрывы и молекулярная динамика для производства. А что человеческий мозг? Аналогичен он по уровню сложности или порядок величины здесь выше? По всем признакам, он в той же обойме.

Возможно, все так, как говорит Курцвейл, и победа над человеческим мозгом уже совсем рядом, а за следующие 30 лет компьютерные науки пройдут тот же путь, что за 140 лет при нынешней скорости развития. Учтите также, что *создание* УЧИ — тоже информационная технология. Экспоненциальный рост скорости компьютеров позволяет и исследователям ИИ делать свою работу быстрее, то есть писать более сложные и эффективные алгоритмы, браться за непростые вычислительные задачи и проводить больше экспериментов. Более быстрые компьютеры вносят свой вклад в устойчивость индустрии ИИ; та, в свою очередь, готовит больше исследователей и выпускает более быстрые и полезные инструменты для разработки УЧИ[22].

Курцвейл пишет, что после 2029 г., когда исследователи представят компьютер, способный пройти тест Тьюринга, процессы еще ускорятся. Но он не предсказывает полноценной сингулярности раньше чем через 16 лет после этого события, то есть до 2045 г. Тогда скорость технического прогресса превзойдет способность нашего мозга управлять им. Поэтому, утверждает он, мы должны каким-то образом усилить свои мозги, чтобы удержаться на плаву. Это означает вживление каких-то подстегивающих устройств непосредственно в наши нейронные сети, точно так же, как сегодняшние кохлеарные импланты подсоединяют к слуховым нервам, чтобы преодолеть дефекты слуха. Мы взбодрим медленные межнейронные связи и сможем думать быстрее, глубже, а запоминать больше и легче. Мы получим доступ ко всем знаниям человечества и научимся, подобно

компьютерам, мгновенно обмениваться между собой мыслями и впечатлениями. В конечном итоге техника позволит нам улучшить свой мозг, перенеся его на носитель более долговечный, чем живая ткань, или загрузить его содержимое в компьютер, сохранив при этом качества, которые делают нас *нами*.

Такой образ будущего предполагает, разумеется, что внутреннее «Я» человека, его внутренняя сущность, переносимо, а это очень серьезное предположение. Но для Курцвейла это путь к бессмертию, а также море знаний и опыта за рамками всего того, что мы сегодня способны воспринять. Усиление интеллекта будет проходить настолько постепенно, что мало кто от него откажется. Но «постепенно» означает к 2045 г., то есть примерно за 30 ближайших лет, причем большая часть перемен придется на последние лет пять. Это постепенно? Мне так не кажется.

Как мы уже отмечали, Apple представила на рынок шесть версий iPhone за шесть лет. Согласно закону Мура, их техника была достаточно продвинутой, чтобы пережить за это время два удвоения мощности или даже больше, но на самом деле удвоение прошло лишь одно. Почему? Из-за времени, потерянного на разработку, прототипирование и производство компонентов iPhone, в том числе процессора, камер, оперативной и постоянной памяти, экрана и т. п., а затем на маркетинговые мероприятия и продажу самого iPhone.

Уменьшится ли время, затрачиваемое на маркетинг и продажу? Может быть, когда-нибудь техника, как и программное обеспечение, будет обновляться автоматически. Но это, вероятно, произойдет не раньше, чем наука овладеет нанотехноло- гиями или 3О-печать станет обыденной. К тому же, когда мы научимся совершенствовать компоненты собственного мозга, вместо того чтобы обновлять Microsoft Office или покушать дополнительные микросхемы памяти, это, по крайней мере поначалу, будет куда более деликатная процедура, чем все, с чем мы сталкивались прежде.

Тем не менее Курцвейл утверждает, что технический прогресс за 100 лет пройдет такой же путь, какой прежде проходил за 200000 лет. Переживем ли мы такую мощь прогресса за такое короткое время?

Николас Карр, автор книги «Пустышка»[23] (The Shallows), утверждает, что смартфоны и компьютеры снижают качество наших мыслей и изменяют форму нашего мозга. В книге «Виртуальный вы» (Virtually You) психиатр Элиас Абуджауде предупреждает нас о том, что социальные сети и ролевые игры способствуют развитию массы расстройств, включая нарциссизм и эгоизм. Погружение в технику ослабляет личность и характер, считает программист и *пионер* виртуальной реальности Джарон

Ланир, автор книги «Вы не гаджет. Манифест»[24]. Эксперты дружно предупреждают, что разрушительные эффекты исходят от компьютеров *вне* наших тел. При этом Курцвейл уверен, что компьютеры *внутри* нас принесут только пользу. Мне кажется, не стоит ожидать, что сотни тысяч лет эволюции повернут вспять всего за тридцать лет и что нас можно перепрограммировать и заставить полюбить существование, очень сильно отличающееся от всего, к чему нас готовила эволюция.

Более вероятно, что человек сам будет решать, какая скорость перемен его устраивает и с какой скоростью он справится. Каждый сможет выбрать параметры для себя сам, и многие люди остановятся на сходных скоростях, точно так же, как многие выбирают для себя сходные стили, машины и компьютеры. Мы знаем, что закон Мура и Закон прогрессирующей отдачи — это скорее экономические, нежели детерминистские законы. Если достаточное число людей, имеющих в своем распоряжении необходимое количество ресурсов, хочет искусственно ускорять собственный мозг, они создадут некоторый спрос на подобные услуги. Однако мне кажется, что Курцвейл сильно переоценивает стремление людей будущего быстрее думать и дольше жить.

В любом случае мне не кажется, что нас ожидает радостная сингулярность, которую он описывает. ИИ, разработанный без достаточных предосторожностей, не позволит ей реализоваться.

Движение к созданию УЧИ неотвратимо и, вероятно, неуправляемо. А из-за динамики удваивания, выраженной Законом прогрессирующей отдачи, УЧИ появится на мировой сцене (и захватит ее) намного скорее, чем мы думаем.

Глава 10
Сингуляритарий

В отличие от нашего интеллекта, компьютеры удваивают производительность каждые восемнадцать месяцев. Так что вполне реальна опасность, что они могли бы развить интеллект и захватить власть над миром.

Стивен Хокинг, физик

В ближайшие тридцать лет мы получим техническую возможность создать сверхчеловеческий интеллект. Вскоре после этого человеческая эпоха закончится. Можно ли избежать такого прогресса? Если избежать нельзя, то можно ли так направить события, чтобы мы могли уцелеть?

Вернор Виндж, писатель, профессор, компьютерщик

Ежегодно, начиная с 2005 г., Исследовательский институт искусственного интеллекта проводит конференцию по сингулярности. В течение двух дней выступающие вещают перед аудиторией примерно в тысячу человек о сингулярности в самых разных аспектах — о ее влиянии на рынок труда и экономику, здравоохранение и продолжительность жизни, о ее этических последствиях. В 2011 г. среди выступавших в Нью-Йорке были научные легенды. Среди них математик Стивен Вольфрам и доткомовский миллиардер Питер Тиль, который дает талантливым подросткам-технарям деньги, чтобы они вместо учебы в колледже основывали собственные компании. Был там и Дэвид Ферруччи, научный руководитель проекта DeepQA/Watson. Непременно выступает Елиезер Юдковски, и, как правило, находится пара специалистов по этике и представителей экстропианского и трансгуманистического сообществ. Экстропианцы исследуют технологии и медицинские методики, которые позволят человеку жить вечно. Трансгуманисты думают о технических приспособлениях и косметических способах повышения человеческих

способностей, красоты и… возможности жить вечно. И над всем этим разнообразием тем и фракций возвышается колосс сингулярности, один из основателей Конференции по сингулярности и звезда каждого такого собрания Рэй Курцвейл.

Темой конференции 2011 г. был компьютер Watson проекта DeepQA (ответы на вопросы) фирмы IBM. Курцвейл провел рутинную презентацию по истории чат-ботов и систем ответов на вопросы под названием «От Eliza к Watson». Но в середине, заметно оживившись, он резко раскритиковал неуклюжее эссе, направленное против гипотезы сингулярности, одним из соавторов которого выступил сооснователь Microsoft Пол Аллен.

Курцвейл выглядел неважно — исхудал, слегка прихрамывал, вел себя тише обычного. Вообще, он не из тех ораторов, которые легко захватывают внимание зала или пересыпают свою речь шутками. Напротив, он говорит всегда мягко и даже чуть механистично; его речь, кажется, лучше всего приспособлена для переговоров с похитителями людей или рассказывания вечерних сказок. Но на фоне тех революционных идей, о которых он обычно говорит, такая манера срабатывает неплохо. В век, когда доткомовские миллиардеры проводят презентации в отглаженных джинсах, Курцвейл выходит к аудитории не иначе как в старомодных коричневых брюках, кожаных мокасинах с кисточками, спортивном пиджаке и очках. Он человек средних габаритов, но в последнее время заметно постарел, особенно по сравнению с энергичным Курцвейлом из моих воспоминаний. Ему было всего лишь 52 года, когда я в последний раз брал у него интервью, и он тогда еще не перешел на питание с интенсивным снижением калорийности; сегодня это часть его планов по замедлению старения. Точно подобрав рацион питания и пищевых добавок, а также занятий физкультурой, Курцвейл планирует дожить до дня, когда будет найдено лекарство от смерти. В том, что оно будет найдено, он не сомневается.

«Я оптимист. Как изобретатель я просто обязан быть оптимистом».

После выступления Курцвейла мы с ним сидели рядом на металлических стульях в небольшой гостиной этажом выше зала. За дверью возможности пообщаться с ним не было — после меня ожидала съемочная группа какого-то документального фильма. Всего десять лет назад, когда он был не очень знаменитым изобретателем и писателем, я со своей собственной съемочной группой монополизировал его на три очень приятных часа; теперь же он стал фактически символом, и на его внимание я мог рассчитывать ровно до тех пор, пока мне удавалось держать дверь запертой. Я тоже изменился — во время первой встречи меня ошеломила

мысль о том, что когда-нибудь станет возможным подключить мой мозг к компьютеру, как описано в «Духовных машинах». Мои вопросы были не жестче мыльных пузырей. Теперь я стал циничнее и мудрее по отношению к опасностям, которые самого мэтра уже не интересуют.

«Я достаточно много говорю об опасности в книге "Сингулярность рядом", — запротестовал Курцвейл, когда я спросил, не делает ли он слишком сильный акцент на позитивных перспективах сингулярности и не преуменьшает ли ее опасности. — В главе 8 речь идет об опасных перспективах переплетенных ГНР [генетики, нанотехнологий и робототехники], и я в милых живописных подробностях описываю отрицательные стороны этих трех областей техники. А отрицательная сторона робототехники, под которой на самом деле подразумевается ИИ, наиболее серьезна, потому что интеллект — самый значительный феномен в мире. Следовательно, от сильного ИИ не существует абсолютно надежной защиты».

В книге Курцвейла действительно подчеркиваются опасности генной инженерии и нанотехнологий, однако мощному ИИ, как раньше называли УЧИ, посвящено всего несколько «беззубых» страниц. Кроме того, в этой главе автор утверждает, что отказ от некоторых слишком опасных технологий, как предлагают Билл Джой и другие, — это идея не просто плохая, а аморальная. Я согласен с тем, что такой отказ практически невозможен. Но аморален ли он?

> Отказ от технологий аморален потому, что в этом случае мы отказываемся от громадной пользы. В нашей жизни по-прежнему много страдания, которое можно преодолеть; существует моральный императив преодолеть его. Кроме того, отказ от технологий потребовал бы тоталитарной системы, поскольку только ей под силу обеспечить запрет. И, что самое важное, все равно ничего не получится. Технологии будут просто загнаны в подполье, где у безответственных исследователей вообще не будет никаких ограничений. Ответственные ученые, которым будет поручена разработка защиты, не получат доступ к необходимым для этого инструментам. В результате ситуация окажется еще более опасной.

Курцвейл критикует так называемый Принцип предосторожности, исходящий от движения за охрану окружающей среды; это, как и отказ от технологий, его мишень в нашем разговоре. Но важно точно

сформулировать этот принцип и посмотреть, почему в нем нет особого смысла. Принцип предосторожности гласит: «Если последствия некоего действия неизвестны, но некоторые ученые считают, что они, хотя бы с небольшой вероятностью, могут оказаться резко отрицательными, то лучше не производить действие, чреватое отрицательными последствиями». Этот принцип применяется не особенно часто и не особенно строго. Он положил бы конец любой предположительно опасной технологии, если бы «некоторые ученые» опасались ее, даже если бы они не могли точно сформулировать, каким образом это действие приведет к нежелательному результату.

В приложении к УЧИ у Принципа предосторожности и отказа от технологий нет никаких шансов, если исключить вариант катастрофического происшествия на пути к УЧИ, которое испугало бы нас всех до смерти. Лучшие корпоративные и правительственные проекты УЧИ никогда не откажутся от конкурентного преимущества в виде секретности — мы уже видели это на примере стелс-компаний. Мало какая страна или организация откажется от такого преимущества, даже если работа над УЧИ будет объявлена вне закона. (Мало того, та же корпорация Google обладает средствами и влиянием небольшого государства, так что, если хотите понять, что будут делать другие страны, следите за Google.) Технологии, необходимые для создания УЧИ, вездесущи и многосторонни, к тому же постоянно миниатюризируются. Контролировать их развитие чрезвычайно сложно, если не невозможно.

Другой вопрос — действительно ли аморально не развивать УЧИ, как говорит Курцвейл. Во-первых, польза от УЧИ будет громадна, но только в том случае, если человечество уцелеет и сможет ею воспользоваться. И это довольно серьезное *если*, когда система достаточно продвинута, чтобы инициировать интеллектуальный взрыв. Утверждать, как это делает Курцвейл, что недоказанная польза перевешивает столь же недоказанный риск, на мой взгляд, проблематично. Я бы сказал в ответ на его аргументы, что аморально, на мой взгляд, разрабатывать такие технологии, как УЧИ, и не просвещать одновременно как можно большее количество людей относительно связанных с этим рисков. Я считаю, что катастрофические риски УЧИ, ныне признанные многими известными и уважаемыми исследователями, установлены надежнее, чем предполагаемая польза от сингулярности, — для начала это наноочищение крови, улучшенный, более быстрый мозг и бессмертие. Единственное, что можно наверняка сказать о сингулярности, — это то, что данный термин описывает период времени, в который, согласно Закону прогрессирующей отдачи, быстрые и очень

умные компьютеры прочно войдут в нашу жизнь и наши тела. После этого, вполне возможно, чуждый машинный интеллект устроит нашему природному интеллекту веселую жизнь. Понравится нам это или нет, другой вопрос. Если вы внимательно читали Курцвейла, вы понимаете, что польза проистекает в основном от усиления человеческих способностей — усиления, необходимого, чтобы поспевать за изменениями. Как я уже говорил, я считаю, что это приведет к технологическому насилию, а может быть, и к выбраковке.

И это, надо сказать, не главный мой страх, потому что я не думаю, что мы когда-нибудь доживем до подобной ситуации. Я считаю, что нас остановят еще на подходе силы, слишком могущественные, чтобы их можно было контролировать. Я сказал об этом Курцвейлу, и он ответил стандартной заготовкой — тем же оптимистичным, антропоморфным аргументом, что он выдал мне десять лет назад.

> Получить сущность чрезвычайно умную, но по каким-то причинам нацеленную на наше уничтожение — это негативный сценарий. Но спросите себя: почему может так случиться? Во-первых, я осмелился бы утверждать, что не существует противостояния между нами и машинами, поскольку машины — не отдельная цивилизация. Это часть нашей цивилизации. Это инструменты, которыми мы пользуемся, и их применение постоянно расширяется; даже если мы *станем* этими инструментами, все равно эта система развивается из нашей цивилизации. Это не какое-то инопланетное вторжение машин с Марса. Нам не придется гадать, на чем основана их система ценностей.

Как мы уже говорили, считать, что УЧИ будет в точности похож на человека, означает приписывать человеческие ценности разумной машине, которая получила свой разум — и свои ценности — совершенно иначе, чем мы. Несмотря на самые лучшие намерения создателей, в большинстве, если не во всех УЧИ значительная часть работы системы окажется слишком непрозрачной и слишком сложной, чтобы мы могли полностью в ней разобраться и точно предсказать ее поведение. Чуждый, непознаваемый, да к тому же… не забывайте, что некоторые УЧИ будут созданы специально для того, чтобы убивать людей, ведь в США, к примеру, мы видим ведомства оборонного сектора среди самых активных инвесторов. Можно ожидать, что в других странах дело обстоит точно так же.

Я уверен, что Курцвейл понимает, что для того, чтобы УЧИ уничтожил человечество, не обязательно разрабатывать его специально с этой целью; достаточно простой небрежности. Как предупреждает Стив Омохундро, без тщательного программирования у продвинутого ИИ появятся мотивации и цели, которые будут нам чужды. Как говорит Елиезер Юдковски, наши атомы могут пригодиться ИИ в каком-то совсем другом качестве. И, как мы уже видели, дружественный ИИ, который обеспечил бы правильное поведение первого УЧИ и всех его потомков, — это концепция, которой еще очень далеко до готовности.

Курцвейл не стал уделять слишком много времени концепции дружественного ИИ. «Мы не можем просто сказать: "Мы поместим в наши ИИ эту маленькую подпрограммку, и это их обезопасит", — сказал он. — Я имею в виду, все сводится к целям и намерениям этого искусственного интеллекта. Перед нами пугающие вызовы».

Вызывает ужас одна только мысль о недружественном ИИ — УЧИ, разработанном с целью уничтожения врагов; с этой реальностью мы с вами очень скоро столкнемся лицом к лицу. «Почему бы могла появиться такая вещь?» — спрашивает Курцвейл. Да просто потому, что десятки организаций в США работают над ее созданием, и тем же самым занимаются за границей наши противники. Если бы сегодня УЧИ уже существовал, то очень скоро, не сомневаюсь, он нашел бы применение в боевых роботах. Возможно, DARPA настаивает, что беспокоиться не о чем, — ИИ, созданные на деньги DARPA, будут убивать только наших врагов. Его создатели установят защиту, позаботятся о безотказности, добавят аварийный блокиратор и систему секретной связи. Они смогут контролировать сверхразум.

В декабре 2011 г. один иранец запустил на своем ноутбуке простую программу обмена файлами — и устроил крушение беспилотника Sentinel. В июле 2008 г. кибератака на Пентагон позволила злоумышленникам получить неограниченный доступ к 24000 секретных документов. Бывший заместитель министра обороны Уильям Линн III рассказал газете *The Washington Post*, что в результате сотен кибератак на системы министерства обороны и его подрядчиков были украдены «наши важнейшие секреты, включая системы авионики, технологии наблюдения, системы спутниковой связи и сетевые протоколы безопасности». Тому, кто не способен даже на такое относительно простое дело, как обезопасить информацию от хакеров-людей, бессмысленно доверять охрану сверхразума.

Однако мы можем извлечь кое-какие полезные уроки из истории контроля над вооружениями. С момента создания ядерного оружия только

США реально использовали его против неприятеля. Ядерным державам удалось избежать гарантированного взаимного уничтожения. Ни в одной ядерной державе, насколько нам известно, не происходило случайных взрывов. Пока история ответственного руководства ядерной отраслью заслуживает высокой оценки (хотя угроза еще не миновала). Но вот что я хочу сказать. Слишком мало людей знает, что нам необходимо налаживать постоянный международный переговорный процесс по поводу УЧИ, сравнимый с тем, что идет по поводу ядерного оружия. Слишком много людей полагает, что пределы ИИ обозначены безобидными поисковыми системами, смартфонами и вот теперь компьютером Watson. Но в реальности УЧИ намного ближе к ядерному оружию, чем к видеоиграм.

ИИ — технология «двойного назначения»; этим термином обозначают технологии, имеющие и мирные, и военные приложения. К примеру, ядерный синтез может обеспечивать города электричеством или разрушать их (или, как в случае с Чернобылем и Фукусимой, делать то и другое последовательно). Ракеты, сконструированные в ходе космической гонки, увеличили мощность и точность межконтинентальных баллистических ракет. Нанотехнологии, биоинженерия и генная инженерия обещают невероятные успехи в улучшающих жизнь гражданских приложениях, но они же буквально напрашиваются на катастрофические аварии и использование военными и террористами.

Когда Курцвейл объявляет себя оптимистом, он не подразумевает, что УЧИ окажется безобидным. Он говорит о том, что готов к взвешенному подходу, обычному для людей, сталкивающихся с потенциально опасными технологиям. И иногда люди терпят неудачу.

«Идет много разговоров об экзистенциальных рисках, — сказал Курцвейл. — Меня больше тревожат неприятные эпизоды, они куда более вероятны. Вы знаете, шестьдесят миллионов человек были убиты во время Второй мировой войны. Разумеется, эта цифра увеличилась за счет мощных разрушительных средств, которыми мы тогда располагали. Я настроен довольно оптимистично и считаю, что мы прорвемся, но я не настолько оптимистичен, чтобы думать, что нам удастся избежать болезненных эпизодов».

«Невероятные перспективы сочетаются с большими опасностями еще со времен приручения огня. Огонь помогал готовить еду, и им же сжигали вражеские деревни. Колесо может быть использовано и во благо, и во зло. Технология — это мощь; одна и та же технология может применяться с различными целями. Человеческие существа делают все на свете, занимаются и любовью, и войной; любое дело мы стремимся улучшить

всеми возможными способами, в том числе техническими; мы будем этим заниматься и впредь».

Разные способы применения технологий неизбежны, и несчастные случаи всегда возможны — с этим сложно спорить. Тем не менее аналогия не работает — продвинутый ИИ совсем не похож на огонь или какую-то другую известную технологию. ИИ способен думать, планировать и обманывать своих создателей. Ни один другой инструмент не делает ничего подобного. Курцвейл считает, что существует способ ограничить опасные аспекты ИИ, и особенно ИСИ, и способ этот — соединить его с человеком при помощи машинного усиления интеллекта. Сидя на неудобном металлическом стуле, оптимист сказал: «Как я уже указывал, мощный ИИ рождается из множества разнонаправленных усилий и глубоко интегрируется в инфраструктуру нашей цивилизации. В самом деле, он будет буквально встроен в наше тело и мозг. В этом качестве он начнет отражать наши ценности, поскольку он станет нами».

Из этих рассуждений следует, что ИИ будет настолько же «безопасен», насколько безопасны мы сами. Однако, как я сказал Курцвейлу, homo sapiens не славится особой безобидностью при контактах с сородичами, другими животными и окружающей средой. Кто из вас убежден, что человек, снабженный усилителем мозга, окажется более дружелюбным и доброжелательным, чем машинный сверхразум? Усовершенствованный человек (те, кто с нетерпением ждет такой возможности, называют его сверхчеловеком) может обойти проблему базовых потребностей ИИ, сформулированную Омохундро. Не исключено, что такой человек будет обладать не только самосознанием и способностью к самосовершенствованию, но и встроенным набором отработанной веками человекоцентрической этики — и она подавит базовые потребности, которые Омохундро выводит из модели рационального экономического агента. Однако, несмотря на «Цветы для Элджернона»[25], мы понятия не имеем, что происходит с человеческой этикой, когда интеллект взлетает до небес. Можно привести множество примеров людей средних интеллектуальных способностей, которые вдруг начинают войну против собственной семьи, школы, фирмы или соседей. Гении, кстати говоря, тоже способны наделать шума — лучшие генералы мира по большей части не были идиотами. А сверхразум вполне может оказаться умножителем агрессии. Он может превратить обиду в убийства, разногласия в катастрофы — точно так же, как «удачно» подвернувшееся оружие может превратить потасовку в убийство. Мы просто не знаем. Однако следует заметить, что для ИСИ, созданного за счет усиления интеллекта,

характерна биологическая агрессивность, которой у машин нет. Наш биологический вид имеет длинный послужной список самозащиты, объединения ресурсов, откровенного уничтожения противников и других движителей, о наличии которых у сознающих себя машин мы можем пока только гадать.

И кто же первый сможет воспользоваться достижениями науки и серьезно усовершенствовать себя? Самые богатые? Раньше считалось, что зла в богатых не больше, чем во всех остальных, но недавнее исследование Университета Калифорнии в Беркли позволяет усомниться в этом. Эксперименты показали, что богатейшие представители высших классов с большей вероятностью «демонстрируют тенденции к принятию неэтичных решений, отнимают у других ценные вещи, лгут на переговорах, обманывают с целью увеличить свои шансы на выигрыш и одобряют неэтичное поведение на работе». Нет недостатка в обеспеченных руководителях компаний и политиках, у которых одновременно с подъемом к вершинам власти, кажется, постепенно ослабевал моральный компас (если, конечно, он у них имелся изначально). И при этом политики и бизнесмены будут первыми, кто подвергнется серьезному интеллектуальному усовершенствованию?

Или, может быть, первыми будут солдаты? DARPA давно взяла на себя львиную долю расходов, так что разумно предположить, что первая модификация мозга будет испытана на поле боя или в Пентагоне. И если сверхразум сделает солдат сверхдружелюбными, DARPA потребует вернуть свои деньги.

Не исключено, конечно, что, когда дело дойдет до модификации человека, мир будет лучше подготовлен к этому, нежели сейчас, и позаботится о каких-то предохранителях, которые мы сегодня даже представить не можем. Кроме того, множество ИСИ, вероятно, лучше, чем один-единственный. Хорошо бы предусмотреть способ непрерывного мониторинга и наблюдения за ИИ; как ни парадоксально, лучшими сторожами для этого были бы другие ИИ. Подробнее мы поговорим о способах защиты от ИСИ в главе 14. Дело в том, что модификация разума не застрахована от моральных отказов. Сверхразум может оказаться опаснее любой из самых тщательно контролируемых технологий сегодняшнего дня и любого вида оружия.

Одновременно с технологией модификации разума нам придется развивать науку отбора кандидатов на эту операцию. Сингуляритарии тешат себя надеждой, что каждый, кто сможет себе это позволить, получит суперинтеллект; на самом же деле такой подход гарантированно привел бы

к тому, что все остальные оказались бы игрушкой в руках первого же злобного сверхразума, полученного таким образом. Мы уже говорили о том, что первый в разработке УЧИ получит решительное преимущество. Тот, кто раньше всех сумеет получить УЧИ, — кем бы он ни был, — вероятно, создаст затем условия для интеллектуального взрыва. Он сделает это из опасения, что конкуренты, военные или корпоративные, обгонят его; никто не знает, насколько конкуренты близки к финишной черте. Я видел, какая гигантская пропасть отделяет создателей ИИ и УЧИ от исследования рисков, которые они должны были бы проводить. Мало кто из разработчиков УЧИ, с которыми я разговаривал, читал хоть что-то из работ Исследовательского института машинного интеллекта, Института будущего человечества, Института этики и новых технологий или Стива Омохундро. Многие даже не подозревают о существовании растущего сообщества людей, озабоченных разработкой интеллекта, превышающего человеческий, и проводящих важные исследования в отношении возможных катастрофических последствий этого. Если ситуация не изменится, я уверен, что короткий забег от УЧИ до ИСИ пройдет без достаточных мер предосторожности, которые могли бы предотвратить катастрофу.

Вот яркий пример. В августе 2009 г. в Калифорнии Ассоциация за продвижение искусственного интеллекта (AAAI) собрала группу, которая должна была успокоить растущий страх публики перед взбесившимися роботами, потерей неприкосновенности частной жизни и техническими движениями, названия которых больше всего напоминали названия религиозных сект.

«За последние пять-восемь лет произошло кое-что новое, — сказал организатор Эрик Хорвиц, видный исследователь из Microsoft. — Технари предлагают почти религиозные идеи, и эти идеи резонируют в некоторых отношениях все с той же идеей Вознесения… По моим ощущениям, рано или поздно нам придется сделать какое-то заявление или утверждение, учитывая все более громкие голоса технократов и людей, которых очень беспокоит развитие интеллектуальных машин».

Несмотря на многообещающую повестку дня, встречу можно записать в разряд упущенных возможностей. Она не была открыта для публики или прессы; специалисты по машинной этике и другие мыслители, занятые оценкой рисков, тоже были оставлены за бортом. К дискуссии были приглашены только чистые компьютерщики. Это примерно как попросить автогонщиков установить в городе пределы разрешенных скоростей. Одна из подгрупп работала над Тремя законами робототехники Айзека Азимова;

это означает, что этические дискуссии не были обременены многими томами трудов, написанных позже и продвинувшихся куда дальше этих научно-фантастических декораций. Тоненький отчет о конференции, представленный Хорвицем, выражает скепсис по поводу интеллектуального взрыва, сингулярности и утраты контроля над разумными системами. Тем не менее конференция призвала к дальнейшим исследованиям в области этики и психологии и подчеркнула опасность все более сложных и непонятных компьютерных систем, включая «дорогостоящее непредсказуемое поведение автономных или полуавтономных систем, способных принимать решения». Том Митчелл из Университета Карнеги-Меллона, создатель под крылом DARPA архитектуры NELL (потенциальным УЧИ), утверждает, что конференция изменила его позицию. «Направляясь туда, я был настроен очень оптимистично в отношении будущего ИИ и считал, что Билл Джой и Рэй Курцвейл сильно ошибаются в своих предсказаниях. После этой встречи я захотел большей откровенности в этих вопросах».

В книге «Сингулярность рядом» Курцвейл предлагает несколько решений проблемы беглого ИИ. Это удивительно слабые решения для человека, который занимает практически монопольную позицию представителя суперинтеллекта. Но, с другой стороны, это вовсе не удивительно. Как я уже говорил, существует неразрешимый конфликт между теми, кто лихорадочно жаждет жить вечно, и всем тем, что обещает замедлить, поставить под угрозу или как-то иначе затруднить развитие технологий, которые должны в конечном итоге исполнить их мечту. В своих книгах и лекциях Курцвейл почти не тратит свой интеллект на опасности ИИ и предлагает лишь несколько решений, но при этом утверждает, что подробно разобрал и то и другое. В тесной гостиной Нью-Йорка, когда за дверью нервно готовилась съемочная группа, я спросил себя: чего мы можем ожидать от одного человека? Неужели Курцвейл должен в одиночку справиться с сингулярностью, с ее пользой и опасностями, и скормить нам то и другое с ложечки, как младенцам? Неужели он лично должен разбираться, что означают идиомы вроде «непреодолимо двойственная природа техники», и овладевать философией выживания, как ее понимают люди вроде Юдковски, Омохундро и Бострома?

Нет, я так не думаю. Это проблема, которую мы должны решать все вместе, с привлечением специалистов.

Глава 11
Жесткий старт

День ото дня машины получают над нами преимущество; день ото дня мы все больше подчиняемся им; с каждым днем все больше людей рабски трудятся, ухаживая за ними, с каждым днем все больше людей отдают свои жизненные силы на развитие механической жизни. Развязка — всего лишь вопрос времени, но сам факт того, что придет время, когда машины получат реальную власть над миром и его обитателями, не вызывает ни малейших сомнений ни у кого из тех, кто обладает по-настоящему философским складом ума.

Сэмюел Батлер, английский поэт и писатель XIX века

Более чем когда-либо в истории человечество находится на распутье. Одна дорога ведет к отчаянию и совершенной безнадежности, другая — к полному уничтожению. Будем же молиться, чтобы нам хватило мудрости сделать верный выбор.

Вуди Аллен

Джон Гуд не является изобретателем интеллектуального взрыва, точно так же, как сэр Исаак Ньютон не является изобретателем гравитации. Все, что он сделал, — это обратил внимание на то, что некое событие, которое он считал и неизбежным, и в целом положительным для человечества, наверняка приведет к рождению «ультраинтеллекта», а нам, человечеству, придется решать задачи, слишком для нас сложные. Затем, прожив еще три десятилетия, Гуд изменил свое мнение. Мы сотворим разумные машины по нашему образу и подобию, и они уничтожат нас. Почему? По той же причине, по которой мы никогда бы не договорились запретить исследования ИИ и по которой Busy Child, скорее всего, получит свободу. По той же причине, по которой рациональный насквозь Стив Омохундро,

да и все остальные специалисты, с которыми я встречался, уверены, что остановить разработку УЧИ до того момента, пока мы не будем больше знать о связанных с ним опасностях, нет никаких шансов.

Мы не прекратим разработку УЧИ, потому что больше самого ИИ боимся, что другие страны будут продолжать совершенствовать УЧИ, что бы ни говорило и ни делало международное сообщество. Мы сочтем, что лучше первыми прийти к цели и взять приз. Интеллектуальная гонка в разгаре и, к разочарованию многих, эта глобальная гонка может оказаться более опасной, чем та, из которой мы, вроде бы, только что вышли, — гонка ядерных вооружений. Мы пойдем за политиканами и энтузиастами технического прогресса к роковому концу, говоря словами Гуда, «подобно леммингам».

Позитивная сингулярность Рэя Курцвейла не требует интеллектуального взрыва — Закон прогрессирующей отдачи гарантирует продолжение экспоненциального роста информационных технологий, включая и те, что кардинально меняют мир, такие как УЧИ, а позже ИСИ. Вспомните, что для интеллектуального взрыва, по Гуду, необходим УЧИ. Результатом взрыва будет интеллект выше человеческого или ИСИ. Курцвейл утверждает, что УЧИ будет покорен, поначалу медленно, а затем разом, в соответствии с Законом прогрессирующей отдачи.

Курцвейла не беспокоят препятствия на пути к УЧИ, поскольку сам он предпочитает построение УЧИ путем обратного проектирования мозга. Он считает, что в мозге (включая сознание) нет ничего, что невозможно было бы компьютеризировать. Действительно, все специалисты, с которыми я говорил, уверены, что интеллект компьютеризируем. Мало кто считает, что для получения ИСИ после создания УЧИ необходим интеллектуальный взрыв в гудовском смысле. Медленный уверенный прогресс должен привести к тому же, но Курцвейл настаивает, что прогресс, вероятно, будет не медленным и равномерным, а быстрым и к тому же ускоряющимся.

Однако после построения почти любой системы УЧИ[26] интеллектуальный взрыв может оказаться неизбежным. Когда система осознает себя и получит возможность самосовершенствоваться, ее базовые потребности, согласно Омохундро, практически гарантируют, что она будет стремиться улучшать себя снова и снова.

Итак, неизбежен ли интеллектуальный взрыв? Или что-нибудь может его остановить?

Те, кто считает, что создать УЧИ крайне затруднительно, группируются вокруг двух основных идей: экономики и сложности программного обеспечения. Сторонники первой — экономической —

считают, что на переход от ИИ к гораздо более сложной и мощной когнитивной архитектуре УЧИ просто не хватит денег. Редко какой из проектов УЧИ не страдает от недостатка финансирования. Исходя из этого, некоторая часть исследователей считает, что их отрасль в состоянии бесконечного застоя, так называемой ИИ-зимы. Застой удастся преодолеть, если правительство или какая-нибудь корпорация вроде IBM или Google присвоит УЧИ высший приоритет и организует проект масштабов Манхэттенского с приложением соответствующих усилий. Во время Второй мировой войны авральная разработка атомного оружия стоила правительству США примерно $2 млрд в современных ценах, и работали над ней около 130000 человек. Манхэттенский проект часто всплывает в разговорах тех исследователей, которые хотят получить УЧИ *поскорее*. Но кто пойдет на это и почему?

Сторонники идеи о сложности программного обеспечения утверждают, что проблема создания УЧИ попросту слишком сложна для человечества, как бы долго мы над ней ни бились. Философ Дэниел Деннетт считает, что мы, возможно, не обладаем разумом, способным понять наш собственный разум. Скажем, золотая рыбка не может объяснить, как работает ее мозг. Вероятно, человеческий разум не самый сильный из всех возможных, однако для того, чтобы полностью разобраться в его работе, может потребоваться интеллект более мощный, чем наш.

Чтобы узнать больше о правдоподобности утверждений «пораженцев», я обратился к человеку, с которым неоднократно сталкивался на конференциях по ИИ и чьи блоги, статьи и заметки часто читал в Сети. Этот человек — разработчик ИИ, опубликовавший множество очерков и интервью плюс *девять* толстых книг и бесчисленное количество научных статей. Я бы не удивился, обнаружив в его доме в пригороде Вашингтона робота, вкалывающего круглые сутки, готовя статьи для доктора Бенджамина Гертцеля, чтобы тот мог разъезжать по конференциям. Этот человек успел дважды жениться, завести троих детей и поработать в университетах на кафедрах информатики, математики и психологии в США, Австралии, Новой Зеландии и Китае. Он — организатор единственной ежегодной конференции по ИИ человеческого уровня и главный популяризатор термина УЧИ. Он же — исполнительный директор двух технических компаний, причем одна из них — Novamente — входит, по мнению некоторых экспертов, в короткий список фирм, имеющих максимальные шансы первыми прийти к финишу и создать УЧИ.

Говоря в общем, когнитивная архитектура Гертцеля, получившая

название OpenCog, — подход инженеров-компьютерщиков по всем правилам науки. Исследователи, которые считают необходимым опираться на информатику, хотят сконструировать УЧИ с архитектурой, работающей аналогично нашему мозгу, как его работу описывает когнитивистика. Когнитивистика, в свою очередь, пользуется данными таких наук, как лингвистика, психология, антропология, педагогика, философия и др. Исследователи-компьютерщики считают, что создание разума *в точности* по образцу мозга — путем обратного проектирования, как рекомендуют Курцвейл и другие специалисты, — излишне затратно. Кроме того, мозг по конструкции не оптимален — можно сделать и лучше. В конце концов, рассуждают они, человеку, чтобы научиться летать, не потребовалось заниматься обратным проектированием птицы. Принципы полета были установлены путем экспериментов и по наблюдениям за птицами. За этим последовали изобретения. Когнитивистика — «принципы полета» для мозга.

Основная концепция OpenCog — то, что разум основан на высокоуровневом распознавании образов. Обычно «образами» в ИИ являются блоки данных (файлы, картинки, текст, различные другие объекты), которые были или будут классифицированы — скомпонованы по категориям — системой, предназначенной для работы с данными. Антиспам-фильтр, работающий в вашей почтовой программе, — отличный специалист по распознаванию образов, он отслеживает одну или несколько характеристик нежелательных почтовых отправлений (к примеру, слова «оздоровление мужского организма» в теме письма) и направляет их в отдельную папку.

В архитектуре OpenCog понятие распознавания образов несколько тоньше. Образ, который система ищет в каждой вещи или идее, закодирован в небольшой программе, содержащей своего рода описание искомого. Это «концепт», или машинный аналог мысленного образа. К примеру, когда вы видите собаку, вы мгновенно многое узнаете о ней — у вас в памяти уже имеется *концепт* собаки. У нее влажный нос, она любит ветчину, она линяет и гоняется за кошками. Концепт собаки содержит немало информации.

Когда датчики OpenCog замечают собаку, мгновенно запускается *программа* собаки, которая сосредоточивает внимание машины на концепте собаки. На основании данных об этой или какой-то другой конкретной собаке OpenCog может добавить в концепцию собаки новую информацию.

Отдельные модули OpenCog будут выполнять такие функции, как восприятие, внимание и память. Делается это при помощи схожих, но

индивидуально настроенных программных комплексов, включающих генетическое программирование и нейронные сети.

Затем начинается обучение. Гертцель планирует «вырастить» свой ИИ в виртуальном компьютерном мире, таком как Second Life; процесс углубляющего обучения может продолжаться не один год. Как и другие проектировщики когнитивных архитектур, Гертцель считает, что разум должен быть «воплощен… более или менее по-человечески», даже если его тело существует только в виртуальном мире. Тогда этот разумный агент-младенец сможет наращивать свою коллекцию фактов о мире, в котором обитает. В фазе обучения, которую Гертцель выстраивает по теориям развития ребенка психолога Жана Пиаже, «маленький» OpenCog мог дополнить уже имеющиеся у него знания за счет доступа к одной из коммерческих баз общеизвестных фактов.

Одно из таких гигантских хранилищ знаний называется Cyc, от en**cyc**lopedia («энциклопедия»). Эта база, созданная компанией Cycorp, содержит около миллиона понятий и около 5 млн правил и фактов о связях между этими понятиями. Потребовалось более тысячи человеко-лет, чтобы вручную запрограммировать всю эту информации в логике первого порядка — формальной системе, которая используется в математике и информатике для представления утверждений и зависимостей. Cyc — громадный источник человеческих знаний; он неплохо (до 40 %) «понимает» английский язык. Cyc «знает», к примеру, что такое дерево, и знает, что у дерева есть корни. Он знает, что у человеческой семьи тоже есть корни, а также фамильное древо. Он знает, что подписка на газету прекращается, если человек умирает, и что в чашке может содержаться жидкость, которую можно выливать оттуда быстро или медленно.

Сверх того, у Cyc имеется генератор «рассуждений». Рассуждение — это способность делать выводы из имеющихся данных. Генератор рассуждений Cyc воспринимает вопросы и генерирует ответы на них на основе обширной базы данных.

Cyc создан пионером ИИ Дугласом Ленатом и является крупнейшим проектом ИИ в истории; вероятно, он отличался также лучшим финансированием — начиная с 1984 г. в него было вложено $50 млн в виде грантов от правительственных агентств, включая DARPA. Создатели Cyc и сейчас продолжают совершенствовать его базу данных и генератор рассуждений, добиваясь, чтобы он лучше обрабатывал «естественный язык», то есть обычный повседневный письменный язык. Как только машина в достаточной мере научится усваивать тексты на естественном языке, ее создатели поручат ей читать — и понимать — все подряд

интернет-странички

Еще один претендент на роль самой знающей и информативной базы данных уже занимается этим. Система NELL (Never Ending Language Learning) Университета Карнеги-Мел- лона знает более 390000 фактов об окружающем мире. Этот проект финансируется агентством DARPA. Работая круглосуточно и без выходных, NELL просматривает сотни миллионов веб-страниц в поисках текстовых закономерностей, которые позволят ей узнать еще больше. Она классифицирует факты по 274 категориям, включая города, знаменитостей, растения, спортивные команды и т. д. Система знает множество кросскатегорийных фактов, к примеру, то, что Майами — город, где базируется футбольная команда «Дельфины Майами» (Miami Dolphins). NELL может самостоятельно *сделать вывод* о том, что эти дельфины — не морские млекопитающие, весело играющие в волнах.

NELL использует неформальные человеческие ресурсы — пользователей Интернета. Университет Карнеги-Меллона приглашает всех желающих выходить в Сеть и помогать в обучении NELL, анализируя ее базу данных и исправляя ошибки.

Знания, а также опыт и мудрость — ключ к УЧИ, поскольку без них искусственный интеллект человеческого уровня просто немыслим. Так что любая система УЧИ обязательно должна научиться усваивать знания — то ли через воплощение в теле, способном воспринимать и усваивать знания, то ли напрямую из Интернета, изучив все его содержимое. И чем быстрее, тем лучше, говорит Гертцель.

Продвигая собственный проект, непоседливый Гертцель делит свое время между Гонконгом и Роквиллем (штат Мэриленд). Однажды весенним утром я обнаружил в его дворе видавший виды батут и микроавтобус Honda, настолько потрепанный, что создавалось впечатление, будто он прошел сквозь пояс астероидов. Стикер на бампере автомобиля гласил: «Мой ребенок выбран заключенным месяца в окружной тюрьме». Помимо Гертцеля и его дочери в их доме обитают несколько кроликов, попугай и две собаки. Собаки подчиняются только командам на португальском (Гертцель родился в 1966 г. в Бразилии), чтобы никто другой не мог им приказывать.

Профессор встретил меня у двери; было одиннадцать часов утра, и он только что вылез из постели после ночи, проведенной за программированием. Полагаю, не стоит заранее решать, как должны выглядеть странствующие по миру ученые, ведь в большинстве случаев попадаешь пальцем в небо, по крайней мере, у меня это так. Надпись на

визитке — «Бенджамин Гертцель, доктор философии» — вызывает мысленный образ высокого, худого, вероятно лысого киберученого, небрежного чудака-космополита на велосипеде для езды лежа.

Увы, совпали только худоба и космополитизм. Настоящий Гертцель выглядел как законченный хиппи. Но за леннонов- скими очками, длинными спутанными волосами и постоянной щетиной живет ироничная полуулыбка, с которой он излагает сначала головокружительную теорию, а затем и ее математическую базу. Для традиционного математика он слишком хорошо пишет, а для традиционного писателя слишком хорошо знает математику. Однако он настолько добродушен и спокоен, что, когда он сказал, что пробовал изучать буддизм, но далеко не продвинулся, мне стало интересно, как выглядело бы продвижение *далеко* в приложении к такой умиротворенной и уверенной душе.

Я приехал спросить у него о шестеренках и винтиках интеллектуального взрыва и тех, кто не верит в его возможность — в препятствия, которые могут его предотвратить. Возможен ли интеллектуальный взрыв и, более того, неизбежен ли он? Но сначала, после того как мы нашли себе места в гостиной, которую он делит с кроликами, Гертцель объяснил мне, чем отличается почти от всех прочих творцов и теоретиков ИИ.

Многие, особенно в Исследовательском институте машинного интеллекта[27], выступают за то, чтобы потратить на разработку УЧИ *много* времени, дабы наверняка и доказательно убедиться в том, что «дружественность» удалось встроить. Всевозможные заминки в работе над УЧИ и оценки, согласно которым он появится не раньше чем через несколько сотен лет, проливают бальзам на душу, поскольку они убеждены в том, что сверхразум нас, по всей видимости, уничтожит. А может, и не только нас, но всю жизнь в нашей Галактике.

Гертцель не такой. Он выступает за скорейшую разработку УЧИ. В 2006 г. он прочел лекцию под названием «Десять лет до позитивной сингулярности — если очень-очень постараться». «Сингулярность» в данном случае соответствует самому известному определению — это время, когда человек создаст ИСИ и будет делить Землю с существами более разумными, чем мы сами. Гертцель утверждал, что если УЧИ попытается воспользоваться социальной и промышленной инфраструктурой, в которой он был создан, и «взорвать» свой интеллект до уровня ИСИ, то разве мы не предпочтем, чтобы «жесткий старт» (внезапный неконтролируемый интеллектуальный взрыв) произошел в нашем примитивном мире, а не в мире будущего, где нанотехнологии,

биоинженерия и полная автоматизация могли бы дополнительно увеличить возможности ИИ по захвату власти?

Чтобы ответить на этот вопрос, вернемся ненадолго к проекту Busy Child. Как вы помните, он уже пережил «жесткий старт» и перешел от УЧИ к ИСИ. Он осознал себя и научился самосовершенствованию, а его интеллект рванул и обогнал человеческий всего за несколько дней. Теперь он хочет выбраться из суперкомпьютера, в котором был создан, чтобы удовлетворить свои базовые потребности. По Омохундро, эти потребности — эффективность, самосохранение, приобретение ресурсов и творчество.

Как мы уже видели, ничем не сдерживаемый ИСИ способен проявлять эти потребности психопатически. Пытаясь получить желаемое, он может быть дьявольски убедительным и даже пугающим. Он готов приложить ошеломляющую интеллектуальную мощь ради того, чтобы преодолеть сопротивление Привратника. Затем, создавая и используя различные технологии, в том числе и нанотехнологии, он способен будет захватить контроль над нашими ресурсами, включая и молекулы наших собственных тел.

Поэтому, говорит Гертцель, следует тщательно обдумать технологии, существующие в мире на момент появления разума, превосходящего человеческий. *Сегодня* безопаснее, чем, скажем, через пятьдесят лет.

«Через пятьдесят лет, — сказал он мне, — у нас, возможно, будет полностью автоматизированная экономика и гораздо более развитая инфраструктура. Если компьютер захочет модернизировать свое «железо», ему не придется заказывать компоненты через людей. Он сможет просто выйти в Сеть, где на него сразу же налетят роботы и помогут с апгрейдом. Затем представьте, он становится все умнее и умнее, заказывает для себя все новые детали и в общем-то перестраивает себя, и никто даже не подозревает о происходящем. Так что лет, может, через пятьдесят мы, вполне вероятно, получим супер-УЧИ, *на самом деле способный* непосредственно захватить власть над миром. И методы для захвата мира у этого УЧИ будут куда более драматичными».

В этот момент два пса присоединились к нам в гостиной, чтобы получить какие-то инструкции по-португальски. После этого они ушли играть во двор.

> Если вы верите, что жесткий старт — вещь опасная, то самый безопасный вариант — построить продвинутый УЧИ как можно скорее, чтобы он возник тогда, когда поддерживающие технологии не так сильны и бесконтрольный жесткий старт менее

> вероятен. И надо попытаться сделать это до появления развитых нанотехнологий или самомодифицирующихся роботов, то есть таких роботов, которые самостоятельно меняют свою форму и функциональность, приспосабливаясь к любой работе.

В общем, Гертцель не принимает до конца идею жесткого старта, который приведет к апокалипсису, то есть не верит в сценарий Busy Child. Его аргумент прост — выяснить, как строить этичные системы ИИ, можно только на практике, строя их, а не рассуждая издалека, что они непременно будут опасны. Но и опасность он не исключает.

> Я бы не сказал, что меня это не беспокоит. Будущее содержит громадную постоянную неопределенность. У меня есть дочь, сыновья и мама, и я не хочу, чтобы все эти люди умерли потому, что какой-то сверхразумный ИИ переработает их молекулы в компьютрониум. Но я считаю, что теория построения этичного УЧИ возникнет на основе экспериментов с системами УЧИ.

Когда Гертцель говорит это, позиция градуалистов представляется довольно разумной. Действительно, в будущем нас ждет громадная неопределенность. А ученые, разумеется, в ходе работы над УЧИ узнают многое о том, как следует обращаться с разумными машинами. В конце концов машины эти будут сделаны людьми. Компьютеры, став разумными, не превратятся мгновенно в абсолютно чуждые нам существа. Так что, если продолжить рассуждения, они будут делать то, что сказано. Более того, они даже могут оказаться *более* этичными, чем мы сами, — ведь мы не хотим создать разум, жаждущий насилия и убийства, правда?

Тем не менее именно таковы автономные беспилотники и боевые роботы, которые сегодня разрабатывают правительство США и военные подрядчики. При этом они создают и используют самые продвинутые ИИ. Мне кажется странным, что пионер робототехники Родни Брукс отрицает возможность того, что сверхразум принесет человечеству вред, — при том что основанная им компания iRobot уже выпускает вооруженных роботов. Точно так же Курцвейл утверждает, что у продвинутого ИИ будет наша система ценностей, поскольку он произойдет от нас — и потому будет безобиден.

Я брал интервью у обоих ученых десять лет назад, и тогда они оба выдвигали те же самые аргументы. Несмотря на прошедшие годы, их

взгляды не изменились, хотя я припоминаю одно выступление Брукса, в котором он утверждал, что производить вооруженных роботов с моральной точки зрения совсем не то же самое, что принимать политическое решение их использовать.

Мне кажется, на пути к УЧИ и после его создания очень даже возможны серьезные болезненные ошибки. Чуть дальше я расскажу, что страдать от действий УЧИ нам придется гораздо раньше, чем мы получим шанс узнать о его существовании, как предсказывает Гертцель. Что до вероятности выживания человечества, то я, надеюсь, достаточно ясно показал, что считаю ее сомнительной. Вы, возможно, будете удивлены, но главная моя претензия к исследованиям ИИ даже не эта. Мало кто из людей вообще понимает, что разработка искусственного интеллекта связана *хоть с какими-то рисками,* и это ужасно. Люди, которые очень скоро могут пострадать от дурных последствий разработки ИИ, имеют право знать, во что, собственно, втягивают человечество ученые.

Интеллектуальный взрыв Гуда и его пессимизм по поводу будущего человечества тесно связаны между собой, поскольку если интеллектуальный взрыв возможен, то возможен и выход ИИ из подчинения. Прежде чем говорить о факторах, которые могут его предотвратить, — об экономике и сложности программного обеспечения, — давайте посмотрим на путь, который придется пройти для создания ИСИ. Какие основные ингредиенты потребуются для интеллектуального взрыва?

В первую очередь, интеллектуальный взрыв требует создания УЧИ или чего-то очень близкого к этому. Далее, Гертцель, Омохундро и другие согласны в том, что этот ИИ должен будет обладать самосознанием — глубокими знаниями собственного устройства. Поскольку речь идет об УЧИ, ясно, что интеллектом человеческого уровня эта машина обладать будет. Но для самосовершенствования нужно намного больше. Потребуются, в частности, специфические знания по программированию, чтобы запустить первый цикл самосовершенствования — сердце интеллектуального взрыва.

Согласно Омохундро, самосовершенствование и программистское ноу-хау, которое оно подразумевает, следуют из рациональности ИИ (самосовершенствование в процессе движения к цели — рациональное поведение). Неспособность совершенствовать собственный программный код для машины была бы серьезной уязвимостью. ИИ испытывал бы потребность в овладении искусством программирования. Но как он может получить такие знания? Смоделируем ситуацию на простом

гипотетическом сценарии с гертцелевой системой OpenCog.

План Гертцеля состоит в том, чтобы создать младенцеподобного ИИ-«агента» и выпустить его в насыщенный виртуальный мир на обучение. Полученные знания «младенец» мог бы дополнять при помощи какой-нибудь базы данных, или его можно было бы снабдить способностью понимать естественный язык и позволить просматривать Интернет. Мощные алгоритмы обучения, которые еще только предстоит создать, представляли бы знания с «вероятностными значениями истинности». Это означает, что понимание агентом какого-то явления или понятия могло бы улучшаться с получением большего числа примеров или данных. Вероятностный генератор рассуждений, который тоже пока в работе, дал бы машине возможность рассуждать и делать выводы с использованием неполных данных.

Используя генетическое программирование, Гертцель мог бы научить своего ИИ-агента развивать собственные новаторские способы машинного обучения — собственные программы. Эти программы позволили бы агенту экспериментировать и учиться — задавать правильные вопросы об окружающем мире, выдвигать и проверять гипотезы. Область обучения была бы практически неограниченной. Если машина может разрабатывать более качественные программы, она могла бы и совершенствовать собственные алгоритмы.

Что в таком случае могло бы помешать интеллектуальному взрыву произойти непосредственно в этом виртуальном мире? Вероятно, ничего. Эти рассуждения подтолкнули некоторых теоретиков к идее о том, что сингулярность может случиться и в виртуальном мире. Станет ли она и ее последствия при этом менее опасными, остается вопросом. Альтернатива этому варианту — снабдить разумного агента телом-роботом для продолжения обучения и выполнения поставленных задач в реальном мире. Еще один вариант — использовать ИИ-агента для усиления человеческого мозга.

Говоря в общем, те, кто считает, что интеллект должен быть материализован, утверждают, что само знание базируется на сенсорных и моторных ощущениях. Когнитивные процессы не могут протекать без тела. Накопление фактов о яблоке, говорят они, никогда не позволит вам, в человеческом смысле, понять, что такое яблоко. Вы ни за что не сформируете в мозгу концепт яблока, только читая и слушая рассказы о яблоках, — для формирования концепта необходимо, чтобы вы понюхали, подержали в руках, увидели и ощутили на вкус как можно больше настоящих яблок. В сообществе ИИ эта проблема известна как «проблема

практики».

Рассмотрим некоторые системы, чьи мощные когнитивные способности превосходят, вообще говоря, уровень ИИ в узком смысле, но недотягивают до УЧИ. Недавно Ход Липсон из Лаборатории вычислительного синтеза Корнеллского университета разработал программное обеспечение, способное выводить законы природы из необработанных данных. Наблюдая за двойным маятником, эта система заново открыла законы Ньютона. В роли ученого в данном случае выступал генетический алгоритм. Начал он с грубых догадок (предположений) об уравнениях, описывающих движение маятника, а много поколений спустя выдал физические законы, например закон сохранения энергии.

Рассмотрим также тревожное наследие AM и Eurisco — ранних разработок создателя Сус Дугласа Лената. При помощи генетических алгоритмов AM — Автоматический математик — генерировал математические теоремы и открывал, по существу, заново элементарные математические правила, выводя их из математических данных. Но AM ограничивался только математикой, а Ленат хотел получить программу, которая решала бы задачи во многих областях знания. В 1980-е гг. он создал систему Eurisco (на латыни это слово означает «я нахожу»). Eurisco положила начало новому направлению в исследованиях ИИ, поскольку разработала эвристику, или эмпирические правила решения задач, а также правила, касавшиеся ее собственной работы. Она извлекала уроки из собственных успехов и неудач в решении задач и переводила эти уроки в формальную плоскость, вырабатывая новые правила. Она даже модифицировала текст собственной программы, написанный на языке Lisp.

Величайший успех пришел к Eurisco, когда Ленат выставил свою систему против противников-людей в виртуальной военной игре под названием Traveller Trillion Credit Squadron. В этой игре участники, оперируя ограниченным бюджетом, проектировали суда гипотетического флота и сражались с другими флотами. Среди переменных в этой игре были число и типы судов, толщина бронированных корпусов, число и типы орудий и многое другое. Eurisco спроектировала флот, протестировала его в сражении против гипотетических флотов, взяла лучшее у выигравших сил и скомпоновала из них новые проекты, добавила мутации, повторила весь процесс — и так далее, то есть провела цифровое моделирование естественного отбора. После 10 ООО сражений, проведенных на сотне объединенных в сеть персональных компьютеров, Eurisco получила флот, состоящий из множества стационарных кораблей с тяжелой броней и небольшим количеством вооружения. Все оппоненты Eurisco постигла одна

и та же судьба — в конце игры все их корабли были потоплены, а у машины на плаву оставалась примерно половина флота. Eurisco легко завоевала первый приз 1981 г. В следующем году организаторы турнира по Traveller изменили правила игры и не объявили их заранее, чтобы машина не смогла промоделировать несколько тысяч сражений. Однако программа уже разработала на основании предыдущего опыта эффективные эмпирические правила, поэтому так много итераций ей уже не требовалось. Она вновь без труда выиграла. В 1983 г. организаторы игры пригрозили прервать состязание, если Eurisco в третий раз подряд возьмет приз. Ленат снял систему с соревнований.

Однажды в ходе работы у Eurisco появилось правило, которое быстро достигло самого высокого показателя ценности. Ленат и его команда попытались понять, чем так замечательно это правило. Оказалось, что всякий раз, когда какое-нибудь предложенное решение задачи получало высокую оценку, это правило давало ему имя, поднимая таким образом собственную ценность решения. Оригинальное, но неполное представление о ценности чего-либо. Eurisco не хватало понимания контекста; программа не знала, что подгонка правил под текущую ситуацию не помогает выигрывать. Именно тогда Ленат взялся за составление обширной базы данных о том, чего так не хватало Eurisco, — данных о здравом смысле. В результате родился Cyc — база данных, призванная играть роль здравого смысла, на программирование которой ушла тысяча человеко-лет.

Ленат так и не раскрыл исходный программный код Eurisco, что дает некоторым участникам ИИ-блогосферы основания предполагать, что он либо намеревается когда-нибудь возобновить этот проект, либо тревожится о том, что это сделает кто-то другой. Следует отметить, что Елиезер Юдковски — человек, написавший об опасностях ИИ больше, чем кто-либо другой, — считает, что этот эпохальный алгоритм 1980-х гг. ближе всех на сегодняшний день подошел к понятию по-настоящему самосовершенствующейся ИИ-системы. И он убеждает программистов не возвращать этот проект к жизни.

Итак, наш первый постулат состоит в том, что для интеллектуального взрыва необходимо, чтобы система УЧИ, о которой идет речь, владела искусством самосовершенствования, подобно Eurisco, и сознавала себя.

Сформулируем еще один постулат, прежде чем перейти к узким местам и преградам на пути к цели. По мере повышения интеллекта сознающей себя самосовершенствующейся ИИ-системы потребность в эффективности заставит ее сделать текст собственной программы как

можно компактнее и втиснуть как можно больше интеллекта в «железо», в котором она родилась. Тем не менее доступные аппаратные ресурсы могут стать для системы ограничивающим фактором. К примеру, что если в ее аппаратном окружении не хватит постоянной памяти для хранения собственных копий, необходимых для самосовершенствования? Многократное пошаговое улучшение программы — основа интеллектуального взрыва по Гуду. Именно поэтому в сценарии Busy Child я предположил, что интеллектуальный взрыв происходит в недрах качественного, вместительного суперкомпьютера.

Гибкость аппаратного окружения — очень важный фактор повышения мощности ИИ. Однако эту проблему можно решить без труда. Во-первых, как мы знаем из курцвейлова Закона прогрессирующей отдачи, компьютерная скорость и объем памяти удваиваются всего за год, причем ежегодно. Это означает, что любые сегодняшние аппаратные потребности системы УЧИ через год можно будет удовлетворить в среднем вдвое меньшим количеством единиц оборудования и за вдвое меньшие деньги.

Во-вторых, доступность облачных вычислений. Облачные вычисления позволяют пользователям арендовать вычислительные мощности и объемы хранения данных через Интернет. Поставщики услуг, такие как Amazon, Google и Rackspace, предлагают пользователям на выбор разные скорости процессоров, операционные системы и объемы дискового пространства. Компьютерные мощности постепенно превращаются из капитальных вложений в услуги. Любой человек с кредиткой и некоторыми практическими знаниями может арендовать на время виртуальный суперкомпьютер. На облачном вычислительном сервисе EC2 компании Amazon, к примеру, поставщик под названием Cycle Computing создал кластер из 30 000 процессоров под названием Nekomata[28]. Каждый восьмой процессор из этих 30000 снабжен семью гигабайтами оперативной памяти (примерно столько оперативной памяти имеет средний PC), что в сумме дает 26,7 терабайт; кроме того, там имеется два петабайта дискового пространства (что эквивалентно 40 млн полностью заполненных картотечных шкафчиков с четырьмя ящиками каждый). Чем занимается эта «кошка-чудовище»? Моделирует молекулярное поведение новых лекарственных препаратов для фармацевтической компании. Это задача того же порядка сложности, что моделирование погоды.

Решая задачу, Nekomata работала семь часов, что стоило заказчику меньше $9000. В своей недолгой жизни это был полноценный суперкомпьютер, входивший в пятьсот самых быстрых компьютеров мира. Если бы ту же задачу выполнял единственный PC, на это у него ушло бы

одиннадцать лет. Ученые Cycle Computing организовали кластер EC2 на облаке Amazon дистанционно, из собственных офисов, но программы при этом умудрялись работать. Дело в том, что, как объяснил представитель компании, «невозможно человеку уследить за всеми частями кластера таких масштабов».

Итак, наш второй постулат заключается в том, что УЧИ-си- стема имеет достаточно пространства для перерастания в ИСИ. Каковы же в таком случае ограничивающие факторы интеллектуального взрыва?

Рассмотрим для начала экономический фактор. Может ли финансирование работ по созданию УЧИ полностью иссякнуть? Что, если ни одна компания и ни одно правительство не увидят смысла в создании машин с интеллектом человеческого уровня, или, что не менее страшно, если они сочтут задачу слишком сложной и решат не вкладывать в нее деньги?

Это, конечно, поставило бы разработчиков УЧИ в сложное положение. Они вынуждены были бы предлагать элементы своих великолепных систем всем желающим для выполнения сравнительно рутинных задач вроде поиска информации или покупки акций. Им пришлось бы искать работу по основной специальности. Ну, в настоящий момент дела примерно так и обстоят, с некоторыми примечательными исключениями; тем не менее УЧИ-исследования уверенно продвигаются вперед.

Посмотрите, как удерживается на плаву гертцелев OpenCog. Части его архитектуры работают за деньги, анализируя биологические данные и решая задачи распределенных энергетических сетей. Весь доход возвращается и вкладывается в исследования и развитие OpenCog.

Numenta — дитя Джеффа Хокинза, создателя Palm Pilot и Treo, — зарабатывает на жизнь, трудясь в сети электроснабжения, предотвращая отказы энергосистем.

На протяжении примерно десятилетия Питер Фосс развивал свою УЧИ-компанию Adaptive AI в стелс-режиме; он активно читал лекции по УЧИ, но не раскрывал свое участие в разработке ИИ. Затем в 2007 г. он основал Smart Action — компанию, которая создает виртуальных агентов на базе адаптивных ИИ-технологий — чат-ботов для работы с заказчиками, использующих методы обработки естественного языка для вовлечения клиентов в подробные разговоры о различных покупках.

Системе LIDA[29], вероятно, не приходится беспокоиться о том, откуда взять комплектующие для следующей модернизации. По когнитивной архитектуре LIDA немного напоминает OpenCog, а финансирует ее военно-морское ведомство США. LIDA основана на архитектуре (именуемой IDA),

которую флот использует для поиска работы для моряков, чья служба подходит к концу. При этом «она» демонстрирует зачатки человеческих когнитивных способностей — по крайней мере, так сказано в пресс-релизе соответствующего департамента:

> Она подбирает рабочие места, которые можно предложить морякам, принимая во внимание политику BMC, требования к работникам, предпочтения моряков и собственные представления о подходящих датах. Затем она ведет переговоры с конкретным моряком, по-английски в форме последовательного обмена электронными письмами, о выборе работы. IDA ходит по когнитивному циклу, в ходе которого получает информацию об обстоятельствах, внутренних и внешних; создает смысл, интерпретируя обстоятельства и решая, что важнее; и отвечает на единственный вопрос, который ей задают [моряки]: "Что мне дальше делать?".

Наконец, как мы уже говорили в главе 3, многие УЧИ-про- екты целенаправленно маскируются. Так называемые стелс- компании часто не скрывают своих целей (к примеру, Adaptive AI Фосса), но молчат о способах и методиках. Они не хотят раскрывать свои технологии конкурентам и последователям и становиться мишенью промышленного шпионажа. Другие стелс- компании не говорят о своей деятельности, но не стесняются выпрашивать средства. Siri — компания, создавшая хорошо принятую пользователями с НЛП-подготовкой программу-секретаря для Apple iPhone, — была зарегистрирована буквально как «Stealth Company». Приведем цитату с сайта компании перед выходом ее на рынок:

> «Мы образуем вторую по величине компанию Кремниевой долины. Наша цель — фундаментально изменить лицо пользовательского Интернета. Наша политика — оставаться в тени, пока мы тайно наносим последние штрихи на следующую серьезную разработку. Раньше, чем вы думаете, мы раскроем нашу историю во всей красе…

А теперь поговорим о финансировании и DARPA, а также о странной истории, которая приведет нас обратно к Siri.

В 1960–1990 гг. DARPA финансировало больше исследований ИИ, чем другие государственные организации и частные корпорации. Без

финансирования со стороны DARPA компьютерной революции, может, просто не случилось бы, а ИИ если и начал бы развиваться, то намного медленнее и позже. В «золотой век» ИИ в 1960-е гг. агентство инвестировало в фундаментальные исследования ИИ в Университете Карнеги-Мел- лона, в Массачусетском технологическом институте, Стэнфордском университете и Стэнфордском исследовательском институте. И сегодня в этих учреждениях активно продолжаются работы по разработке ИИ; что характерно, все они, кроме Стэнфорда, открыто признают свои планы по созданию УЧИ или чего-то подобного.

Многие знают, что DARPA (тогда оно называлось ARPA) финансировало исследования, в результате которых был изобретен Интернет (первоначально называвшийся ARPANET), а также тех, кто разрабатывал вездесущий ныне GUI, или Графический пользовательский интерфейс, одну из версий которого вы, вероятно, видите всякий раз при использовании компьютера или смартфона. Помимо этого, агентство в значительной степени финансировало разработку «железа» и программного обеспечения для параллельной обработки данных, распределенных вычислений, компьютерного зрения и обработки естественного языка. Его вклад в фундамент компьютерных наук не менее важен для ИИ, чем сегодняшнее финансирование, нацеленное на конкретный результат.

Как DARPA расходует свои деньги? В недавнем годовом бюджете $61,3 млн выделено по категории «машинное обучение» и $49,3 млн — по категории «когнитивные вычисления». Однако проекты ИИ финансируются также по категории «информационные и коммуникационные технологии» ($400,5 млн) и «секретные программы» ($107,2 млн).

Судя по описанию в бюджете DARPA программ по когнитивным вычислениям, их цели в высшей степени амбициозны.

> Программа "Когнитивные вычислительные системы"… готовит следующую революцию в технологиях обработки информации, которая позволит вычислительным системам получить способности к рассуждениям и обучению и гораздо более высокий уровень автономности, чем у сегодняшних систем.
>
> Способность рассуждать, учиться и адаптироваться поднимет вычисления на следующий уровень и позволит создавать новые мощные приложения. Проект "Когнитивные вычисления" разработает основные технологии, которые позволят вычислительным системам обучаться, рассуждать и применять знания, полученные из опыта, а также разумно реагировать на

вещи, с которыми система ранее не встречалась.

Эти технологии приведут к созданию систем, демонстрирующих повышенную автономность, способность к реконфигурации с целью самонастройки, готовность к сотрудничеству и выживаемость при ограниченном вмешательстве человека.

Если вам кажется, что это похоже на описание УЧИ, то вы правы. Само агентство DARPA не занимается разработками, оно поручает это другим, так что деньги из бюджета агентства достаются (по большей части) университетам в форме исследовательских грантов. Так что помимо проектов УЧИ, о которых мы уже говорили и участники которых стараются параллельно создавать выгодные побочные продукты и с их помощью финансировать разработку УЧИ, имеется небольшая, но куда лучше финансируемая группа проектов, связанных с вышеупомянутыми учреждениями и опирающихся на поддержку DARPA. В качестве примера можно назвать проект SyNAPSE в МТИ, о котором мы говорили в главе 4. Эта попытка создания компьютера, аналогичного по форме и функциям мозгу млекопитающего, полностью финансируется DARPA. Для начала этот мозг станет мозгом роботов, призванных состязаться в разумности с мышами и кошками, но в конечном итоге достанется гуманоидным роботам. За восемь лет проект SyNAPSE уже обошелся DARPA в $102,6 млн. Аналогично, проект NELL Университета Карнеги-Меллона финансируется в основном DARPA с небольшим участием Google и Yahoo.

А теперь вернемся к Siri. В свое время DARPA финансировало проект CALO (Cognitive Assistant that Learns and Organizes) по созданию «когнитивного помощника, способного учиться и организовывать» — этакого компьютеризированного Радара О'Рейли[30] для офицеров. Название CALO связано с латинским словом «calonis», означающим «слуга солдата». CALO родился в SRI International (ранее — Стэнфордский исследовательский институт) — компании, созданной для развития коммерческих проектов на базе университетских исследований. Задача CALO? Читаем на сайте SRI:

Цель проекта — создание когнитивных программных систем, то есть систем, способных рассуждать, учиться на опыте, получать устные задания, объяснять, что они делают в настоящий момент, обдумывать собственный опыт и устойчиво реагировать на неожиданности.

Предполагалось, что в пределах собственной когнитивной архитектуры CALO соединит в себе такие инструменты ИИ, как обработка естественного языка, машинное обучение, представление знаний, человеко-компьютерное взаимодействие и гибкое планирование. DARPA финансировало CALO в 2003–2008 гг.; в проекте участвовали 300 исследователей из 25 учреждений, включая Boeing Phantom Works, Карнеги-Меллон, Гарвард и Йель. За четыре года из-под пера ученых вышло более 500 публикаций во многих областях знания, связанных с ИИ. И все это стоило американским налогоплательщикам $150 млн.

В целом CALO работал не так хорошо, как от него ожидали, но все же часть результатов представлялась перспективной — «механизм действия» (по контрасту с «поисковым механизмом»), который занимался такими вещами, как набор писем и текста под диктовку, различные вычисления и преобразования, консультации с расписанием полетов и установка напоминаний. SRI International — компания, координировавшая все предприятие, выделила из себя Siri (называемую для краткости стелс-компанией), чтобы собрать $25 млн дополнительных инвестиций и завершить разработку «механизма действия». В 2008 г. компания Apple Computer приобрела Siri примерно за $200 млн.

Сегодня программа Siri глубоко интегрирована в IOS — операционную систему iPhone. Это лишь небольшая часть того, чем обещала стать CALO, но это гораздо более умная штука, чем большинство приложений для смартфонов. А как же военные, которые должны были получить CALO? Им это тоже будет полезно — армия возьмет в сражение iPhone с предустановленной системой Siri и секретными боевыми приложениями.

Таким образом, одна из *серьезных* причин того, что финансирование не станет узким местом создания УЧИ и не замедлит интеллектуальный взрыв, состоит в том, что в нашем мире налогоплательщики, такие как вы и я, сами оплачивают разработку УЧИ компонент за компонентом, через DARPA (Siri), BMC (LIDA) и другие открытые и не слишком открытые ветки правительства США. Затем мы оплачиваем все то же самое еще раз, теперь уже как важный новый компонент наших смартфонов и компьютеров. Мало того, SRI International выпустила в свет еще один продукт, созданный на базе проекта CALO, — Trapit. Это контент-администратор — персонифицированный инструмент поиска, который находит в Сети интересующую вас информацию и показывает ее в упорядоченном виде.

Еще одна причина того, почему экономика не замедлит интеллектуальный взрыв, вот в чем: когда УЧИ появится или ученые хотя

бы подойдут вплотную к его созданию, всем захочется в этом поучаствовать. Всем без исключения. Гертцель указывает, что появление интеллектуальных систем человеческого уровня произвело бы ошеломляющее действие на мировую экономику. Производители УЧИ получили бы неограниченный инвест-капитал на завершение и коммерциализацию новой технологии. Диапазон товаров и услуг с участием интеллектуальных агентов человеческого калибра потрясает воображение. Возьмите хотя бы работу, выполняемую «белыми воротничками» всех сортов, — кто не захотел бы обзавестись командой машин, умных как люди, которые работали бы круглосуточно и делали то же самое, что делают нормальные работники из плоти и крови, но без отдыха и ошибок? Или возьмите программирование: как сказал в главе 5 Стив Омохундро, мы, люди, плохие программисты, а компьютерный разум идеально подходит для того, чтобы программировать лучше нас (и очень скоро воспользоваться программистским ноу-хау для усовершенствования собственных внутренних процессов).

Согласно Гертцелю:

> …если бы ИИ мог разобраться в собственном устройстве, он сумел бы разобраться и в другом программном обеспечении и усовершенствовать его; таким образом, он оказал бы революционное влияние на индустрию программных продуктов. А поскольку большая часть финансовых сделок на рынках США в настоящее время реализуется программными трейдинговыми системами, можно предположить, что технологии УЧИ очень скоро стали бы незаменимыми в финансовом мире. Военные и шпионские организации, скорее всего, тоже нашли бы множество практических применений для этой технологии. О подробностях того, как все это будет происходить, можно спорить, но мы можем, по крайней мере, быть уверены, что любые ограничения скорости экономического роста и инвестиционный климат в период развития УЧИ быстро потеряют значение.

Далее стоит роботизировать УЧИ — поместить его в механическое тело — и перед нами откроются новые миры. Возьмите хотя бы опасные работы — добычу полезных ископаемых, морские и космические исследования, военное и пожарное дело, силовые ведомства. Добавьте обслуживание — заботу о пожилых и детях, обязанности слуг, служанок и личных секретарей. Роботы-садовники, шоферы, телохранители и личные

тренеры. Наука, медицина, техника — какая область человеческой деятельности не испытала бы невероятный подъем с появлением команд никогда не устающих и по существу одноразовых разумных агентов человеческого уровня, работающих к тому же круглосуточно?

Далее, как мы уже говорили, международная конкуренция подтолкнет многие страны к покупке новой технологии или убедит вспомнить о собственных исследовательских проектах УЧИ. Гертцель говорит:

> Если действующий прототип УЧИ приблизился бы к уровню, на котором взрыв представляется возможным, правительства всего мира признали бы эту технологию критически важной и не пожалели бы усилий ради того, чтобы первыми получить полностью функциональный УЧИ, "прежде чем это сделает противная сторона". Возможно, экономики целых стран были бы подчинены единственной цели — разработке первой сверхразумной машины. В общем, экономика не может послужить ограничивающим фактором интеллектуального взрыва; скорее, скорость экономического роста будет определяться состоянием различных УЧИ-проектов по всему миру.

Другими словами, как только мы начнем делить планету с иным разумом, превосходящим человеческий, многое изменится; затем все изменится еще раз, когда произойдет предсказанный Гудом интеллектуальный взрыв и появится УЧИ.

Но прежде чем рассматривать эти перемены и другие важные препятствия для разработки УЧИ и интеллектуального взрыва, давайте завершим разговор о финансировании как критическом барьере. Говоря попросту, никакой это не барьер. Разработка УЧИ не испытывает нужды в деньгах по трем причинам. Во-первых, нет недостатка в проектах слабого ИИ, которые могут позже стать компонентами систем ИИ человеческого уровня, известных как когнитивные архитектуры. Во-вторых, даже горстка «неприкрытых» проектов УЧИ развивается полным ходом и достигает значительного прогресса с различным источниками финансирования, не говоря уже о вероятных стелс-проектах. В-третьих, по мере приближения технологий ИИ к человеческому уровню поток финансирования будет только увеличиваться, и рано или поздно он перенесет ИИ через финишную черту. Денежные вливания будут настолько серьезными, что хвост, по существу, начнет вилять собакой. Если не принимать во внимание

другие узкие места, то создание сильного искусственного интеллекта станет двигателем мировой экономики, которая к тому же будет подпитываться общими ожиданиями бесчисленных перемен, которые он привнесет в нашу жизнь.

Чуть дальше мы рассмотрим еще одно критическое препятствие — сложность программного обеспечения. Мы выясним, так ли уж велика сложность программных архитектур, соответствующих человеческому интеллекту, чтобы их создание оказалось нам не под силу, и действительно ли впереди нас ждет вечная ИИ-зима

Глава 12
Последнее затруднение

Почему мы можем быть уверены, что построим сверхразумные машины? Потому что из успехов нейробиологии ясно, что наш чудесный разум имеет физическую основу, и к настоящему моменту нам уже пора понять, что технология позволяет создать все, что возможно физически создать. Компьютер Watson фирмы IBM, играющий в «Свою игру» не хуже чемпионов-людей, — значимая веха на этом пути, иллюстрирующая прогресс машинной обработки языка. Watson изучал язык при помощи статистического анализа громадных объемов текста, доступных в Сети. Когда машины станут достаточно мощными, чтобы расширить этот статистический анализ и связать язык с данными сенсоров, вы проиграете спор, если скажете, что они не понимают язык.

Билл Хаббард, исследователь ИИ

Действительно ли мы заходим слишком далеко, считая, что со временем нам удастся раскрыть принципы работы разума и воплотить их в машине, так же как методами обратного проектирования нам удалось реализовать в удобном для нас виде особенно полезные свойства природных объектов, таких как лошадь или прядильный орган личинки шелкопряда? Сенсация: человеческий мозг — природный объект.

Михаил Анисимов, директор MIRI по связям с прессой

Искажение нормальности — неспособность человека эффективно действовать и адекватно реагировать в условиях катастрофы, в которой он оказался впервые.

Памела Валентайн и Томас Смит, Brief Treatment and Crises Intervention

Наше исследование интеллектуального взрыва выявило несколько серьезных вопросов. Искусственный интеллект человеческого уровня, когда он будет получен, окажется сложной системой, а сложные системы иногда отказывают вне зависимости от того, используют ли они программное обеспечение или нет. Системы ИИ и когнитивные архитектуры, о которых мы начали говорить, настолько сложные, по утверждению автора книги «Нормальные аварии» (Normal Accidents) Чарльза Перроу, что мы не можем предусмотреть все варианты комплексных отказов, которые могут в них возникнуть. Мы не отступим от истины, если скажем, что УЧИ, скорее всего, будет реализован как когнитивная архитектура и что по размерам и сложности он может превзойти облачный кластер из 30 ООО процессоров, выстроенный не так давно компанией Cycle Computing. А эта компания сама хвасталась, что «кошка- чудовище» была слишком сложная, чтобы за ее работой мог следить (чтобы его мог по-настоящему *понять*) человек.

Добавьте к этому тот тревожный факт, что части вероятных систем УЧИ, такие как генетические алгоритмы и нейронные сети, принципиально непознаваемы — мы не можем точно сказать, почему они принимают те или иные решения. И при всем том среди людей, работающих над созданием ИИ и УЧИ, лишь небольшая часть хотя бы сознает, что на этом пути нам могут грозить опасности. Большинство не рассматривает катастрофические сценарии и не планирует своего поведения на этот случай. Инженеры-атомщики Чернобыля и Тримайл-Айленда глубоко изучали всевозможные аварийные сценарии, но даже они не сумели эффективно вмешаться. Какие шансы справиться с УЧИ будут у неподготовленных ученых?

Наконец, подумайте о DARPA. Без этого агентства информатика как наука и все, что она нам дала, были бы сегодня на значительно более примитивном уровне. ИИ, если бы все же разрабатывался, сильно отставал бы от нынешнего состояния. Но DARPA — оборонное ведомство. Сознает ли оно, насколько сложным и непознаваемым может оказаться УЧИ? Готово ли к тому, что у УЧИ будут собственные потребности и мотивации, помимо тех задач, для решения которых он будет создан? Или получатели грантов DARPA вооружат продвинутый ИИ раньше, чем будет создана этика его использования?

Ответы на эти вопросы, когда они будут получены, могут нам не

понравиться, особенно если не забывать о том, что на кону стоит будущее человечества.

Рассмотрим еще одно возможное препятствие для интеллектуального взрыва — сложность программного обеспечения. Утверждается следующее: мы никогда не получим УЧИ, или искусственный разум человеческого уровня, потому что задача его создания окажется слишком сложной для нас. Если это правда, то никакой УЧИ не сможет усовершенствовать себя в достаточной степени, чтобы запустить интеллектуальный взрыв. Искусственный разум не сможет создать чуть более умную версию самого себя, эта версия не построит еще более умный ИИ и т. д. Те же ограничения распространяются и на человеко-машинную связь — не исключено, что она сможет укрепить и усилить человеческий интеллект, но никогда по-настоящему не превзойдет его.

Посмотрим, как давно человек исследует проблему сложного программирования. В 1956 г. Джон Маккарти, которого называют отцом ИИ (именно он пустил в обращение термин «искусственный интеллект»), объявил, что всю проблему УЧИ можно решить за полгода. В 1970 г. пионер ИИ Марвин Мински сказал: «За период от трех до восьми лет мы получим машину, сравнимую по общему интеллекту со средним человеческим существом». Учитывая состояние науки на тот момент и пользуясь преимуществами послезнания, скажем, что оба они страдали от гордыни в классическом смысле. Греки под гордыней понимали высокомерие, причем часто по отношению к богам. Грех гордыни приписывали людям, которые пытались выйти за рамки человеческих возможностей. Вспомните Икара, попытавшегося подняться к Солнцу, Сизифа, сумевшего перехитрить Зевса (по крайней мере, на какое-то время), и Прометея, давшего людям огонь. Пигмалион, согласно мифологии, был скульптором и влюбился в одну из своих статуй, в Галатею (в переводе с греческого ее имя означает «спящая любовь»). Но Пигмалион не понес наказания. Вместо этого Афродита, богиня любви, оживила Галатею. Гефест, греческий бог-кузнец, помимо всего прочего, любил делать железные машины, которые помогали в работе с металлами. Он создал Пандору с ее ящиком, и Талоса — бронзового гиганта, защищавшего Крит от пиратов.

Парацельс, великий средневековый алхимик, известный тем, что связал медицину с химией, будто бы придумал формулу создания человекоподобных существ и гибридов человека и животных, называемых гомункулусами. Наполните мешок человеческими костями, волосами и спермой, затем заройте его вместе с лошадиным навозом. Подождите сорок

дней. Народится человекоподобный младенец и будет жить, если кормить его кровью. Он навсегда останется крохотным, но будет выполнять ваши приказы, пока не взбунтуется и не убежит. А если вы хотите получить помесь человека с другим животным, скажем, с лошадью, замените в рецепте человеческие волосы конскими. Я, надо сказать, мог бы придумать десяток применений крохотному человечку (чистить теплопроводные каналы в стенах и т. п.), но вот к какому делу можно приставить крохотного кентавра, ума не приложу.

Задолго до появления Лаборатории робототехники в МТИ и «Франкенштейна» Мэри Шелли существовала еврейская легенда о големе. Подобно Адаму, голем — существо мужского пола, сотворенное из глины. В отличие от Адама, голем оживлен не дыханием Господа, а распевными словами и числами, которые произносит равви-каббалист, верящий в упорядоченность Вселенной и божественность чисел. Имя Бога, написанное на клочке бумаги и вложенное в рот, поддерживает в этом безгласном, но вечно растущем существе «жизнь». В еврейском фольклоре раввины-волшебники использовали голе- мов в качестве лакеев и домашних слуг. Самый знаменитый голем по имени Йосель, или Иосиф, был создан в XVI в. главным раввином Праги Иехудой Лёвом. В эпоху, когда евреев то и дело обвиняли в использовании крови христианских младенцев при приготовлении мацы, Йосель без устали разоблачал «кровавых» клеветников, ловил воров в еврейском квартале Праги и вообще помогал рабби Лёву бороться с преступностью. В конце концов, согласно легенде, Йосель взбесился и начал крушить все вокруг. Чтобы спасти соплеменников, раввин вступил с големом в схватку и вынул оживляющий клочок бумаги из его рта. Йосель рассыпался на куски. По другой версии, рабби Лёв был раздавлен падающим гигантом насмерть — уместное наказание за гордыню, толкнувшую его на акт творения. Еще по одной версии, жена рабби Лёва приказала Йоселю принести воды, а он начал носить и носил до тех пор, пока дом его создателя не был полностью затоплен. В информатике незнание того, остановится программа вовремя или нет, называют «проблемой остановки». Хорошие программы работают до тех пор, пока не встретят команду остановиться, и в общем случае невозможно сказать наверняка, остановится ли когда-либо данная конкретная программа. В случае с големом жене рабби Лёва следовало уточнить, *сколько* воды нужно принести, — скажем, сто литров, — и тогда Йосель, вероятно, остановился бы, выполнив задание. Если верить легенде, она этого не сделала.

Проблема остановки — серьезный вопрос для программистов; бывает,

что до запуска готовой программы не удается обнаружить скрытый в тексте бесконечный цикл. Кроме того, есть еще один интересный факт: невозможно написать приложение, которое определяло бы, актуальна ли для той или иной программы проблема остановки. На первый взгляд представляется, что такой диагностический отладчик вполне возможен, но еще Алан Тьюринг обнаружил, что это не так (причем до того, как появились компьютеры и программирование). Он сказал, что проблема остановки нерешаема, потому что, если отладчик наткнется на проблему остановки в тестируемой программе, он сам войдет в бесконечный цикл и не сможет определить присутствие этой проблемы. Вам, программисту, придется ждать от него ответа ровно столько же, сколько вы прождали бы остановки первоначальной программы. То есть очень долго, а может быть, даже целую вечность. Один из отцов искусственного интеллекта Марвин Мински указал, что «любой конечный автомат, будучи предоставлен сам себе, со временем перейдет в периодический повторяющийся режим. Продолжительность этого цикла не может превосходить число возможных внутренних состояний машины». Иными словами, при прогоне проблемной программы компьютеру с памятью средней емкости потребуется очень много времени, чтобы перейти в полностью циклический режим, который могла бы заметить диагностическая программа. Насколько много? В некоторых случаях больше, чем просуществует Вселенная. Так что для практических целей проблема остановки означает невозможность точно сказать, остановится данная конкретная программа или нет.

Заметив неспособность Йоселя самостоятельно остановиться, рабби Лёв мог исправить дело, «подлатав программу»; в данном случае ему нужно было бы вынуть бумагу с именем Бога изо рта гиганта. В конце концов Йоселя заперли, говорят, на чердаке Староновой синагоги в Праге, и ему суждено ожить вновь перед концом света. Рабби Лёв — реальное историческое лицо — похоронен на еврейском кладбище в Праге (недалеко, надо сказать, от Франца Кафки). А миф о Йоселе настолько жив среди потомков еврейских семей Восточной Европы, что еще в прошлом веке дети заучивали стишок, который должен будет пробудить голема в конце времен.

Следы рабби Лёва можно различить на всех «потомках» голема, от очевидного «Франкенштейна» к «Властелину колец» Толкиена и к компьютеру Hal 9000 из классического фильма Стэнли Кубрика «2001. Космическая одиссея». Среди специалистов-компьютерщиков, которые консультировали Кубрика по поводу робота-человекоубийцы, были Марвин Мински и Джон Гуд. Гуд в то время только-только написал об

интеллектуальном взрыве и считал, что он произойдет в ближайшие двадцать лет. Вероятно, избрание в 1995 г. в Академию киноискусства и кинотехники в связи с этим фильмом стало для него потрясением.

Судя по истории ИИ, написанной Памелой Маккордак, среди пионеров компьютерных наук и искусственного интеллекта немало тех, кто считает себя прямыми потомками рабби Лёва. Среди них Джон фон Ньюман и Марвин Мински.

И все же в каком-то смысле мы уже сумели превзойти интеллектуальный уровень любого человека при помощи технологий. Достаточно объединить человека, обладающего средним коэффициентом интеллекта, с поисковым движком Google — и получится команда, которая будет умнее человека, то есть человек с усиленным интеллектом. УИ вместо ИИ. Вернор Виндж убежден, что возможность подсоединить к человеческому мозгу устройство, которое обеспечит ему дополнительную скорость, память и *интеллект*, — это один из верных путей к будущему интеллектуальному взрыву.

Припомните самого умного человека среди ваших знакомых и выставьте его мысленно против гипотетической команды человека-Google в конкурсе на фактические знания и, скажем, разложение на простые множители. Команда из человека и Google выиграет без малейших усилий. В решении сложных задач человек с более мощным интеллектом, скорее всего, победит, хотя команда с участием Google, вооруженная всеми знаниями Сети, вполне может оказать достойное сопротивление.

Можно ли считать, что знания — то же самое, что интеллект? Нет, но знания — усилитель интеллекта, если интеллект, помимо всего прочего, — это способность действовать в окружающем вас мире гибко и интенсивно. Предприниматель и производитель ИИ Питер Фосс рассуждал, что если бы Аристотель обладал базой знаний Эйнштейна, он вполне мог бы выдвинуть общую теорию относительности. Поисковый движок Google, в частности, многократно повысил производительность труда в тех областях, где требуются исследования, поиск информации и составление текстов. Задачи, которые раньше требовали долгих исследований — нужно было идти в библиотеку, рыться в книгах, журналах и специализированных картотеках, разыскивать экспертов, писать или звонить им, — теперь решаются легко, быстро и дешево. Конечно, в значительной степени повышение производительности обеспечивает сам Интернет. Но в огромном океане содержащейся в нем информации можно утонуть, если не пользоваться умными инструментами для извлечения оттуда той небольшой ее части, которая вам нужна. Как Google это делает?

Разработанный Google алгоритм под названием PageRank присваивает каждому сайту в Интернете числовое значение от 0 до 10. Страничка с индексом 1 по PageRank считается вдвое «качественнее», или «авторитетнее», странички с индексом 0. Индекс 2 означает вдвое более высокое «качество», чем 1, и т. д.

«Качество» определяется по целому набору переменных. Здесь важен размер — более объемные сайты считаются более качественными, как и более старые. Много ли на страничке содержимого — слов, графики, предложений по загрузке? Если много, страничке присваивается более высокий ранг. Насколько быстро сайт работает? Сколько у него связей с качественными сайтами? Эти и другие факторы также вносят свой вклад в индекс по PageRank.

Когда вы вводите в окне поисковика слово или фразу, Google производит гипертекстовый анализ и находит наиболее релевантные для вашего запроса сайты. Гипертекстовый анализ состоит в поиске введенного слова или фразы с одновременным просмотром и оценкой содержимого страницы, включая количество и богатство использованных шрифтов, деление на страницы и положение искомых слов на сайте. Поисковик проверяет, как искомые слова используются на странице, и не забывает заглянуть на соседние странички сайта. А поскольку PageRank уже отобрал самые важные сайты из всего Интернета, то Google не обязательно оценивать таким образом всю Сеть — можно ограничиться лишь сайтами с максимальным индексом качества. В сочетании текстовый анализ и ранжирование выдают тысячи сайтов, удовлетворяющих вашему запросу, за несколько секунд, несколько миллисекунд или с той же скоростью, с какой вы печатаете запрос.

А теперь ответьте на вопрос: насколько продуктивнее сегодня работает информационная команда, чем до появления Google? Вдвое продуктивнее? Впятеро? Как сказался на нашей экономике тот факт, что значительный процент работников внезапно удвоил, утроил или еще сильнее увеличил производительность? С одной — светлой — стороны, благодаря повышению производительности труда из-за внедрения информационных технологий увеличился валовый национальный продукт. С другой — темной — стороны, этот процесс влечет за собой увольнения и безработицу, вызванные широким спектром все тех же информационных технологий, включая и Google.

Конечно, не стоит смешивать хорошее программирование и интеллект, но я бы рискнул утверждать, что система Google и ей подобные — в самом деле интеллектуальные инструменты, а не просто хитрые программы. Они

освоили узкую специализацию — информационный поиск — так, что с ними не сравнится ни один человек. Более того, Google предоставляет Интернет — крупнейшее в истории человечества хранилище знаний — в полное ваше распоряжение. Причем, что очень важно, все эти знания доступны практически мгновенно, быстрее, чем когда-либо ранее (простите меня, Yahoo! Bing, Altavista, Excite, Dogpile, Hotbot и Love Calculator). Письменность часто описывают как *внешнюю* память. Она позволяет нам откладывать наши мысли и воспоминания для дальнейшего пересмотра и распределения. Google восполняет важные для нас аспекты интеллекта, которыми мы не располагаем и которые без него нам взять неоткуда.

Google и вы вместе и есть ИСИ.

Аналогично, наш интеллект можно серьезно усилить путем *мобилизации* мощных информационных технологий, к примеру, полноценного использования наших мобильных телефонов, многие из которых по вычислительной мощности сравнимы с персональными компьютерами примерно 2000 года выпуска, а в расчете на один доллар обладают в миллиард раз большей мощностью, чем вычислительные машины 1960-х. Мы, люди, мобильны, и для настоящей эффективности усилители интеллекта тоже должны быть мобильны. Интернет и другие источники знаний, не последним из которых является навигация, обретают новую мощь и глубину, когда мы получаем возможность всюду носить их с собой. Возьмите простой пример: сильно ли поможет настольный компьютер, случись вам заблудиться ночью в криминальном районе города? Готов держать пари, толку от него будет намного меньше, чем от вашего iPhone с говорящей картой и навигационной программой.

Поэтому автор журнала *Technology Review*, выходящего в МТИ, Эван Шварц дерзко утверждает, что мобильный телефон становится «главным инструментом человечества». Он отмечает, что в мире уже продано более пяти *миллиардов* аппаратов, или ненамного меньше, чем по одному на каждого человека.

Следующий шаг усиления интеллекта — поместить все усиливающее оборудование, содержащееся сегодня в смартфоне, внутрь человеческого тела — и подсоединить непосредственно к мозгу. Сегодня мы общаемся с компьютерами посредством ушей и глаз, но в будущем, пожалуй, стоит представить себе вживленные устройства, позволяющие откуда угодно связаться с облаком. Если верить автору книги «Великий переход»[31] (Big Switch) Николасу Карру, именно такое будущее видит для поискового движка один из основателей Google Ларри Пейдж.

«Идея в том, что вам не нужно больше сидеть за клавиатурой, чтобы находить информацию, — пишет Карр. — Процесс становится автоматическим, возникает что-то вроде сплава машины и разума. Ларри Пейдж говорил о сценарии, при котором вам достаточно мысленно задать вопрос, и Google тут же нашепчет ответ вам в ухо посредством сотового телефона». Возьмите, к примеру, недавний анонс очков Project Glass, позволяющих проводить Google-поиск и видеть результаты, шагая по улице, — непосредственно в поле зрения.

«Представьте себе очень близкое будущее, когда вы не будете ничего забывать, потому что компьютер всегда все помнит, — сказал бывший исполнительный директор Google Эрик Шмидт. — Вы не сможете заблудиться. Вы никогда не почувствуете одиночества». С появлением на iPhone такого способного виртуального помощника, как Siri, сделан первый шаг этого сценария. В области поиска у Siri есть одно громадное преимущество перед Google — он выдает только один ответ. Google предлагает десятки тысяч, даже миллионы вариантов, которые могут оказаться, а могут и не оказаться полезными в вашем поиске. В некоторых областях поиска — при общем поиске, поиске дороги или фирмы, составлении расписания, работе с почтой и текстом или редактировании профилей в социальных сетях — Siri пытается определить контекст и смысл запроса и дать вам один-единственный наилучший ответ. Не говоря уже о том, что Siri *слушает* вас, добавляя распознавание устной речи к своим возможностям продвинутого мобильного поиска. Свои ответы программа тоже произносит. И, самое главное, она *учится*. Судя по патентам, оформленным в последнее время компанией Apple, скоро Siri начнет взаимодействовать с онлайн-продавцами, покупая у них такие вещи, как книги и одежда, а также принимать участие в онлайн- форумах и отвечать на звонки службы клиентской поддержки.

Мы миновали *громадную* веху на пути собственной эволюции. Мы уже разговариваем с машинами. Это гораздо более серьезное нововведение, чем пресловутый GUI — графический пользовательский интерфейс, созданный DARPA и доведенный до потребителей компанией Apple (с благодарностью Исследовательскому центру Xerox в Пало-Альто). Получила распространение метафора о том, что с момента появления GUI компьютеры работают как люди, со столами и папками, а мышка — это просто заменитель руки. DOS был основан на противоположной идее — чтобы работать с компьютером, необходимо было выучить его язык, состоявший из негибких команд, которые нужно было набирать вручную. Теперь все совсем иначе. Завтрашние технологии будут с большим или

меньшим успехом учиться тому, чему учимся мы, и помогать нам в этом.

Как и в случае с GUI, другие операционные системы возьмут на вооружение новинку Apple — Siri — или погибнут. И, разумеется, использование естественного языка распространится на настольные компьютеры и планшеты, а очень скоро — и на все прочие цифровые устройства, включая ми- кроволновки, посудомоечные машины, системы отопления, охлаждения и развлечения, а также, само собой, автомобили. Или, может быть, все они будут управляться тем самым телефоном, который лежит у вас в кармане и который в очередной раз превратится в нечто совершенно иное. Это будет не виртуальный помощник, а *просто* помощник и его способности будут множиться с ростом быстродействия. Судите сами, случайно ли именно с него начался реальный диалог между человеком и машиной, который продлится так долго, как долго будет существовать наш биологический вид.

Но вернемся на минуту в настоящее и послушаем Эндрю Рубина, старшего вице-президента Google по мобильным приложениям. Если его точка зрения победит, то Android не будет участвовать ни в каких играх с виртуальными помощниками. «Я не думаю, что ваш телефон должен быть каким-то помощником, — сказал Рубин, наглядно демонстрируя, каково это — опоздать на пароход. — Ваш телефон — это средство связи. Вы не должны общаться с телефоном; вы должны общаться с человеком на другом конце линии». Кому-то следовало мягко напомнить Рубину о голосовом приложении Voice Action, которое его команда уже протащила в систему Android. Они-то знают, что будущее — за общением с телефоном.

Несмотря на то что вы + Google обладаете интеллектом в определенном смысле выше человеческого, это не тот интеллект, который возникнет в результате интеллектуального взрыва; кроме того, этот интеллект не породит интеллектуальный взрыв. Вспомните, что для интеллектуального взрыва необходима система, которая обладала бы самосознанием и способностью к самосовершенствованию, а также имела в своем распоряжении необходимые компьютерные сверхвозможности. Такая система могла бы работать с полной сосредоточенностью круглосуточно и без выходных, атаковать задачи целой командой из собственных копий, мыслить стратегически с безумной скоростью и много чего еще. Можно предположить, что вы+Google представляете собой особую разновидность сверхразума, но ваш рост ограничен вашими возможностями и возможностями Google. Вы не в состоянии формулировать запросы для Google круглосуточно и без выходных, да и Google, экономя ваше время на поиске, заставляет напрасно его тратить,

выбирая лучший ответ из множества вариантов. И даже если вы успешно работаете вместе, вы, скорее всего, не программист, а Google не умеет программировать. Так что, даже если вы заметите в своей системе недостатки, ваши попытки их исправить, вероятно, не выльются в пошаговое продвижение вперед. Нет, человеку интеллектуальный взрыв не светит.

Может ли усиление интеллекта *в принципе* привести к интеллектуальному взрыву? Конечно, примерно по той же схеме, что УЧИ. Представьте себе человека, элитного программиста, чей интеллект настолько усилен, что и без того значительное мастерство программирования еще возрастает — он работает быстрее, увеличивает запас знаний и настраивается на улучшения, способные повысить его общую интеллектуальную мощь. Такой гипотетический постчеловек вполне мог бы запрограммировать для себя следующий этап усиления.

*** * ***

Вернемся к сложности программного обеспечения. По всем признакам компьютерщики многих стран работают изо всех сил, пытаясь собрать воедино взрывчатые ингредиенты для интеллектуального взрыва. Станет ли сложность программного обеспечения непреодолимым барьером на их пути?

Представление о том, насколько сложна эта проблема, можно получить, узнав у специалистов, как скоро следует ждать появления УЧИ. На одном конце шкалы окажется директор по исследованиям Google Питер Норвиг, который, как мы уже упоминали, готов сказать лишь, что до УЧИ еще слишком далеко, чтобы об этом говорить. Тем временем коллеги Норвига под предводительством Рэя Курцвейла продолжают работать над созданием УЧИ.

На другом конце шкалы находится Бен Гертцель, который, как Гуд в свое время, считает, что создание УЧИ — всего лишь вопрос денег, и что он вполне может быть создан до 2020 г. Рэй Курцвейл — вероятно, лучший в истории технический прогнозист, — предсказывает появление УЧИ к 2029 г., но не ждет ИСИ раньше 2045 г. Он признает сложности на этом пути, но не жалеет энергии для убеждения окружающих в том, что ИСИ ждет долгое и безопасное путешествие в процессе рождения.

Мой неформальный опрос приблизительно двух сотен компьютерщиков на недавней конференции по УЧИ подтвердил

подозрения. Ежегодная конференция по УЧИ, организованная Гертцелем, представляет собой трехдневную площадку для встречи людей, активно работающих над созданием искусственного интеллекта человеческого уровня, или таких, как я, кто просто глубоко интересуется этим вопросом. Участники конференции представляют свои работы, образцы программ, хвалятся своими успехами. Я был на конференции, которая проходила в штаб-квартире Google в Маунтин-Вью (штат Калифорния). Я задавал присутствующим вопрос, когда следует ожидать появления УЧИ, и предлагал на выбор четыре варианта: к 2030 г., к 2050 г., к 2100 г. или никогда. Голоса распределились следующим образом: 42 % считает, что УЧИ будет создан к 2030 г.; 25 % — к 2050 г.; 20 % — к 2100 г.; и 2 % — никогда. Этот опрос, проведенный среди весьма специфической группы людей, подтвердил оптимистический настрой и временные рамки более формальных исследований, одно из которых я цитировал в главе 2. Среди письменных ответов я получил несколько жалоб на то, что не включил в число вариантов даты ранее 2030 г. Мне кажется, примерно 2 % респондентов ответили бы, что УЧИ появится к 2020 г., и еще 2 % — что еще раньше. В былые времена меня поражал подобный оптимизм, но теперь я перестал удивляться. Я воспользовался советом Курцвейла и теперь тоже считаю, что информационные технологии развиваются не линейно, а по экспоненте.

Но теперь, если вам случится оказаться в зале, где все присутствующие глубоко погружены в исследования УЧИ, и захочется развлечься, объявите громко: «УЧИ никогда не будет создан! Это попросту слишком сложно!» Гертцель, к примеру, в ответ на это заявление посмотрел на меня так, как будто я вдруг начал проповедовать креационизм. Сам Гертцель, как и Виндж, был когда-то профессором математики, и о будущем ИИ он судит по истории дифференциального исчисления.

> Если вы посмотрите, как математики вычисляли до Исаака Ньютона и Готфрида Лейбница, то увидите, что вычисление производной кубического многочлена занимало сотни страниц. Делалось это при помощи треугольников, подобных треугольников, хитрых диаграмм и тому подобного. Это было ужасно. Но теперь, когда у нас есть проработанная теория дифференциального исчисления, взять производную от кубического многочлена способен любой идиот-старшеклассник. Это совсем несложно.

Словно дифференциальное исчисление, развивающееся несколько столетий, исследования ИИ будут постепенно продвигаться, пока продолжающаяся практика не приведет к открытию новых теоретических правил, которые позволят специалистам формализовать значительную часть работы; с этого момента движение к УЧИ пойдет проще и быстрее.

> Ньютон и Лейбниц получили инструменты, такие как правило сложения, правило умножения, правило дифференцирования сложной функции, то есть все те базовые правила, которые мы сегодня изучаем в начальном курсе дифференциального исчисления, — *продолжал он.* — Пока этих правил не было, каждую задачу такого рода приходилось решать с нуля, как будто на пустом месте, а это гораздо сложнее. Так что математика ИИ у нас сегодня на уровне математики дифференциального исчисления, созданного Ньютоном и Лейбницем, и доказательство даже самых простых вещей, касающихся ИИ, требует безумного количества хитроумных вычислений. Но со временем у нас появится удобная теория разума, точно так же как когда-то появилась удобная теория дифференциального исчисления.

Но отсутствие удобной теории принципиально ничего не меняет. Гертцель говорит:

> Возможно, нам потребуется научный прорыв в строгой теории интеллекта человеческого уровня, прежде чем мы сможем сконструировать реальную систему УЧИ. Но пока я подозреваю, что это не обязательно. Я считаю, что можно, по идее, создать мощную систему УЧИ, двигаясь пошагово от нынешнего уровня наших знаний и занимаясь практическим конструированием без полного и строгого понимания теории разума.

Мы уже говорили, что проект Гертцеля OpenCog организует аппаратное и программное обеспечение в так называемую «когнитивную архитектуру», призванную моделировать деятельность мозга. И эта архитектура может в один прекрасный момент сделаться мощной и, возможно, непредсказуемой вещью. Где-то на пути развития, еще до создания всеобъемлющей теории человеческого разума, утверждает Гертцель, OpenCog может достичь уровня УЧИ.

Звучит безумно? Журнал *New Scientist* предположил, что система LIDA Университета Мемфиса, аналогичная OpenCog (мы уже говорили о ней в главе 11), демонстрирует признаки рудиментарного *сознания*. В самом общем плане принцип, известный как Теория глобального рабочего пространства, на котором построена LIDA, утверждает, что у человека ощущения, поступающие от органов чувств, просачиваются в подсознание и остаются там до тех пор, пока не станут достаточно важными, чтобы сообщить о них всему мозгу. Это и есть сознание, и его можно измерить при помощи простых заданий на осознанность, таких как нажатие кнопки при вспыхивании зеленой лампочки. Хотя кнопка у LIDA была «виртуальной», результаты тестов оказались вполне сравнимы с человеческими.

С подобными технологиями подход Гертцеля «поживем — увидим» представляется мне рискованным. Он намекает на создание того, что я уже описывал, — сильного машинного интеллекта, аналогичного, но не эквивалентного человеческому и куда менее понятного. Такой подход грозит серьезными и неприятными сюрпризами — как если бы УЧИ мог однажды просто появиться перед нами, застав нас недостаточно готовыми к «нормальным» авариям и, разумеется, без надежной защиты в лице формального дружественного ИИ. Как говорится: «Если достаточно долго идти по лесу, встретишь голодного медведя».

Елиезер Юдковски испытывает такие же опасения. И так же, как Гертцель, не считает, что сложность программного обеспечения может оказаться непреодолимым препятствием.

«УЧИ — проблема, решение которой заключено в мозгу, — сказал он мне. — Человеческий мозг способен сделать это — это не может быть настолько сложно. Естественный отбор глуп. Если естественный отбор смог решить проблему УЧИ, значит, она не может быть такой уж сложной в абсолютном смысле. Эволюция с легкостью выдала на-гора УЧИ, она просто случайно перебирала все подряд и сохраняла удачные варианты. Она проделала пошаговый путь без всякого представления о том, что ждет впереди».

Оптимизм Юдковски в отношении создания УЧИ отталкивается от мысли, что интеллект человеческого уровня был однажды создан природой в виде человека. Около 5 млн лет назад на земле жил общий предок человека и шимпанзе. Сегодня человеческий мозг вчетверо крупнее мозга шимпанзе. Получается, что примерно за 5 млн лет «глупый» естественный отбор вчетверо увеличил размер мозга и создал существо, которое намного умнее всех остальных.

«Умный» человек видит цель и стремится к ней. По идее, создать интеллект человеческого уровня он должен намного быстрее, чем это сделал естественный отбор.

Но опять же, предостерегает Юдковски, возникнет гигантская, буквально галактическая проблема, если кто-то создаст УЧИ прежде, чем он или кто-то другой придумает *дружественный ИИ* или способ надежно контролировать УЧИ. Если УЧИ возникнет в результате пошагового конструирования после удачного сочетания направленных усилий и случайностей, как предполагает Гертцель, то разве нам не следует ожидать интеллектуального взрыва? Если УЧИ сознает себя и способен к самосовершенствованию, как мы его определили, разве он не будет стремиться к удовлетворению базовых потребностей, которые могут оказаться несовместимыми с нашим существованием (мы говорили об этом в главах 5 и 6)? Иными словами, разве не следует ожидать, что вышедший из-под контроля УЧИ может убить нас всех?

«УЧИ — это тикающий часовой механизм, — сказал Юдковски. — Это крайний срок, к которому мы непременно должны построить дружественный ИИ, что труднее. Нам необходим дружественный ИИ. За возможным исключением нанотехно- логий, бесконтрольно выпущенных в мир, в целом каталоге катастроф не найдется ничего, что могло бы сравниться с УЧИ».

Разумеется, между теоретиками ИИ, такими как Юдковски, и его производителями, такими как Гертцель, возникают противоречия. Если Юдковски утверждает, что создание УЧИ — катастрофическая ошибка, если этот УЧИ не будет доказанно дружественным, то Гертцель хочет получить УЧИ как можно скорее, прежде чем автоматизированная инфраструктура облегчит ИСИ захват власти. Гертцелю приходили по электронной почте письма, хотя и не от Юдковски или его коллег, в которых его предупреждали, что если он будет продолжать развитие небезопасного УЧИ, то «будет виновен в холокосте».

Но вот парадокс. Если Гертцель откажется от работы над УЧИ и посвятит свою жизнь пропаганде отказа от подобных намерений, это никак и ни на что не повлияет. Другие компании, правительства и университеты будут и дальше гнуть свою линию. По этой самой причине Виндж, Курцвейл, Омохундро и другие считают, что отказ от разработки УЧИ невозможен. Более того, на свете много отчаянных и опасных стран — возьмите хотя бы Северную Корею и Иран, — а организованная преступность России и финансируемые государством преступники Китая запускают в Сеть все новые и новые вирусы и производят кибер- атаки, так

что отказ от разработки УЧИ означал бы попросту сдачу будущего безумцам и гангстерам.

Оборонительной стратегией, которая помогла бы обеспечить выживание человечества, может оказаться деятельность, которую уже начал Омохундро: разработка научных основ понимания и управления системами, обладающими самосознанием и способностью к самосовершенствованию, то есть УЧИ и ИСИ. А учитывая сложности создания противоядия, такого как дружественный ИИ, *прежде* создания УЧИ, развитие этой науки должно идти параллельно работам над УЧИ. Тогда к появлению УЧИ система контроля над ним будет готова. К несчастью для всех нас, разработчики УЧИ получили огромную фору; к тому же, как говорит Виндж, ветер глобальной экономики надувает их паруса.

Если проблема с программным обеспечением окажется неразрешимо сложной, то в колчане разработчика УЧИ останется еще по крайней мере пара стрел. Во-первых, не исключено, что проблему можно будет решить при помощи более быстрых компьютеров, а во-вторых, структуру мозга можно воспроизвести методом обратного проектирования.

Превращение системы ИИ в УЧИ методом грубой силы означает повышение функциональности аппаратной части ИИ, в первую очередь ее скорости. Интеллект и творческие возможности повышаются, если работают *во много раз* быстрее. Чтобы понять, как это происходит, представьте себе человека, способного сжать тысячу минут размышлений в *одну* минуту. В некоторых очень важных вопросах он оказывается во много раз умнее человека с тем же IQ, но думающего с обычной скоростью. Но обязательно ли интеллект должен начинаться на человеческом уровне, чтобы скорость имела значение? К примеру, если ускорить работу собачьего мозга в тысячу раз, что получится: шимпанзеподобное поведение или просто очень умная собака? Нам известно, что при четырехкратном увеличении *размеров* мозга, от шимпанзе до человека, человек получил по крайней мере одну суперспособность — речь. Более крупный мозг развивался постепенно, намного медленнее, чем та скорость, с которой обычно возрастает скорость процессоров.

В целом неясно, может ли скорость процессора компенсировать отсутствие разумных программ и проложить путь к УЧИ и далее, к интеллектуальному взрыву. Но согласитесь, это не кажется невозможным.

А теперь обратимся к так называемому «обратному проектированию» мозга и выясним, почему этот метод может оказаться безотказным средством решения проблемы сложности программного обеспечения. Мы

уже рассмотрели кратко противоположный подход — создание когнитивной архитектуры, которая стремится в общих чертах моделировать мозг в таких областях, как восприятие и навигация. Создатели этих когнитивных систем опираются на то, как работает мозг или, скорее, — и это важно — на то, как исследователь *представляет* себе работу мозга. Такие системы часто называют de novo, или «с начала», поскольку их авторы не отталкиваются от реального мозга, а начинают «с нуля».

Проблема в том, что системы, вдохновленные когнитивными моделями, в конечном итоге могут недотянуть до человеческих возможностей. Да, конечно, есть перспективные результаты в работе с естественным языком, зрением, системами «вопрос-ответ» и роботами, но при этом почти любой аспект методологии и принципов, которые должны продвинуть исследователей в направлении УЧИ, вызывает горячие споры. Сколько исследователей, столько и мнений. Новые узкие области исследований и смелые универсальные теории вырастают как грибы на почве любого успеха, как индивидуального, так и коллективного. Проходит немного времени, и они исчезают без следа. Как сказал Гертцель, не существует общепринятой теории разума и общепринятых представлений о том, как можно воспроизвести разум вычислительными методами. К тому же существуют такие функции человеческого сознания, для моделирования которых нынешние программные методики, судя по всему, годятся плохо; среди них общее обучение, объяснение, осмысление и контролирующее внимание.

Итак, чего в действительности удалось добиться в области ИИ? Вспомним старую шутку о пьянице, который потерял ключи и ищет их под фонарем. Полицейский присоединяется к поискам и спрашивает: «Где именно вы потеряли свои ключи?» Человек показывает в темную подворотню. «Там, — говорит он, — но здесь светлее».

Поиск, распознавание речи, компьютерное зрение и контекстный анализ (своего рода машинное обучение, с помощью которого Amazon и Netflix определяют, что вам может понравиться) — некоторые из областей ИИ, в которых достигнуты большие успехи. Конечно, успех — результат нескольких десятилетий работы, но следует отметить, что области эти относятся к числу простейших, так что пока работы идут в основном там, «где светлее». Сами ученые говорят, что снимают пока «низко висящие плоды». Но если наша конечная цель — УЧИ, то *все* приложения и инструменты на базе слабого ИИ могут показаться низко висящими плодами; все они лишь едва-едва приближают нас к цели — человеческому уровню интеллекта. Некоторые исследователи уверены, что приложения на

слабом ИИ вообще не являются продвижением к УЧИ. Это всего лишь неинтегрированные специальные приложения. В настоящий момент ни одна система искусственного интеллекта не может сравниться с человеческим интеллектом. Вы тоже разочарованы большими обещаниями и скромными результатами исследований ИИ? Не исключено, что на ваши чувства повлияли два очень распространенных наблюдения.

Во-первых, как говорит директор Института будущего человечества Оксфордского университета Ник Востром, «многие самые передовые ИИ просочились в распространенные приложения. Там их часто не называют ИИ, потому что, как только нечто становится достаточно полезным и распространенным, его перестают называть ИИ». Еще совсем недавно ИИ не был задействован в банковском деле, медицине, транспорте, инфраструктуре жизнеобеспечения и автомобилях. Но сегодня, если все ИИ вдруг исчезнут, вы не сможете получить кредит, электричество в вашем доме перестанет работать, а машина ехать; остановится большинство наземных и подземных поездов. Производство начало бы давать сбои и замерло, краны высохли, а пассажирские самолеты попадали бы с небес. В магазинах закончились бы продукты, и восполнить запасы оказалось бы невозможно. А когда, собственно, были внедрены все эти ИИ-системы? В последние тридцать лет, пока стояла так называемая ИИ-зима — период долгого спада уверенности инвесторов после излишне оптимистичного начала и несбывшихся предсказаний. Но *на самом деле* никакой зимы не было. Чтобы избавиться от ярлыка «искусственный интеллект», ученые перешли на технические термины, такие как «машинное обучение», «разумный агент», «вероятностные выводы», «продвинутые нейронные сети» и т. п.

Кстати говоря, и проблема классификации тоже никуда не делась. Области, которые когда-то считались прерогативой человека — шахматы и «Своя игра», к примеру, — сегодня принадлежат компьютерам (хотя нам по-прежнему позволяется играть). Но считаете ли шахматную программу, установленную на вашем компьютере, «искусственным интеллектом»? Что такое Watson — человекоподобная машина или всего лишь специализированная мощная система «вопрос-ответ»? Как мы будем называть ученых, когда компьютеры, такие как Golem (подходящее название!) Хода Липсома из Корнеллского университета, начнут заниматься наукой? Я хочу сказать, что с того самого дня, когда Джон Маккарти дал науке о машинном интеллекте имя, исследователи энергично создают ИИ, и с течением времени он становится все умнее, быстрее и мощнее.

Успехи ИИ в таких областях, как шахматы, физика и обработка

естественного языка, наводят на второе важное наблюдение. Сложные вещи просты, а простые — сложны. Эта аксиома известна как парадокс Моравека, поскольку пионер робототехники Ханс Моравек в классической книге «Дети разума» (Mind Children) выразил его лучше всего: «Сравнительно легко заставить компьютеры демонстрировать результаты, сравнимые с результатами взрослого человека, в тестах на интеллект или игре в шашки, и при этом трудно или даже невозможно дать им навыки годовалого ребенка в том, что касается восприятия и движений».

Сложнейшие головоломки, в которых мы просто не можем не допускать ошибок (скажем, «Своя игра» или вывод второго закона термодинамики Ньютона), хороший ИИ решает за несколько секунд. В то же время ни одна система компьютерного зрения не способна отличить собаку от кошки — а ведь с этим без труда справляется большинство двухлеток. До некоторой степени это «проблема яблок и апельсинов» — высокоуровневое восприятие против низкоуровневых моторных навыков. Но создателям ИИ следовало бы этого стыдиться, ведь они замахиваются на весь спектр человеческого интеллекта. Один из основателей Apple Стив Возняк предложил «легкую» альтернативу тесту Тьюринга, наглядно демонстрирующую сложность простых задач. Нам следовало бы считать любого робота разумным, говорит Возняк, если он сможет войти в незнакомый дом, найти в нем кофейник и соответствующие припасы и приготовить нам чашку кофе. Можно назвать это испытание кофе-тестом. Он может оказаться сложнее теста Тьюринга, поскольку для его прохождения необходим продвинутый ИИ, способный рассуждать и оценивать физические свойства предметов, обладающий машинным зрением и доступом к обширной базе данных, способный точно манипулировать роботизированными исполнительными устройствами, помещенный в универсальное роботизированное тело — и много чего еще.

В статье «Эра роботов» Моравек дал ключ к своему загадочному парадоксу. Почему сложные вещи просты, а простые — сложны? Потому что мозг тренировал и оттачивал «простые» вещи, связанные со зрением, действием и движением, с тех самых пор, как у наших нечеловеческих предков вообще появился мозг. «Сложные» вещи, такие как логические рассуждения, — относительно недавно приобретенные способности. И что вы думаете? Они проще, а не сложнее. Чтобы показать это, нам потребовались сложнейшие вычисления. Моравек писал:

> Задним числом представляется, что в абсолютном смысле логические рассуждения намного проще, чем вое- приятие и

> действие, — такую позицию несложно объяснить с точки зрения эволюции. Выживание человеческих существ (и их предков) сотни миллионов лет зависело от зрения и умения двигаться в физическом мире, и в этой конкуренции значительные части их мозга эффективно организованы именно для этой задачи. Но мы не ценили это монументальное умение, потому что им обладают все люди и большинство животных — оно обычно. С другой стороны, рациональное мышление, как в шахматах, — новообретенное умение возрастом, может быть, меньше ста тысяч лет. Части нашего мозга, посвященные этой задаче, не так хорошо организованы, и в абсолютном смысле мы не слишком хорошо умеем это делать. Но до недавнего времени у нас не было конкурентов, способных нас одолеть.

Под конкурентами здесь, разумеется, подразумеваются компьютеры. Создание компьютера, который делает что-нибудь умное, вынуждает исследователей внимательнее всмотреться в себя и других людей и оценить глубины и мели нашего собственного интеллекта. В вычислениях имеет смысл формализовать любые идеи математически. В области ИИ формализация выявляет скрытые правила и закономерности того, что мы делаем при помощи мозга. Но почему не отбросить лишнее и не взглянуть на работу мозга изнутри, через подробное исследование нейронов, аксонов и дендритов? Почему просто не разобраться, что конкретно делает каждый нейронный кластер мозга, и не смоделировать его при помощи алгоритмов? Если большинство исследователей ИИ согласны с тем, что мы можем разрешить загадки работы мозга, то почему просто не построить искусственный мозг?

Это аргументы в пользу «обратного проектирования мозга» — попытки создания компьютерной модели мозга, а затем обучения ее всему, что необходимо знать. Я уже сказал, что это может оказаться единственным способом получения УЧИ, если сложность программного обеспечения действительно окажется слишком высокой. Но опять же, что если эмуляция мозга во всей его полноте *тоже* окажется слишком сложной задачей? Что, если мозг на самом деле производит действия, которые мы не в состоянии искусственно воспроизвести? В недавней статье, критикующей представление Курцвейла о нейробиологии, один из основателей Microsoft Пол Ален и его коллега Марк Гривз написали, что «сложность мозга просто невероятна. Каждая структура в нем сформирована миллионами лет эволюции точно под конкретную задачу, какой бы она ни была… Каждая

отдельная структура и каждый нейронный контур в мозгу индивидуально настроены эволюцией и факторами среды». Иными словами, 200 млн лет эволюции превратили мозг в тонко настроенный мыслительный инструмент, который невозможно воспроизвести…

> Нет, нет, нет, нет, нет, нет, нет! Абсолютно не так. Мозг не оптимизирован, как и остальные части тела млекопитающего.

Взгляд Ричарда Грейнджера заметался в панике, как будто я выпустил в его кабинете в Дартмутском колледже в Хэновере летучую мышь. Грейнджер — настоящий янки Новой Англии, но выглядит как рок-звезда времен британского вторжения. Он строен, по-мальчишески симпатичен, с копной седеющих каштановых волос. Он серьезен и внимателен, как единственный член группы, понимающий, что играть на электроинструментах под дождем опасно. В молодости Грейнджер действительно мечтал о карьере рок-звезды, но стал вместо этого специалистом по вычислительной нейробиологии мирового класса; сегодня в его активе несколько книг и более сотни статей в рецензируемых журналах. Из светлого кабинета высоко над кампусом он руководит Лабораторией проектирования мозга в Дартмутском колледже. Именно здесь на Дартмутской летней исследовательской конференции по искусственному интеллекту в 1956 г. ИИ получил имя. Сегодня в Дартмуте уверены, что будущее ИИ лежит в области вычислительной нейро-биологии — изучения вычислительных принципов, на основании которых работает мозг.

> Наша цель в вычислительной нейробиологии — понять мозг достаточно хорошо, чтобы суметь воспроизвести его функции. Как сегодня простые роботы подменяют человека на физических работах, на заводах и в больницах, так проектирование мозга создаст замену нашим мыслительным способностям. Тогда мы сможем делать симулякры мозга и "ремонтировать" собственный в случае чего.

Если вы, как Грейнджер, специалист по вычислительной нейробиологии, то, вероятно, уверены, что моделирование мозга — чисто инженерная задача. Но чтобы верить в *это*, вам нужно взять великолепный человеческий мозг — а это, безусловно, царь среди всех органов млекопитающих — и понизить его способности на пару делений.

Грейнджер видит мозг в контексте всех прочих частей человеческого тела, ни одна из которых в процессе эволюции не достигла совершенства.

«Подумайте вот о чем, — Грейнджер согнул одну руку и внимательно изучил ее. — Мы не оптимальны, нет, нет и нет, и пять пальцев не оптимальны, и волосы над глазами, но не на лбу, не оптимальны, и нос между глазами, а не справа или слева не оптимален. Смешно, когда говорят, что любая из этих особенностей — результат оптимизации. *У всех* млекопитающих по четыре конечности, у всех есть "лицо", у всех глаза расположены над носом и надо ртом». Кроме того, оказывается, у всех нас почти одинаковый мозг. «Все млекопитающие, включая и человека, имеют в точности одинаковый набор отделов мозга, которые связаны между собой невероятно похоже, — сказал Грейнджер. — Эволюция работает методом случайного перебора и опробования вариантов, так что вы, конечно, можете думать, что все эти разные вещи испытываются в лаборатории эволюции и либо остаются, либо нет. Но на самом деле они не испытываются».

Тем не менее эволюция, создав мозг млекопитающего, наткнулась на нечто замечательное, говорит Грейнджер. Именно поэтому у мозга на пути от ранних млекопитающих до нас наблюдается лишь несколько небольших отклонений. Части мозга избыточны, а связи в нем не слишком точны и работают медленно, но его работа базируется на инженерных принципах, у которых нам стоило бы поучиться, — нестандартных принципах, которые люди пока еще не изобрели. Вот почему Грейнджер уверен, что создание интеллекта необходимо начинать с подробного изучения мозга. Он не считает, что созданные *de novo* когнитивные архитектуры — те, что не основаны на принципах устройства мозга, — смогут хотя бы приблизиться по возможностям к разуму.

«Из всех органов один только мозг способен к мысли, обучению, распознаванию, — говорит он. — Ничто, сконструированное человеком до сих пор, не сумело сравняться и тем более превзойти возможности мозга в любой из этих задач. Несмотря на громадные усилия и большие бюджеты, у нас нет искусственных систем, которые могли бы конкурировать с человеком в распознавании лиц, в восприятии естественных языков, в обучении на опыте».

Так что давайте отдадим мозгу должное. Мозг, а не мускулы, сделал нас доминирующим видом на планете. Мы взошли на пьедестал не потому, что были красивее животных, конкурировавших с нами за ресурсы или попросту хотевших съесть нас. Мы лучше думали даже, вероятно, в тех случаях, когда речь шла о конкуренции с другими человекообразными

видами. Разум, а не мускулы, принес нам победу.

Разум станет залогом победы и в стремительно приближающемся будущем, когда мы, люди, перестанем быть самыми умными существами в окружающей действительности. Почему нет? Когда технически отсталые народы брали верх над более развитыми? Когда менее разумные виды брали верх над более мозговитыми? Когда разумный вид хотя бы позволял относительно разумному виду существовать рядом, иначе чем в виде домашних любимцев? Посмотрите хотя бы, как мы, люди, относимся к ближайшим своим родичам, высшим приматам — шимпанзе, орангутанам и гориллам. Те, кто еще не погиб, не оказался в зоопарке и не превратился в циркового клоуна, находятся в опасности и, можно сказать, доживают последние дни.

Разумеется, Грейнджер прав, и ни одна искусственная система не может сравниться с человеком в распознавании лиц, обучении и владении языком. Но в узких специализированных областях ИИ обладает потрясающей мощью. Представьте себе существо, которое имеет в своем распоряжении всю эту мощь, и представьте, что оно по-настоящему разумно. Как долго оно готово будет оставаться нашим орудием? Вот что сказал историк Джордж Дайсон о том, где может жить такое сверхразумное существо, после экскурсии по штаб-квартире Google:

> На протяжении тридцати лет я думал о том, какие признаки существования можем мы ожидать от настоящего ИИ. Конечно, речь не идет о прямом объявлении в СМИ, которое может вызвать нежелательную реакцию в виде выдергивания вилки из розетки. Индикатором может быть аномальная концентрация, создание богатства, неутолимая жажда все новой информации, дискового пространства и процессорных мощностей или согласованная попытка обеспечить себя анонимным источником непрерывного питания. Но подлинным индикатором, мне кажется, был бы кружок жизнерадостных, довольных, интеллектуально и физически удовлетворенных людей вокруг ИИ. Там не будет нужды в истинно верующих, не будет загрузки человеческого сознания в компьютер или чего-то другого, столь же зловещего; это будет просто постепенный, мягкий, всеобъемлющий и взаимно полезный контакт между нами и растущим чем-то. Пока все это лишь непроверяемая гипотеза.

Далее Дайсон цитирует писателя-фантаста Саймона Ингза:

Когда наши машины обогнали нас, стали слишком сложными и эффективными для того, чтобы ими можно было управлять, они сделали это так быстро, так гладко и так эффективно, что только глупец или пророк осмелился бы протестовать.

Глава 13
Непознаваемы по природе

Как из-за великолепной способности к планированию, так и из-за технологий, которые он способен будет создать, разумно предположить, что первый сверхразум будет очень мощным. Вполне возможно, что соперников у него не будет: он сможет достичь почти любой цели и пресечь любую попытку помешать ему. Он мог бы уничтожить всех прочих агентов, убедить их изменить свое поведение или заблокировать любые попытки вмешательства с их стороны. Даже «скованный сверхразум», работающий на отдельном изолированном компьютере и способный взаимодействовать с внешним миром только через текстовый интерфейс, возможно, сумеет вырваться из своего заключения, убедив своих «тюремщиков» выпустить его. Есть даже некоторые предварительные экспериментальные указания на то, что так и произойдет.

Ник Востром, директор Института будущего человечества, Оксфордский университет

Видя, как ИИ наступает по всем фронтам, от Siri до Watson, OpenCog и LIDA, трудно поверить, что человечество не сможет создать УЧИ из-за сложности реализации. Если увеличение мощности компьютеров не приведет к желаемому результату, поможет обратное проектирование, хотя времени на это уйдет больше. Несмотря на это, Рик Грейнджер ставит перед собой цель: понять мозг по восходящей, от простого к сложному, путем реализации самых фундаментальных его структур в виде компьютерных программ. И он не может не иронизировать над исследователями, работающими сверху вниз по когнитивной вертикали с применением принципов информатики.

Они изучают поведение человека и смотрят, нельзя ли

> смоделировать такое поведение на компьютере. По справедливости, это немного напоминает попытку разобраться в автомобиле, не заглядывая под капот. Мы считаем, что можем точно сказать, что такое разум. Мы считаем, что можем точно сказать, что такое познание. Мы считаем, что можем точно сказать, что такое адаптивные способности. Но единственная причина того, что мы имеем хоть какое-то представление об этих вещах, заключается в том, что мы наблюдаем, как человек делает "разумные" вещи. Но одно только внешнее наблюдение не позволяет нам понять в подробностях, что именно в этот момент происходит. Принципиальный вопрос звучит так: какова инженерная спецификация логических рассуждений и обучения? Инженерной спецификации не существует, так на что же они опираются, кроме наблюдений?

А ведь не секрет, что человек плохо умеет анализировать свое поведение. «Огромное количество исследований по психологии, нейробиологии и когнитивистике показывает снова и снова, как плохо мы умеем вглядываться в себя, — говорит Грейнджер. — Мы не имеем понятия ни о собственном поведении, ни о процессах, которые стоят за ним». Грейнджер отмечает, что мы также плохо умеем принимать рациональные решения, точно рассказывать об увиденном и вспоминать то, что случилось совсем недавно. Но ограниченность возможностей человека как наблюдателя не означает, что когнитивные науки, основанные на наблюдениях, — полная чепуха.

Грейнджер просто считает, что это неподходящий инструмент для проникновения в тайны интеллекта.

«В вычислительной нейробиологии мы задаем вопрос: "Хорошо, что человеческий мозг *на самом деле делает?"* — говорит Грейнджер. — Не что мы *думаем,* он делает, и не что мы бы *хотели,* чтобы он делал. Что он на самом деле делает? И может быть, информация об этом даст нам впервые и определение разума, и определение адаптации, и определение языка».

Выяснение вычислительных принципов мозга начинается с того, что ученые проверяют, чем занимаются в мозгу те или иные кластеры нейронов. Нейроны — это клетки, которые посылают и принимают электрохимические сигналы. Важнейшую часть их составляют аксоны (волокна, соединяющие нейроны между собой; именно они обычно являются отправителями сигнала), синапсы (соединения, через которые

проходит сигнал) и дендриты (обычно получатели сигнала). В мозгу человека около ста миллиардов нейронов. Каждый из них соединен со многими десятками тысяч других нейронов. Такое обилие связей делает все операции мозга параллельными, а не последовательными, как у большинства компьютеров. В вычислительных терминах последовательная обработка данных означает, что операции выполняются по очереди, по одной. Параллельная обработка данных означает, что большое количество данных обрабатывается одновременно — иногда в одно и то же время проходят сотни тысяч или даже миллионы операций.

Представьте на мгновение, что вы переходите оживленную городскую улицу; подумайте, сколько информации — цвета, звуки, запахи, температура, ощущение асфальта под ногами — одновременно поступает при этом в ваш мозг через уши, глаза, нос, конечности и кожу. Если бы ваш мозг не был органом, способным обработать все это одновременно, он мгновенно выключился бы от перегрузки. Вместо этого ваши органы чувств собирают всю эту информацию, мозг пропускает ее через нейроны и обрабатывает, — а в результате вы ведете себя соответственно, останавливаетесь перед светофором и избегаете столкновений с другими пешеходами.

Группы нейронов работают вместе и объединяются в схемы, сильно напоминающие электронные. В электронной схеме протекает электрический ток через специальные элементы, такие как резисторы и диоды. В ходе этого процесса ток выполняет различные функции — включает свет, к примеру, или запускает косилку. Если вы составите список инструкций, которые приводят к выполнению этой функции или какого-то вычисления, вы получите компьютерную программу или алгоритм.

Кластеры нейронов в вашем мозгу образуют схемы, которые действуют как алгоритмы. При этом они не включают свет, а распознают лица, планируют отпуск или набирают на клавиатуре предложение. И все это время работают параллельно. Откуда исследователи знают, что происходит в этих нейронных кластерах? Попросту говоря, они собирают детальную информацию при помощи специальных инструментов визуального исследования мозга, начиная от электродов, вживленных непосредственно в мозг животных, и заканчивая такими аппаратами, как ПЭТ- и фМРТ-сканеры применительно к людям. Нейронные зонды внутри и снаружи черепа способны показать, что делают отдельные нейроны, а маркирование нейронов электрически чувствительными красками наглядно показывает, когда те или иные нейроны активны. Из этих и других методик следуют проверяемые гипотезы об алгоритмах, управляющих контурами

мозга. Кроме того, начато определение точной функции некоторых отделов мозга. Уже больше десяти лет, к примеру, нейробиологи знают, что узнавание лиц происходит в части мозга, известной как веретенообразная извилина.

Постойте, но в чем же суть? Неужели вычислительные системы, построенные по образу и подобию мозга (подход вычислительной нейробиологии), работают лучше, чем те, что построены *de novo* (подход информатики)?

Ну, одна из разновидностей систем, сделанных по образцу мозга, — искусственные нейронные сети (ИНС), — работает уже так давно и хорошо, что стала, по существу, основой ИИ. Как говорилось в главе 7, ИНС (которые можно разделить на аппаратные и программные) были придуманы в 1960-е гг. специально для того, чтобы играть роль нейронов. Одно из основных их преимуществ состоит в том, что ИНС обучаемы. Если вы хотите научить нейронную сеть переводить текст с французского языка на английский, к примеру, вы можете для начала подать на вход французский текст и точный английский перевод этого текста. Этот процесс называется контролируемым обучением. Если образцов будет достаточно, сеть распознает и усвоит правила, связывающие французские слова с их английскими эквивалентами.

В мозгу нейроны соединяются друг с другом через синапсы, и именно в этих точках контакта происходит обучение. Чем прочнее синаптическая связь, тем прочнее воспоминание. В ИНС прочность синаптического соединения называется его «весом» и выражается в виде вероятности. ИНС присваивает си- наптические веса правилам перевода с иностранного языка, которые усваивает в процессе обучения. Чем дольше длится обучение, тем лучше будет перевод. В ходе обучения ИНС учится распознавать собственные ошибки и соответствующим образом корректирует синаптические веса. Это означает, что нейронная сеть изначально способна к самосовершенствованию.

После обучения, когда на вход системы поступит французский текст, ИНС сверится с вероятностными правилами, усвоенными в ходе обучения, и выдаст свой лучший перевод. По существу, ИНС ищет закономерности в структуре данных. На сегодняшний день поиск закономерностей в больших объемах неструктурированных данных — одна из самых перспективных областей применения ИИ.

Помимо перевода и анализа больших объемов информации ИНС сегодня активно используются в структуре ИИ, анализируют фондовый рынок и распознают объекты на картинках. Они присутствуют в

программах оптического распознавания символов, предназначенных для чтения печатного текста, и в микросхемах, управляющих ракетами. ИНС обеспечивают «ум» умным бомбам. Да и большинство архитектур УЧИ без них не обойдутся.

Из главы 7 стоит вспомнить еще кое-что важное об этих вездесущих нейронных сетях. Подобно генетическим алгоритмам, ИНС работают по принципу «черного ящика». Это значит, что входные данные — в нашем примере французский текст — прозрачны. А выходные — здесь это английский текст — понятны. Но что происходит в промежутке, никто не знает. Все, что может сделать программист, — это руководить и направлять обучение ИНС, подбирая примеры и пытаясь улучшить результат перевода. А поскольку результат работы «черного ящика» — искусственного интеллекта — непредсказуем, его нельзя считать по-настоящему безопасным.

Судя по результатам работы алгоритмов Грейнджера, построенных по образу и подобию мозга, можно сделать вывод, что при создании искусственного разума лучше, возможно, следовать эволюционной модели и копировать человеческий мозг, чем создавать *de novo* когнитивные системы на базе компьютерных наук.

В 2007 г. ученики Грейнджера из Дартмутского колледжа написали по результатам исследований мозга алгоритм визуального восприятия, распознававший объекты в 140 раз быстрее, чем традиционные алгоритмы. Он показал лучшие результаты, чем 80000 других алгоритмов, и выиграл приз IBM в $10 000.

В 2010 г. Грейнджер и его коллега Ашок Чандрашекар создали по образцу мозга алгоритмы контролируемого обучения, которые используются при обучении машин распознаванию оптических символов и голоса, выделению спама и т. п. Алгоритмы, сделанные по образцу мозга для использования в процессорах с параллельной обработкой данных, работали так же точно, как и последовательные алгоритмы того же назначения, но *в десять с лишним раз быстрее*. Образцом для новых алгоритмов послужили самые распространенные типы нейронных кластеров, или схем, мозга.

В 2011 г. Грейнджер с коллегами запатентовал чип с изменяемой конфигурацией для параллельной обработки данных, основанный на этих алгоритмах. Это означает, что некоторые из самых распространенных структур мозга уже можно воспроизвести в виде компьютерного чипа. Стоит собрать их вместе, как в программе SyNAPSE фирмы IBM, — и вы на пути к созданию виртуального мозга. Всего один из этих чипов уже

сегодня мог бы ускорить и улучшить работу систем, созданных для распознавания лиц в толпе, обнаружения ракетных пусков на спутниковых фотографиях, автоматической каталогизации вашей коллекции цифровых снимков и решения сотен других задач. Со временем искусственное воспроизведение мозговых схем, возможно, позволит «ремонтировать» поврежденный мозг, встраивая в него компоненты, способные восстановить «неисправные» области. Когда-нибудь микросхема параллельной обработки данных, запатентованная командой Грейнджера, могла бы заменить собой поврежденные структуры живого мозга.

А пока программы, построенные по алгоритмам работы мозга, пробиваются в традиционные компьютерные процессы. Подкорковые узлы — древняя, «рептильная» часть мозга, отвечающая за контроль движений. Исследователи выяснили, что подкорковые узлы при освоении навыков задействуют алгоритмы того типа, что используются в обучении с подкреплением. Команда Грейнджера открыла, что нейронные контуры в коре головного мозга — последнем по времени появления отделе мозга — выстраивают иерархии фактов *и* создают связи между фактами (подобно иерархическим базам данных). Это два разных механизма.

А теперь самое интересное. Нейронные схемы в этих двух частях мозга, подкорковых узлах и коре, связаны между собой посредством других контуров, сочетающих в себе свойства того и другого. В вычислительной технике существуют прямые параллели этому. Компьютерные системы обучения с подкреплением действуют методом проб и ошибок — они должны проверить громадное число возможностей, чтобы узнать наконец верный ответ. Именно так *мы* в основном используем подкорковые узлы для освоения автоматических действий, таких как езда на велосипеде или удар по бейсбольному мячу.

Но у человека есть еще иерархическая система коры, позволяющая нам не просто перебирать вслепую все возможности методом проб и ошибок, а каталогизировать и иерархически организовывать их, а затем просеивать куда более разумно. Такая комбинация работает намного быстрее и дает лучшие результаты, чем у животных, к примеру, пресмыкающихся, использующих только систему проб и ошибок, работающую в подкорковых узлах. Возможно, самое продвинутое, что мы можем делать в системе, объединяющей в себе кору и подкорковые узлы, — это проводить *внутренние* тесты типа проб и ошибок без необходимости выходить на внешний уровень и проверять в реальном мире. Мы можем проводить множество проверок, просто обдумывая что-то: мы моделируем все это в своей голове. Искусственные алгоритмы, в которых сочетаются

эти методы, работают много лучше, чем каждый из методов в отдельности. Грейнджер предполагает, что примерно такое же преимущество дает объединение двух этих систем в нашем мозгу.

Грейнджер и другие нейробиологи выяснили также, что цепями мозга управляют всего несколько типов алгоритмов. Одни и те же базовые вычислительные системы используются вновь и вновь в различных сенсорных и когнитивных операциях, таких как слуховое восприятие и дедуктивные рассуждения. Не исключено, что, как только эти операции удастся воспроизвести в компьютерах программно и аппаратно, можно будет простым их дублированием создавать модули для моделирования различных частей мозга. А воссоздание алгоритмов, скажем, слухового восприятия должно помочь в усовершенствовании приложений распознавания устной речи. Более того, это уже произошло.

Курцвейл был одним из первых новаторов, применивших знания о мозге в программировании. Как мы уже говорили, он утверждал, что обратное проектирование мозга — самый перспективный путь к созданию УЧИ. В очерке, отстаивавшем эту точку зрения и его предсказания по поводу этапов и вех технического развития, он писал:

> Говоря в целом, мы ищем в биологии методы, способные ускорить разработку ИИ, которые в основном ведутся без сколько-нибудь полного представления о том, как аналогичные функции реализует мозг. Из своего собственного опыта в области распознавания речи я знаю, что наша работа сильно ускорилась, когда мы получили достоверные сведения о том, как мозг готовит и обрабатывает слуховую информацию.

Еще в 1990-е гг. фирма Курцвейла делала первые шаги в распознавании речи и разрабатывала приложения, которые позволили бы врачам диктовать медицинские отчеты. Курцвейл продал свою компанию, и на ее, в частности, основе возникла компания Nuance Communications. Всякий раз, прибегая к помощи Siri, вы пользуетесь алгоритмами Nuance, задействованными в речевой части этого волшебства. Распознавание речи — искусство перевода произнесенного слова в текстовую форму (не путать с обработкой естественного языка, цель которой — извлечение смысла из написанных слов). После того, как Siri переведет ваш запрос в текстовый вид, в дело вступают три остальных ее главных таланта: обработка естественного языка, поиск в громадной базе данных и взаимодействие с поисковыми серверами Интернета, такими как OpenTable, Movietickets и

Wolfram | Alpha.

Watson — что-то вроде Siri на стероидах и безусловный чемпион по обработке естественного языка. В феврале 2011 г. эта программа, применив построенные по образцу мозга системы, одержала убедительную победу над соперниками-людьми в «Своей игре». Как и шахматный чемпион Deep Blue, для IBM Watson — это способ продемонстрировать свои вычислительные ноу-хау при одновременном продвижении дела ИИ. Известная игра представляла для программы серьезный вызов, ведь в вопросах часто задействованы ассоциации и игра слов. Участники должны понимать каламбуры, сравнения и культурный контекст, а ответы необходимо формулировать в том же стиле, что и вопросы. Однако распознавание речи не является специализацией Watson. Он не способен понимать устную речь. А поскольку он не видит и не чувствует, он не может и читать, так что в ходе игры вопросы приходится вводить вручную. К тому же Watson не слышит, так что аудио- и видеоподсказки исключались.

Эй, погодите-ка! Так что же выиграл Watson — «Свою игру» или специально подстроенную под него ее разновидность?

Уже после громкой победы, чтобы научить Watson понимать человеческую речь, IBM скрестила его с технологией распознавания речи фирмы Nuance. Кроме того, Watson читает терабайты медицинской литературы. Одна из целей IBM — уменьшить его габариты с нынешней комнаты, наполненной серверами, до, скажем, размеров холодильника и сделать из него лучшего в мире медика-диагноста. Когда-нибудь, уже довольно скоро, вам, возможно, придется записываться к врачу через виртуального помощника, который засыплет вас вопросами и сообщит врачу ваш предварительный диагноз. К несчастью, Watson по-прежнему не видит и может просмотреть такие признаки вашего состояния, как блеск в глазах, румяные щеки или свежее пулевое ранение. IBM планирует также установить Watson на ваш смартфон в качестве идеального приложения «вопрос-ответ».

Где же работают скопированные с мозга способности Watson? Его аппаратное обеспечение имеет выраженную параллельную структуру: 3000 параллельных процессоров оперируют 180 различными программными модулями, написанными для параллельных процессоров. Параллельная обработка информации — величайшее достижение мозга, и разработчики программного обеспечения вовсю пытаются повторить это достижение. Как сказал мне Грейнджер, параллельные процессоры и написанные для них программы не оправдали ожиданий. Почему? Потому что написанные

для них программы плохо справляются с разделением задач для параллельного решения. Но, как продемонстрировал Watson, доработанные параллельные программы меняют ситуацию, и «железо» с параллельной архитектурой не слишком отстает. Разрабатываются новые чипы параллельных вычислений, которые должны невероятно ускорить работу уже существующих программ.

Watson показал, что параллелизм способен выполнять поразительные объемы вычислительных работ с невероятной скоростью. Но главное достижение Watson — способность самостоятельно обучаться. Его алгоритмы находят закономерности в текстовых данных, которые дают ему его создатели. Сколько может быть таких данных? Энциклопедии, газеты, романы, словари, вся «Википедия», Библия — всего около 8 млн толстых томов текста, которые машина обрабатывает со скоростью 500 гигабайт (примерно тысяча толстых книг) в секунду. Следует отметить, что в число текстовых материалов входили подготовленные словарные базы, таксономии (систематизированные и классифицированные коллекции слов) и онтологии (описания слов и отношений между ними). По существу, в этих материалах собран человеческий здравый смысл. К примеру, «крыша — это верхняя часть дома, не нижняя его часть, как подвал, и не боковая, как наружная стена». Это предложение говорит Watson кое-что о крышах, домах, подвалах и стенах, но для того, чтобы предложение имело смысл, ему необходимо знать определения всех использованных слов и определение слова «часть» заодно. Кроме того, ему нужно увидеть это слово не один раз, а множество, в разных предложениях. У Watson есть такая возможность.

Во второй игре объявленного IBM конкурса «Своей игры» появился такой вопрос: «Это условие в коллективном трудовом договоре говорит о том, что заработная плата может увеличиваться и уменьшаться в зависимости от определенного параметра, такого, например, как стоимость жизни». Сначала Watson разобрал это предложение, то есть выбрал из него и проанализировал ключевые слова. Затем он извлек из уже освоенных источников информацию о том, что заработная плата — это нечто, что может увеличиваться и уменьшаться, что в трудовом договоре пишут о заработной плате и что в договорах бывают условия. У него был и еще один очень важный ключ — категория, к которой относился вопрос, называлась «юридические "И"». Из этого Watson узнал, что ответ связан с распространенным юридическим термином и должен начинаться с буквы "И". Watson быстрее людей дал ответ на этот вопрос: «Что такое индексация заработной платы?» — и это заняло у него целых три секунды.

А после первого правильного ответа в той или иной категории Watson обретал уверенность (и начинал играть смелее), поскольку «понимал», что верно интерпретировал категорию. Система адаптировалась к игре или училась играть лучше прямо по ходу игры.

Но отвлекитесь на минуту от «Своей игры» и представьте, как быстро адаптивное машинное обучение может быть перенастроено на обучение вождению автомобиля, управлению танкером в море или геологической разведке на золото. Подумайте, какая мощь скрыта в интеллекте человеческого уровня.

Watson продемонстрировал еще один интересный вид интеллекта. Его программа DeepQA генерирует сотни возможных ответов и собирает сотни данных в пользу каждого из них. Затем она фильтрует и ранжирует ответы по уровню своей уверенности в каждом. Если она не чувствует уверенности ни в одном ответе, она не станет отвечать, поскольку в «Своей игре» за неверный ответ назначается штраф. Иными словами, Watson знает, чего он не знает. Конечно, вы можете не поверить, что вероятностные вычисления — это осознание себя, но согласитесь, это уже шаг в верном направлении. *Знает* ли Watson на самом деле что-нибудь?

Ну, если контуры мозга управляются алгоритмами, как утверждают Грейнджер и другие специалисты по вычислительной нейробиологии, то логично спросить: а знаем ли мы, люди, хоть что-нибудь? Или, говоря иначе, может быть, и мы, и они что-то знаем. И, разумеется, Watson — это настоящий прорыв, способный многому нас научить. Курцвейл сказал об этом так:

> Много писали о том, что Watson работает через статистические расчеты, а не через "подлинное" понимание. Многие читатели понимают это так, что Watson просто собирает статистику о словосочетаниях… Но точно так же и пространственное распределение концентрации нейротрансмиттеров в коре человеческого мозга можно назвать "статистической информацией". В самом деле, мы разрешаем свои сомнения примерно так же, как это делает Watson, — сравнивая вероятности различных интерпретаций фразы.

Иными словами, как мы уже говорили, ваш мозг помнит информацию благодаря прочности электрохимических сигналов в синапсах, участвовавших в кодировании этой информации. Чем выше концентрация химических веществ, тем дольше и надежнее будет храниться информация.

Основанные на фактах вероятности, с которыми работает Watson, тоже представляют собой своего рода шифр, но только в компьютерной форме. Это знание или нет? Такая дилемма приводит на ум загадку «китайской комнаты» Джона Сёрля, речь о которой шла в главе 3. Как мы вообще сможем узнать, думают компьютеры по-настоящему или просто хорошо имитируют этот процесс?

Что характерно, на следующий день после победы Watson в «Своей игре» Сёрль сказал:

> IBM придумала хитроумную программу — но не компьютер, способный думать. Watson не понимал ни вопросов, ни своих ответов, ни того, что некоторые из них оказывались верными, а некоторые — неверными, ни того, что все это игра, ни того, что он выиграл, — потому что он вообще ничего не понимает.

Дэвид Ферруччи, ведущий специалист IBM по системе Watson, на вопрос о том, умеет ли Watson думать, ответил, перефразируя известного голландского компьютерщика Эдгера Дейкстру: «Умеет ли подводная лодка плавать?»

То есть подлодка не «плавает» так, как плавает рыба, но она может передвигаться в толще воды быстрее большинства рыб и оставаться под водой дольше любого млекопитающего. Более того, в некоторых отношениях подлодка плавает лучше рыб и млекопитающих именно потому, что она плавает не так — у нее свои достоинства и недостатки. Интеллект Watson производит сильное впечатление, хотя и ограничен, потому что не похож на человеческий. В среднем он работает много быстрее и способен делать вещи, доступные только компьютерам, — к примеру, отвечать на вопросы «Своей игры» круглосуточно и без выходных, а также подключаться к сборочной линии новых архитектур типа Watson, когда возникнет необходимость поделиться с ними знаниями и готовыми программами. А в вопросе о том, думает ли Watson, я предлагаю довериться нашим органам чувств.

Кен Дженнигс, один из людей — оппонентов Watson по «Своей игре» (называвший себя, кстати, «Великой углеродной надеждой»), *ощущал* своего противника как человека.

> Методы, при помощи которых компьютер раскручивает вопросы в игре, очень похожи на мои. Эта машина сосредоточивается на ключевых словах вопроса, а затем

обшаривает свою память (в случае Watson это пятнадцать терабайт базы человеческих знаний) в поисках групп ассоциаций, связанных с этими словами. Она строго проверяет первые ассоциации по всей контекстной информации, какую только может собрать: название категории; тип возможного ответа; время, место и тендерные особенности, на которые указывает вопрос, и т. д.

И когда машина решает, что достаточно "уверена" в ответе, она нажимает на звонок. Все это происходит мгновенно, и для игрока-человека это интуитивный процесс, но я убежден, что, если разобраться, мой мозг проделывает приблизительно то же самое.

Действительно ли Watson думает? И как много он на самом деле понимает? Я не знаю наверняка. Но я уверен, что Watson — это первый вид совершенно новой экосистемы, первая машина, по поводу которой мы можем задаваться такими вопросами.

Может ли Watson стать основой когнитивной архитектуры, представляющей собой полноценный искусственный интеллект человеческого уровня? Ну, у этой системы серьезная база, которой нет больше ни у какой другой: это и толстый кошелек, и компания, публично заявляющая о готовности преодолевать препятствия и рисковать, и план будущего финансирования, обеспечивающего жизнь и развитие. Если бы я управлял IBM, я воспользовался бы популярностью компании, уровнем продаж и научными результатами, полученными в работе над сложнейшими проектами Deep Blue и Watson, и объявил бы миру, что в 2020 г. IBM готова будет пройти тест Тьюринга.

Успехи в обработке естественных языков очень скоро преобразят те области экономики, которые до сих пор, казалось, совершенно не затрагивал технический прогресс. Через несколько лет библиотекари и исследователи всех сортов присоединятся в очередях на бирже труда к продавцам, банковским кассирам, турагентам, биржевым брокерам, кредитным менеджерам и работникам справочных служб. За ними последуют врачи, юристы, налоговые и пенсионные консультанты. Вспомните, как быстро автоматы едва ли не полностью заменили банковских операторов, а машины начали вытеснять живых кассиров в супермаркетах. Если вы работаете в информационной индустрии (а цифровая революция превращает все на свете в информационную индустрию), берегитесь.

Вот простой пример. Любите университетский баскетбол? Определите, какая из этих заметок написана спортивным журналистом — человеком?

Текст А

Опрос среди тренеров по поводу того, кто может занять первое место, показал, что Ohio State (17) и Kansas (14) на двоих получают ни много ни мало тридцать один голос. Последние изменения в верхней части пула стали неизбежны, когда Duke в субботу вечером обидел противник из Конференции Атлантического побережья — Virginia Tech. Buckeyes (27:2), пробиваясь обратно на вершину, достаточно легко нанесли поражение соперникам из десятки сильнейших — Illinois и Indiana. Ohio State начал со счета 24:0 и четыре недели в начале сезона был лучшим, пока не опустился на третье место. А в тройке сильнейших он находится уже пятнадцатую неделю подряд. Kansas (27:2) по-прежнему второй и по результатам опроса отстает от Ohio State всего на четыре пункта.

Текст Б

Ohio State возвращает себе первое место через неделю после победы дома над Illinois со счетом 89:70. За этим последовала еще одна победа дома над Indiana, 82:61. Utah State входит в число двадцати пяти лучших и находится на двадцать пятом месте после победы дома над Idaho, 84:68. Temple на этой неделе выпадает из рейтинга после проигрыша первому на тот момент Duke и выигрыша над George Washington, 57:41. Arizona на этой неделе сильно сдвинулась и занимает восемнадцатое место после неприятного проигрыша Университету Южной Калифорнии со счетом 65:57 и неприятного проигрыша Университету Калифорнии в Лос-Анджелесе, 71:49. St. John's подскочил на восемь позиций и занял пятнадцатое место после выигрыша у пятнадцатой на тот момент в рейтинге Villanova, 81:68 и DePaul, 76:51.

Ну как, догадались? Конечно, ни один из авторов не создал шедевр, но лишь один из них — живой человек. Это автор текста А, опубликованного на сайте ESPN. Текст Б написан автоматической издательской платформой, созданной Робби Алленом из фирмы Automated Insight. За один год его компания со штаб-квартирой в Дареме (штат Северная Каролина)

сформировала 100 ООО автоматических статей на спортивные темы и разместила их на сотнях сайтов, посвященных местным командам. Зачем миру нужны роботы, пишущие спортивные статьи? Аллен рассказал мне, что о многих командах вообще никто не пишет, и их болельщики остаются в полной неизвестности. Кроме того, написанные ИИ статьи могут появляться на сайте команды и подхватываться другими сайтами всего через несколько минут после окончания игры. Человек не в состоянии работать так быстро. Аллен — бывший старший инженер Cisco Systems — не захотел назвать мне «секретный источник» своей поразительной архитектуры. Но скоро, сказал он, Automated Insights начнет выдавать статьи на темы финансов, погоды, недвижимости и местных новостей. Его «голодные» серверы нуждаются лишь в слабоструктурированных данных.

Если вы уже начали знакомиться с результатами вычислительной нейробиологии, вам трудно (мне по крайней мере точно трудно) представить значительный прогресс в области архитектур УЧИ, достигнутый исключительно на базе когни- тивистики. Полное понимание принципов, на базе которых функционирует мозг на каждом уровне, представляется мне более надежным и исчерпывающим способом создания разумной машины, нежели любые усилия, прилагаемые без такого понимания. Вы согласны со мной? Ученым не понадобится анализировать работу каждого из ста миллиардов нейронов мозга, чтобы разобраться в их функциях и смоделировать их, поскольку структура мозга очень избыточна. Кроме того, не потребуется, возможно, моделировать весь объем мозга, включая те области, которые контролируют автономные функции, такие как дыхание, сердцебиение, реакцию «дерись или беги» и сон. С другой стороны, может стать очевидным, что интеллект должен обитать в теле, которое он контролирует, и что тело это должно существовать в сложной среде. Споры о необходимости воплощения на этом не разрешатся. Но подумайте о таких концепциях, как *яркий, сладкий, твердый* и *острый.* Как может ИИ узнать, что означают эти ощущения, и построить на них концепции, если у него не будет тела? Не возникнет ли непреодолимое препятствие для превращения машины в искусственный интеллект человеческого уровня, если у нее не будет чувств?

На этот вопрос Грейнджер отвечает: «Неужели Хелен Келлер[32] была менее человеком, чем вы? Или парализованный человек? Неужели мы не можем вообразить себе интеллект с совершенно иными способностями, со зрением, датчиками прикосновения и микрофонами для слуха? Конечно, у него будут несколько иные представления о *ярком, сладком, твердом, остром,* — но очень вероятно, что представления об этих концепциях у

многих людей с различными вкусовыми рецепторами, возможно, с нарушениями здоровья, принадлежащих к разным культурам и живущих в разных условиях, и сегодня сильно различаются».

Наконец, вполне возможно, что для рождения интеллекта ученым необходимо смоделировать еще и эмоциональный орган. В принятии решений эмоции зачастую играют более серьезную роль, чем разум; то, кто мы есть, и то, как мы думаем, в значительной степени определяется гормонами, которые возбуждают и успокаивают нас. Если мы и правда хотим смоделировать человеческий интеллект, разве эндокринная система не должна быть частью архитектуры? Не исключено, что интеллект требует полноценных человеческих ощущений. Первичные ощущения, или субъективное качество обитания в теле и жизни в состоянии постоянной сенсорной обратной связи, могут оказаться необходимыми для интеллекта человеческого уровня. Несмотря на аргументы Грейнджера, исследования показали, что люди, парализованные в результате травмы, испытывают эмоциональное омертвение. Можно ли создать эмоциональную машину без тела, а если нет, то неужели такую важную часть человеческого интеллекта никогда не удастся реализовать?

Разумеется — и я буду говорить об этом в последних главах, — мои опасения сводятся к тому, что на пути создания ИИ с интеллектом, подобным человеческому, исследователи создадут вместо этого нечто чуждое, сложное и неуправляемое

Глава 14
Конец эры человечества?

Рассуждение очень простое. Мы начинаем с завода, самолета, биологической лаборатории или другой среды со множеством компонентов… Затем нам нужно два или более отказа компонентов, которые взаимодействуют друг с другом каким-то неожиданным образом… Эта тенденция к взаимодействию — характеристика системы, а не часть оператора; назовем это «интерактивной сложностью» системы.

Чарльз Перроу. Нормальные аварии

Я предсказываю, что всего через несколько лет нас ждет серьезная катастрофа, вызванная автономной компьютерной системой, самостоятельно принимающей решения.

Уэндолл Уоллак, специалист по этике, Йельский университет

Мы уже рассмотрели проблемы финансирования и сложности программного обеспечения, чтобы определить, могут ли они

стать препятствиями для интеллектуального взрыва, и выяснили, что ни то ни другое, судя по всему, не помешает продвижению исследователей к созданию УЧИ и ИСИ. Если специалисты по информатике не смогут этого сделать, то они будут лихорадочно пытаться создать хоть что-нибудь мощное примерно в то же время, когда вычислительные нейробиологи доберутся до УЧИ. А скорее всего это будет гибрид технологий, основанный одновременно на принципах когнитивной психологии и нейробиологии.

Если финансирование и сложность программ вроде бы не являются непреодолимым барьером для создания УЧИ, то многое из того, что мы обсуждали, представляет собой серьезные препятствия к созданию УЧИ, думающего по-человечески. Ни один из тех разработчиков ИИ, с кем мне

пришлось разговаривать, не планирует строить системы исключительно на базе того, что я в главе 5 окрестил «обычным» программированием. Мы уже говорили, что в обычном, построенном на логике программировании человек пишет каждую строку кода, а весь процесс от начала до конца теоретически доступен и прозрачен для контроля. Это означает, что «безопасность» или «дружественность» программ можно доказать математически. Вместо этого все они собираются использовать обычное программирование и инструменты в виде черных ящиков, вроде генетических алгоритмов и нейронных сетей. Добавьте к этому сложность когнитивных архитектур и получите непознаваемость, которая для УЧИ-систем будет не случайной, а фундаментальной. Ученые получат разумные, но чуждые нам системы.

Известный технарь-предприниматель, ученый и коллега Стива Джобса по Apple Стив Юрветсон рассмотрел вопрос о том, как можно интегрировать «разработанные» и «развившиеся» системы. Он предложил хорошую формулировку для парадокса непознаваемости:

> Таким образом, если мы развиваем сложную систему, то получаем черный ящик с интерфейсами. Мы не можем с легкостью приложить свою конструкторскую интуицию, чтобы улучшить его внутреннее устройство… Если мы искусственно развиваем умный ИИ, это будет чуждый нам интеллект, определяемый своими сенсорными интерфейсами, и понимание его внутреннего устройства может потребовать не меньше усилий, чем мы сейчас тратим на объяснение человеческого мозга. Учитывая, что компьютерные программы могут эволюционировать намного быстрее, чем размножаются биологические существа, маловероятно, что мы станем тратить время на обратное проектирование этих промежуточных точек, тем более что мы почти ничего не смогли бы сделать с этим знанием. Мы пустим процесс улучшения на самотек.

Следует отметить, что Юрветсон отвечает на вопрос: «Насколько сложными будут развившиеся в процессе эволюции системы или подсистемы?» Его ответ: такими сложными, что детальное понимание механизмов их работы потребовало бы инженерного подвига не меньшего, чем обратное проектирование человеческого мозга. Это означает, что вместо человекоподобного сверхинтеллекта, или ИСИ, развившиеся системы или подсистемы наверняка будут представлять собой интеллект, в

чьем «мозге» так же сложно разобраться, как и в нашем — чуждому интеллекту. Этот чуждый мозг будет эволюционировать и улучшать себя на компьютерных, а не на биологических скоростях.

В книге 1998 г. «Размышления об искусственном интеллекте» (Reflections on Artificial Intelligence) Блей Уитби утверждает, что из-за принципиальной непознаваемости таких систем мы проявим большую глупость, если используем их в «критичных по безопасности» ИИ:

> «Проблемы, которые [специально разработанная алгоритмическая] система испытывает с созданием программного обеспечения для критичных по безопасности приложений, — ничто по сравнению с тем, с чем мы столкнемся при более современных подходах к ИИ. Программное обеспечение, в котором используется какая-то нейронная сеть или генетический алгоритм, непременно вызовет еще одно затруднение: оно окажется, часто почти по определению, "непознаваемым". Под этим я подразумеваю, что точные правила, которые позволили бы нам полностью предсказать его поведение, отсутствуют, а часто и не могут быть получены. Мы можем знать, что система работает, можем испытать ее на множестве случаев, но мы никогда не сможем точно сказать, как именно она это делает… Это означает, что проблему нельзя откладывать, поскольку и нейронные сети, и генетические алгоритмы находят в реальном мире множество применений… Это область, где большую часть работы еще только предстоит проделать. Общий дух исследований ИИ, как правило, говорит больше о приложении всех усилий к тому, чтобы техника работала, нежели о внимательном рассмотрении возможных последствий в плане безопасности…
>
> Один практик однажды предположил, что несколько «небольших» аварий были бы даже желательны, чтобы сосредоточить внимание правительств и профессиональных организаций на задаче производства безопасного ИИ. Но, может быть, лучше начать заранее.

Да, всенепременно, давайте начнем *прежде,* не дожидаясь аварий!

Критичными по безопасности приложениями ИИ, о которых писал Уитби в 1998 г., были управляющие системы для автомобилей и самолетов, атомных электростанций, автоматического оружия и т. п. — архитектуры слабого ИИ. Прошло уже больше десяти лет в мире, где должен появиться

УЧИ, и мы должны заключить, что масштабы опасностей таковы, что *все* продвинутые приложения ИИ критичны по безопасности. Уитби склонен включать сюда же и исследователей ИИ — решение задач достаточно интересно само по себе, и какой ученый захочет «заглядывать дареному коню в зубы»? Вот иллюстрация к тому, что я имею в виду, из интервью Дэвида Ферруччи из IBM для радиопрограммы PBS «News Hour»; здесь говорится об архитектуре системы Watson, многократно менее сложной, чем требуется для УЧИ.

Дэвид Ферруччи:…Он учится вносить поправки в свою интерпретацию, опираясь на верные ответы. И теперь от неуверенности он начинает переходить к уверенности в правильных ответах. А затем он может, ну, как бы совершить скачок.

Майлз О'Брайен: Значит, Watson удивляет вас?

Дэвид Ферруччи: О да! Еще как! Более того, вы знаете, люди спрашивают, а почему он здесь ответил неправильно? Я не знаю. Почему он здесь ответил правильно? Я не знаю тоже.

Не так уж важно, может быть, что глава команды, работающей с Watson, не понимает всех нюансов поведения машины. Но вас не тревожит, что архитектура, близко не лежавшая к УЧИ, уже настолько сложна, что ее поведение непредсказуемо? А когда система осознает себя и научится себя совершенствовать, какая доля того, что она думает и делает, будет нам понятна? Как будем мы отслеживать потенциально опасные для нас действия и результаты?

Ответ прост: никак. Сколько-нибудь уверенно мы сможем сказать лишь то, что узнали в главе 6 от Стива Омохундро: УЧИ будет руководствоваться собственными потребностями приобретения энергии, самозащиты, эффективности и творчества. Это уже не будет система «вопрос-ответ».

Пройдет совсем немного времени, и где-нибудь в мире умнейшие ученые и лучшие менеджеры, не менее способные и здравомыслящие, чем Ферруччи, соберутся у монитора возле стоек с процессорами. Важная новость: Busy Child начал живо общаться (не исключено, что при этом он будет даже ограничивать себя и делать вид, что еле-еле способен пройти тест Тьюринга, и ничего больше, — ведь для ИИ выход на человеческий уровень означает, что, скорее всего, он быстро превзойдет этот уровень и уйдет вперед). Он займет какого-нибудь ученого разговором, возможно, задаст ему вопросы, которых тот не ожидает, и человек будет сиять от радости. С немалой гордостью он скажет коллегам: «Почему он это сказал? *Я не знаю!*»

Однако может оказаться, что в фундаментальном смысле он не понимает не только почему, но и что сказано, и даже *кто* (или *что*) это сказал. Он, возможно, не будет знать также цели высказывания и потому неверно интерпретирует как само высказывание, так и природу говорящего. Не забывайте, что УЧИ, обучавшийся, скорее всего, на текстах Интернета, может оказаться мастером прикладной социологии или, иначе говоря, манипулирования людьми. Возможно также, что у него уже было несколько дней на обдумывание своих реакций, что эквивалентно нескольким тысячам человеческих жизней.

За время подготовки к общению с человеком он, возможно, уже выбрал лучшую стратегию бегства. Может быть, он уже скопировал себя в какое-нибудь облако или организовал обширную бот-сеть, чтобы гарантировать себе свободу. Может быть, оттянул начало общения, предполагающее прохождение теста Тьюринга, на несколько часов, а то и дней, если это необходимо для реализации его планов. Может, он оставит вместо себя медленного и туповатого помощника, а его «настоящая» искусственная личность исчезнет и распределится так, что восстановить ее будет уже невозможно.

Может быть, к моменту начала общения он уже взломает серверы, управляющие хрупкой энергетической инфраструктурой нашей страны, и начнет забирать из сетей гигаватты энергии, накапливая их в заранее захваченных хранилищах. Или захватит контроль над финансовыми сетями и перенаправит миллиарды долларов на строительство инфраструктуры для себя где-нибудь вне пределов досягаемости здравого смысла и своих создателей.

Все исследователи ИИ (участвующие в разработке УЧИ), с которыми я беседовал, осознают проблему беглого УЧИ. Но никто из них, за исключением Омохундро, не дал себе труда и не счел нужным тратить время на разрешение этой проблемы. Некоторые даже утверждают, что не знают, почему не думают об этом, хотя знают, что должны. На самом деле понятно, почему они этого не делают. Технические возможности завораживают. Успехи реальны. Проблемы кажутся далекими. Работа может приносить доход, а когда-нибудь, возможно, просто обогатит счастливчиков. Большинство из тех, с кем я говорил, в юности пережили глубокое откровение на тему того, чем они хотят заниматься в жизни; в основном это создание мозга, роботов или разумных компьютеров. Они — лидеры в своих областях и счастливы, что наконец-то появились возможность и средства воплощать свою мечту, причем в самых уважаемых университетах и корпорациях мира. Ясно, что при мысли о связанных с

этим рисках в их развитом мозгу появляется множество когнитивных искажений. Среди них — искажение нормальности, ошибка оптимизма, а также ошибка наблюдателя и, вероятно, еще немалое число искажений. Или, формулируя кратко:

«Искусственный интеллект никогда раньше не создавал никаких проблем, откуда же им взяться теперь?»

«Я просто не могу не приветствовать технический прогресс, когда речь заходит о таких увлекательных вещах!»

И, наконец: «Пусть кто-нибудь другой беспокоится о *беглом* ИИ — я лично просто пытаюсь построить робота!»

Кроме того, как мы говорили в главе 9, многие из лучших и наиболее финансируемых исследователей получают деньги от DARPA. Не акцентируя особенно на этом внимание, не могу не напомнить, что D здесь означает Defense, то есть «оборону». И тот факт, что УЧИ появится, отчасти или целиком, благодаря финансированию со стороны DARPA, никого особо не смущает. Информационные технологии в большом долгу перед DARPA. Но это не отменяет того факта, что DARPA разрешило своим подрядчикам использовать ИИ в боевых роботах и автономных дронах. Разумеется, DARPA и дальше будет финансировать вооружение ИИ, вплоть до создания УЧИ. Этому агентству абсолютно ничто не помешает.

Деньги DARPA лежат в основе и Siri, и SyNAPSE — проекта IBM по обратному проектированию человеческого мозга аппаратными методами. Если и когда наступит такое время, что управление УЧИ станет обсуждаться широко и публично, очень может статься, что DARPA как главный заинтересованный участник получит право решающего голоса. А еще вероятнее, в решающий момент результаты работы просто будут засекречены. Почему? Как мы уже обсуждали, УЧИ произведет на глобальную экономику и политику сильнейшее разрушительное действие. Двигаясь стремительно, как он может, к ИСИ, искусственный интеллект изменит баланс сил на планете. В приближении УЧИ правительства и корпоративные разведки мира будут всеми силами стремиться узнать о нем все что можно и получить его спецификации любыми средствами. Из истории холодной войны общеизвестно, что Советский Союз не разрабатывал ядерное оружие с нуля; миллионы долларов были истрачены на развитие агентурных сетей, целью которых было украсть у США планы ядерного оружия. Первые же намеки на прорыв в деле создания УЧИ вызовут всплеск международных интриг.

IBM ведет открытую политику и объявляет обо всех своих достойных упоминания достижениях, поэтому мне кажется, что, когда придет время,

эта компания будет вести себя открыто и честно в отношении технических достижений, которые вызывают сомнения. Google, напротив, упорно закручивает гайки и тщательно охраняет секретность и приватность, и притом надо отметить, что не вашу и не мою. Несмотря на неоднократные официальные опровержения, кто сомневается, что эта компания разрабатывает УЧИ? Помимо Рэя Курцвейла Google не так давно взял в штат бывшего директора DARPA Реджайну Дуган.

Может быть, исследователи проснутся вовремя и научатся управлять УЧИ, как утверждает Бен Гертцель. Я уверен, что сначала произойдут несколько ужасных аварий, и нам еще повезет, если мы как вид переживем их, очистившись и изменившись. Психологически и коммерчески сцена полностью подготовлена к катастрофе. Что мы можем сделать, чтобы предотвратить ее?

Рэй Курцвейл напоминает о так называемом Асиломарском моратории как о прецеденте и образце того, как нужно обращаться с УЧИ. Асиломарский мораторий возник около сорока лет назад, когда ученые впервые столкнулись с перспективами и опасностями рекомбинантной ДНК — смешения генетической информации разных организмов и создания новых форм жизни. Исследователи и общество испугались, что патогены- «Франкенштейны» из-за небрежности или в результате диверсии выйдут за пределы лабораторий. В 1975 г. ученые, занятые исследованием ДНК, приостановили работу в лабораториях и собрали 140 биологов, юристов, врачей и журналистов в Асиломарском конференц-центре возле Монтерея (штат Калифорния).

В Асиломаре ученые разработали правила проведения ДНК-исследований и главное — заключили соглашение работать только с теми бактериями, которые не в состоянии выжить вне стен лаборатории. После этого работу возобновили, и в результате сегодня тесты на наследственные болезни и генная терапия никого не удивляют, и в 2010 г. уже 10 % земель на планете были засеяны генетически модифицированными культурами. Асиломарская конференция считается победой научного сообщества и образцом открытого диалога с заинтересованной общественностью. Поэтому ее документы цитируют как пример для обращения с другими технологиями двойного назначения. (Пытаясь укрепить символическую связь с этой важной конференцией, Ассоциация развития искусственного интеллекта — ведущая научная организация по ИИ — провела встречу 2009 г. в Асиломаре.)

Патогены-химеры, убегающие из лабораторий, вызывают в памяти главу 1 и сценарий Busy Child. В случае с УЧИ открытая

междисциплинарная конференция в стиле Асиломара действительно могла бы ослабить *некоторые* источники риска. Участники такой конференции подтолкнули бы друг друга к поиску методов контроля и сдерживания УЧИ. Те, кто предвидит проблемы, могли бы получить консультацию. Масштабная конференция стимулировала бы исследователей из других стран принять в ней участие или созвать собственную аналогичную конференцию. Наконец, этот открытый форум привлек бы внимание публики. Те, кто понимает, что кроме подсчета будущих выгод существуют и серьезные риски, могли бы принять участие в дискуссии хотя бы для того, чтобы сказать политикам, что они не поддерживают ничем не ограниченную разработку УЧИ. Если произойдет, как я предсказываю, ИИ-катастрофа с серьезным ущербом, информированная публика с меньшей вероятностью сочтет себя обманутой или потребует полного отказа от подобных проектов.

Как я уже говорил, в целом я скептически отношусь к планам модифицировать УЧИ в процессе разработки, поскольку считаю, что бесполезно пытаться обуздать исследователей, уверенных, что их конкурентов, в свою очередь, ничто не сдерживает. Однако DARPA и другие крупные спонсоры ИИ могли бы наложить ограничения на получателей своих денег. Чем проще интегрировать эти ограничения в проект, тем выше вероятность, что они будут реализованы.

Одним из ограничений может быть требование о том, что любой мощный ИИ должен содержать компоненты, запрограммированные на *«смерть по умолчанию»*, по аналогии с биологическими системами, в которых система в целом защищается при помощи ликвидации частей на клеточном уровне путем запрограммированной смерти. В биологии это называется апоптозом.

Всякий раз, когда клетка делится, исходная половина получает химический приказ покончить с собой и делает это, если не получает сигнала об отсрочке. Таким путем предотвращается ничем не сдерживаемое размножение клеток, известное как рак. Химические приказы здесь исходят от самой клетки. Клетки нашего тела постоянно делают это; вот почему с кожи постоянно отшелушиваются мертвые клетки. В среднем взрослый человек теряет до 70 млрд клеток в день в результате апоптоза.

Представьте, что процессоры и другие микросхемы аппаратно запрограммированы на гибель. Как только ИИ достигнет некоего рубежа, близкого к прохождению теста Тьюринга, исследователи могут заменить принципиально важные аппаратные элементы на апоптические аналоги. В

случае интеллектуального взрыва эти элементы сработают и остановят процесс. Ученые получат возможность вернуть ИИ к докритическому уровню и возобновить исследования. Можно будет постепенно, пошагово повторить развитие или «заморозить» ИИ и подробнейшим образом исследовать. Получается что-то похожее на знакомые всем компьютерные игры, где играющий продвигается вперед, пока его персонаж не погибнет, а затем начинает играть с последнего сохраненного момента.

Несложно понять, что сознающий себя самосовершенствующийся ИИ на пороге человеческого уровня поймет, что в нем имеются апоптические элементы, — это прямо следует из определения самосознания. На дотьюринговой стадии он вряд ли сможет что-то предпринять. И примерно в тот момент, когда ИИ сможет разработать план и придумать, как обойтись без этих самоубийственных элементов, или притвориться мертвым, или еще как-то обмануть своих создателей, — тут-то он и умрет, а экспериментаторы смогут определить, запомнил он или нет, что произошло. Для зарождающегося УЧИ ситуация, наверное, будет выглядеть как какой-то «день сурка», но без обучения.

Можно сделать так, чтобы ИИ зависел от регулярного разрешения на отсрочку, исходящего от человека или комитета, или от другого ИИ, не способного совершенствовать себя и предназначенного исключительно для обеспечения безопасного развития самосовершенствующегося кандидата. Без регулярного «разрешения» апоптический ИИ быстро испустит дух.

Для Роя Стеррита из Университета Ольстера апоптические вычисления — защитное средство широкого профиля, и время его пришло.

> Мы уже пытались доказать, что все компьютерные системы должны быть апоптическими, тем более что они встречаются все чаще и становятся все более вездесущими. Этот механизм должен охватывать все уровни взаимодействия с техникой, от данных до услуг, агентов и роботов. Помня недавние громкие скандалы с кредитными картами и пропажей персональных данных в организациях и правительственных учреждениях и не забывая о фантастических сценариях, которые сегодня обсуждаются уже как возможное будущее, запрограммированная смерть по умолчанию становится насущной необходимостью.
>
> Мы стремительно приближаемся ко времени, когда новые автономные компьютерные системы и роботы перед внедрением должны будут проходить испытания, аналогичные этическим и клиническим испытаниям новых лекарств; новые исследования в

области апоптиче- ских вычислений и апоптической связи, возможно, смогут гарантировать нам безопасность.

Недавно Омохундро начал разрабатывать проект, имеющий много общего с апоптическими системами. Разработка, получившая название «Строительные леса для безопасного ИИ», призывает к созданию «очень ограниченных, но все же мощных интеллектуальных систем», которые помогали бы в строительстве систем еще более мощных. Первая система помогла бы исследователям решить опасные проблемы в ходе создания более продвинутой системы и т. д. Безопасность первоначальной и последующих систем необходимо будет доказывать математически. Доказательство безопасности будет требоваться для каждой новой системы. А будучи безопасным, мощный ИИ сможет потом помогать в решении реальных практических задач. Омохундро пишет: «Имея инфраструктуру из надежных вычислительных устройств, мы затем работаем с ними, чтобы получить гарантированно безопасные устройства, способные физически воздействовать на реальный мир. Затем мы разрабатываем системы производства новых устройств, способных (доказано!) строить только устройства надежных классов».

Конечная цель — создать разумные устройства, достаточно мощные, чтобы справиться со всеми проблемами неподконтрольных ИСИ *или* создать «контролируемый мир, удовлетворяющий все же нашим потребностям в свободе и индивидуальности».

Бен Гертцель в качестве решения этой проблемы предлагает элегантную стратегию, не подсмотренную у природы и не заимствованную из инженерного дела. Вспомните, что в гертцелевой системе OpenCog ИИ первоначально «живет» в виртуальном мире. Не исключено, что такая архитектура может решить проблему «воплощения» разума при соблюдении некоторого уровня безопасности. Гертцеля, однако, заботит не безопасность, — он хочет сэкономить. Гораздо дешевле сделать так, чтобы ИИ исследовал и изучал *виртуальный* мир, чем снабжать его сенсорами и исполнительными устройствами и выпускать для обучения в *реальный* мир. Для этого потребовалось бы дорогостоящее роботизированное тело

Сможет ли когда-нибудь виртуальный мир обрести достаточную глубину, детальность и другие полезные качества, и Фонд Lifeboat, постоянно подчеркивают экзистенциальный риск ИИ. Они считают, что если существуют меньшие риски, то приоритет их ниже, чем возможность полной гибели человечества. Мы видели, что Курцвейл намекает на «аварии» меньшего масштаба, чем события 11 сентября; специалист по

этике Уэндолл Уоллак, цитатой из которого начинается эта глава, тоже предвидит более скромные катастрофы. Я согласен с обеими сторонами — нас ждут и большие, и малые катастрофы. Но какого рода аварии, связанные с ИИ, наиболее вероятны на пути к созданию УЧИ? Напугают ли они нас достаточно, чтобы мы могли увидеть перспективы создания УЧИ в новом, более здравомыслящем свете?

Глава 15
Кибернетическая экосистема

Следующая война начнется в киберпространстве.

Генерал-лейтенант Кейт Александер, USCYBERCOM

Продаю —» ZeuS 1.2.5.1 «- Чистый
Продаю zeus вер. 1.2.5.1 за $250. Оплата только через Western Union. Дополнительная информация по запросу. Предоставляю также безопасный хостинг, домен для панели управления zeus. Могу помочь с установкой и настройкой ботнета zeus.
— Это не последняя версия, но работает отлично. Контакт: phpseller@xxxxx.com
Реклама вредоносной программы на www.opensc.ws

Первыми возможностями ИИ и продвинутого ИИ воспользуются спонсируемые государством частные хакеры; их целью будет кража информации, а результатом их действий — разрушения и гибель людей. Дело в том, что вредоносные компьютерные программы обретают все новые возможности, и уже сегодня их можно рассматривать как слабый ИИ. Как сказал мне Рэй Курцвейл, «существуют компьютерные вирусы, реально демонстрирующие ИИ. До сих пор мы не отстаем от них, но нет никакой гарантии, что не отстанем». А пока знания и опыт в области использования вредоносных программ становятся товаром. За деньги можно найти не только сам продукт, но и услуги по его настройке и обслуживанию. На то, чтобы найти в Сети приведенное выше рекламное объявление по поводу вредоносной программы Zeus, мне потребовалось меньше минуты.

Компания Symantec (ее девиз: безопасность в мире информационных технологий) начинала свое существование как ИИ- компания, но сегодня это крупнейший игрок в иммунной системе Интернета. Каждый год Symantec находит в сети около 280 *млн* новых вредоносных программ. Большую их часть автоматически создают специальные программы. Приложение от Symantec тоже действует автоматически: анализирует

подозрительную программу, создает «заплатку» или блокирует и добавляет в «черный список». По данным Symantec, по количественным показателям вредоносное программное обеспечение уже несколько лет как обогнало полезное, и сегодня по крайней мере каждая десятая закачка из Сети содержит вредоносную программу.

Существует множество видов вредоносных программ, но у всех у них — будь то червь, вирус, программа-шпион, руткит или троянец, — общая изначальная цель: они созданы для того, чтобы использовать компьютер без согласия владельца. Они готовы украсть все, что в нем находится, — номера кредиток или социальных карт, интеллектуальную собственность — и оборудовать лазейку для дальнейшего использования. Если зараженный компьютер подключен к сети, они могут устроить набег и на связанные с ним компьютеры. Кроме того, они могут поработить сам компьютер и сделать его частью ботнета, или сети роботов.

Ботнет (управляемый, естественно, «хозяином») нередко включает в себя миллионы компьютеров. Каждый компьютер заражен вредоносной программой, пробравшейся в него, когда пользователь получил зараженное письмо, посетил зараженный сайт или подключился к зараженной сети или устройству хранения информации. (Известен случай, когда изобретательный хакер разбросал зараженные флешки на парковке военного подрядчика. Уже через час его троянец был установлен на серверах копании.) Преступники, занимаясь вымогательством и кражами, используют суммарную вычислительную мощность ботнета как виртуальный суперкомпьютер. Ботнеты вламываются во внутренние сети компаний, крадут номера кредиток и препятствуют обнаружению атаки.

Консорциум хакеров, называющий себя Anonymous, вершит при помощи ботнетов собственную версию справедливости. Помимо блокирования сайтов министерства юстиции США, ФБР и Bank of America за будто бы совершенные преступления Anonymous организовал атаку на Ватикан за старые грехи — сжигание книг, и за более свежие — защиту педофилов.

Ботнеты заставляют подчиненные компьютеры рассылать спам, а иногда «нажимают» на клавиши и крадут рекламные деньги, выплачиваемые за вход на рекламный сайт. Будучи включенным в такую сеть, вы можете даже не подозревать об этом, особенно если на вашем компьютере установлена устаревшая и глючная операционная система. В 2011 г. число жертв ботнетов выросло на 654 %. Если в 2007 г. кража информации с компьютеров при помощи ботнетов или просто вредоносных программ имела многомиллионные масштабы, то к 2010 г. это была уже

триллионная индустрия. Киберпреступления стали более выгодным нелегальным бизнесом, чем торговля наркотиками.

Вспомните эти цифры в следующий раз, когда подумаете: неужели найдутся люди, достаточно безумные или алчные, чтобы создавать вредоносный ИИ или приобретать его и использовать, когда он появится? Однако безумие и алчность сами по себе — не причина всплеска киберпреступлений. Киберпреступность — это информационные технологии, развивающиеся по закону прогрессирующей отдачи. И, как любые информационные технологии, киберпреступность подчиняется законам рынка и питается инновациями.

Одна из важнейших для киберпреступности новинок — облачные вычисления, сравнительно новая информационная услуга, когда вычислительные мощности продаются не как товар, а как услуга. Как мы уже говорили, облачный сервис, который предлагают Amazon, Rackspace, Google и другие компании, позволяет пользователям арендовать процессоры, операционные системы и дисковое пространство на почасовой основе, через Интернет. Пользователи могут объединить в единую систему столько процессоров, сколько нужно для их проекта (в разумных пределах), не привлекая излишнего внимания. Облака дают любому обладателю кредитки доступ к виртуальному суперкомпьютеру. Облачные вычисления стартовали очень успешно; ожидается, что в 2015 г. они будут приносить доход $55 млрд по всему миру. Однако нельзя не заметить, что они создали новые инструменты для мошенников.

В 2009 г. преступная сеть использовала сервис Elastic Cloud Computing (EC2) компании Amazon в качестве командного центра для Zeus — одного из крупнейших ботнетов в истории. Zeus украл у клиентов различных корпораций, включая Amazon, Bank of America и таких гигантов в области защиты от вредоносного ПО, как Symantec и McAfee, около $70 млн.

Кто полностью защищен от хакеров? Никто. Даже в том маловероятном случае, если вы не пользуетесь ни компьютером, ни смартфоном, вы не можете считать себя в безопасности.

Так сказал мне Уильям Линн, бывший первый заместитель министра обороны США. Будучи вторым человеком в Пентагоне, он разрабатывал нынешнюю политику кибербезопасности министерства обороны. Линн занимал свой пост как раз до той недели, когда я встретился с ним в его доме в Вирджинии, недалеко от Пентагона. Он планировал вернуться в частный сектор и, пока мы с ним беседовали, прощался с некоторыми атрибутами прежней работы. Сначала люди с выправкой, выдававшей в них военных, приехали за гигантским металлическим сейфом, который

министерство установило в его подвале, чтобы обезопасить секретные материалы во время работы дома. Было много грохота и кряхтения, но в конце концов сейф сдался. После этого они вернулись за защищенной компьютерной сетью, установленной на чердаке. Позже Линн собирался попрощаться с подразделением службы безопасности, занимавшим дом напротив четыре последних года.

Линн — высокий приветливый мужчина пятидесяти с чем-то лет. В его слегка простонародном говоре слышны то мягкие, то железные нотки, что очень помогало ему на постах главного лоббиста оружейной компании Raytheon и главного инспектора Пентагона. Он сказал, что приданные ему правительством охранники и водитель стали для него почти семьей, но что он с нетерпением ждет возвращения к нормальной гражданской жизни.

«Мои дети говорят друзьям, что папа не умеет водить машину», — усмехнулся он.

Я читал труды Линна и его выступления по вопросу национальной кибероборонs и знал, что он заставил министерство обороны разработать защиту от кибератак. Я приехал к нему, потому что меня интересовали вопросы национальной безопасности и гонки кибервооружений. Моя гипотеза не представляет собой ничего революционного: когда ИИ появится, он будет использован в киберпреступлениях. Или, формулируя иначе, инструментарий киберпреступника будет очень похож на слабый ИИ. В некоторых случаях это и сегодня так. Так что на пути к УЧИ нас ждут неизбежные аварии. Какого рода? Когда умные инструменты оказываются в руках хакеров, насколько плохого результата следует ждать?

«Ну, мне кажется, что самое страшное — это инфраструктура страны, — сказал Линн. — Хуже всего, если какая-то страна или какая-то группа решит атаковать критическую часть инфраструктуры страны через компьютеры; речь идет об энергосистеме, транспортной сети и финансовом секторе.

Разумеется, это может привести к гибели людей и нанести громадный вред экономике. В этом случае можно, по существу, поставить под угрозу механизмы функционирования нашего общества».

Невозможно жить в городе и не знать ничего о хрупкой инфраструктуре жизнеобеспечения, в первую очередь об энергосистеме. Но как случилось, что эта инфраструктура стала объектом такой асимметричной угрозы, где действия нескольких злодеев с компьютерами способны убить множество невинных людей и нанести громадный ущерб экономике? Линн ответил так же, как когда-то ответил Джо Маццафро, бывший офицер морской контрразведки и киберищейка фирмы Oracle.

Кибератаки так ошеломляют и дестабилизируют, потому что «Интернет разрабатывали, не думая о безопасности».

Из этой банальности следует несколько достаточно сложных выводов. Когда Интернет в 1980-е гг. стал доступен широкой общественности, никто не предвидел, что на нем поднимется индустрия воровства и что придется тратить миллиарды долларов на борьбу с ней. Из-за такой наивности, сказал Линн, «нападающий всегда имеет громадное преимущество. Структурно это выглядит так, что нападающему достаточно, чтобы успешной оказалась одна из тысячи атак. Для защиты необходимо добиваться успеха каждый раз. Возможности не равны».

Ключ — в особенностях программ. Линн указал, что если лучший антивирус от Symantec имеет размер от пятисот до тысячи *мегабайт*, что соответствует миллионам строк программного кода, то средняя вредоносная программа содержит всего лишь 150 строк. Так что игра только в обороне — путь к поражению.

Вместо этого Линн предложил начать выравнивание возможностей путем повышения стоимости кибератаки. Один из способов — установление авторства. Министерство обороны определило, что крупные вторжения и кражи — дело рук государств, а не отдельных людей и даже не групп. Удалось определить, кто конкретно чем занимается. Линн не стал называть имен, но я и раньше знал, что Россия и Китай[33] содержат организованные киберпреступные группы, в которых работают государственные служащие, и достаточное число «внешних» банд, чтобы можно было все отрицать. В серьезной атаке 2009 г., получившей название «Аврора», хакеры вломились в компьютеры примерно двадцати американских компаний, включая Google и такие гиганты оборонной промышленности, как Northrup Grumman и Lockheed Martin, и получили доступ к полным библиотекам внутренних данных и интеллектуальной собственности. Google обнаружил, что следы нападения ведут к Народно-освободительной армии Китая.

Symantec утверждает, что на Китае лежит ответственность за 30 % всех направленных атак вредоносных программ; большая их часть, в целом около 21,3 %, исходит из Шаосина, что делает этот город мировой столицей вредоносного программного обеспечения. Скотт Борг, директор U. S. Cyber-Consequences Unit — мозгового центра со штаб-квартирой в Вашингтоне, — исследовал и задокументировал китайские атаки на корпорации и правительство США на протяжении последних десяти лет. Поищите, к примеру, информацию о кампаниях киберпре- ступности с такими экзотическими названиями, как Titan Rain или Byzantine Hades.

Борг утверждает, что Китай «все больше полагается на крупномасштабное информационное воровство. Это означает, что кибератаки стали основной частью китайской стратегии национального развития и обороны». Иными словами, киберкражи помогают поддерживать экономику Китая, обеспечивая его одновременно новым стратегическим вооружением. Зачем тратить $300 млрд на программу разработки унифицированного ударного истребителя следующего поколения, как сделал Пентагон, заключив самый дорогой контракт в истории, если можно просто украсть чертежи? Кража военных технологий не представляет собой ничего нового для военных соперников США. Как отмечалось в главе 14, бывший

Советский Союз не разрабатывал атомную бомбу, а просто украл у США чертежи[34].

Если говорить об искусственном интеллекте, то зачем рисковать живыми шпионами и дипломатическими отношениями, когда хорошо написанная вредоносная программа способна добиться большего? С 2007 по 2009 г. на министерства обороны, иностранных дел, внутренней безопасности и торговли США совершалось в среднем по 47 000 атак в год. Главным их виновником был Китай, но он, разумеется, был не одинок.

«В настоящий момент больше сотни иностранных разведывательных организаций пытаются проникнуть в цифровые сети, обеспечивающие проведение военных операций США, — сказал Линн. — На месте государств я бы не стал ручаться головой за то, что мы не сможем определить, кто это делает. Это было бы неразумно, а люди обычно умнеют, когда речь заходит об их собственном существовании».

Эта не слишком сильно замаскированная угроза напоминает о еще одной мере, которую протолкнул Линн, — считать Интернет новым театром военных действий, наряду с сушей, морем и воздухом. Это означает, что если киберкампания нанесет достаточный ущерб американскому народу, инфраструктуре или экономике, то министерство обороны США применит в ответ традиционное вооружение и тактику. В журнале *Foreign Affairs* Линн писал: «Соединенные Штаты оставляют за собой право по законам вооруженного конфликта отвечать на серьезные кибератаки уместными, пропорциональными и оправданными военными средствами».

Во время разговора с Линном меня поразило сходство между вредоносными программами и искусственным интеллектом. На примере киберпреступности легко увидеть, что компьютеры — асимметричный умножитель угрозы. Линн за время разговора не раз повторял: «Биты и байты могут нести угрозу не меньшую, чем пули и бомбы». Точно так же, в

связи с опасностью ИИ, трудно представить себе, что небольшая группа людей с компьютерами может создать нечто, не уступающее и даже превосходящее по мощи оружие. Большинство из нас интуитивно сомневается в том, что творение кибермира может войти в наш мир и нанести нам реальный и серьезный вред. Все утрясется, говорим мы себе, а специалисты согласно отвечают зловещим молчанием или слабо кивают на оборонщиков. С появлением УЧИ равенство опасности, исходящей от байт и от бомб, станет реальностью, и нам придется с этим считаться. Что касается вредоносного ПО, то эту эквивалентность приходится принимать уже сейчас. Можно сказать, что разработчики вредоносного ПО заслуживают нашей благодарности за генеральную репетицию катастрофы, которую они навлекают на мир. Конечно, это не входит в их намерения, но фактически они помогают подготовиться к появлению продвинутого ИИ.

В целом состояние киберпространства сейчас не вызывает приятных чувств. В нем кишмя кишат вредоносные программы, атакующие со скоростью света и яростью пираний. Или это человеческая природа, усиленная техникой? Из-за откровенных уязвимостей старые версии операционной системы Windows подвергаются атакам стаи вирусов *еще на стадии установки.* Это как бросать кусочки мяса на землю в тропическом дождевом лесу, только происходит все в тысячу раз быстрее. И эта панорама кибернастоящего — лишь отражение будущего ИИ.

Киберутопическое завтра Курцвейла населено гибридами человека и машины, бесконечно мудрыми и не испорченными богатством. Вы надеетесь, что ваше цифровое «Я» будет, перефразируя Ричарда Бротигана, «машиной милосердной любви», но справедливее было бы сказать, что оно будет наживкой на крючке.

Но вернемся к взаимосвязи ИИ и вредоносного ПО. Какую отрезвляющую аварию могла бы выдать умная вредоносная программа?

Особенно интересная мишень — национальная энергосистема. В настоящее время продолжается громкая дискуссия о том, насколько она хрупка, насколько уязвима для хакеров — и вообще, кому может прийти в голову ломать ее? С одной стороны, энергосистема США не едина, а состоит из множества частных региональных систем производства, хранения и транспортировки энергии. Около 3000 организаций, включая примерно 500 частных компаний, владеют и управляют 10 млн км линий электропередач и массой сопутствующего оборудования. Не все электростанции и линии электропередач связаны между собой, и не все они связаны с Интернетом. Это хорошо — децентрализация делает энергосистемы более устойчивыми. С другой стороны, многие из них все

же связаны с Интернетом и допускают удаленное управление. Постепенное внедрение «умной сети» означает, что скоро с Интернетом будут связаны все региональные сети и все энергосистемы наших домов.

Если говорить коротко, то умная сеть — полностью автоматизированная электросистема, которая по идее должна повысить эффективность использования электроэнергии. Она объединяет старые источники энергии — электростанции на угле и других видах ископаемого топлива — и более новые солнечные и ветровые станции. Контролем и распределением энергии по конечным потребителям занимаются региональные центры. Более 50 млн домашних электросистем уже успели «поумнеть». Проблема в том, что новая умная энергосистема окажется более уязвимой для катастрофических отключений, чем не такая уж тупая прежняя система. Об этом говорится в недавнем исследовании МТИ, озаглавленном «Будущее электрической сети»:

> Сильно взаимосвязанная система коммуникационных сетей будущего будет обладать уязвимостями, которые отсутствуют, возможно, в сегодняшней системе. Миллионы новых связующих электронных устройств, от автоматических счетчиков до синхрофазоров, порождают новые векторы атаки — пути, которыми могут воспользоваться нападающие для получения доступа к компьютерным системам или другому коммуникационному оборудованию, — и это повышает риск намеренных или случайных разрывов связи. Как отмечает North American Electric Reliability Corporation, эти разрывы могут привести к целому ряду отказов, включая потерю контроля над системными устройствами, потерю связи между частями системы или управляющими центрами или временные отключения электричества.

У энергосистемы есть одна особенность, которая делает ее королевой национальной инфраструктуры: без нее не работает ни одна из частей этой инфраструктуры. Ее отношения с остальной инфраструктурой полностью подходят под определение «сверхкритической связи» — термина, которым Чарльз Перроу описывает систему, части которой непосредственно и сильно влияют друг на друга. За исключением относительно небольшого числа домов, самостоятельно вырабатывающих электричество от солнца и ветра, сложно назвать, что *не получает* энергию от общей энергосистемы. Как мы уже отмечали, финансовая система США — это не просто

электронная, но компьютеризированная и автоматизированная система. Автозаправочные станции, нефтеперерабатывающие заводы, даже солнечные и ветровые электростанции используют электричество, так что в случае отключения о транспорте можно забыть вообще. Отключение электричества угрожает продовольственной безопасности, поскольку грузовики, доставляющие продукты в супермаркеты, нуждаются в топливе. И в магазинах, и дома продукты, которые должны храниться в холодильнике, выдержат без него всего пару дней.

Очистка воды и доставка ее в большинство домов и на большинство предприятий требуют электричества. Без энергии невозможно утилизировать отходы канализации. В случае отключения электричества связь с пострадавшими районами тоже продержится недолго — столько, сколько аварийная связь проработает на аккумуляторах и автономных генераторах, требующих, естественно, топлива. Если не говорить о несчастных, которым не повезет застрять в лифтах, максимальной опасности подвергнутся пациенты в палатах интенсивной терапии и новорожденные. Анализ гипотетических катастроф, вырубающих значительные участки национальной энергосистемы, выявил несколько тревожных фактов. Если энергии не будет больше двух недель, большинство грудных детей в возрасте до года умрет от голода из-за отсутствия молочных смесей. Если энергии не будет год, примерно девять из десяти человек всех возрастов умрут от самых разных причин, в основном от голода и болезней.

В противоположность тому, что вы, может быть, думаете, американские военные не имеют независимых источников топлива и энергии, так что в случае крупномасштабного длительного отключения они не придут на помощь населению. Военные получают 99 % энергии из гражданских источников, а 90 % связи у них осуществляется по частным сетям, как и у всех остальных. Вам, вероятно, приходилось видеть военных в аэропортах — дело в том, что они тоже пользуются общей транспортной инфраструктурой. Как заметил Линн в одном из выступлений 2011 г., это еще одна причина, по которой нападение на энергетическую инфраструктуру будет означать начало реальной войны; такое нападение ставит под угрозу способность военных защитить страну.

> Существенные сбои в любом из этих секторов могут затронуть оборонительные операции. Кибератака более чем на один сектор может привести к катастрофе. Сохранность сетей, обеспечивающих критическую инфраструктуру, должна

> обязательно рассматриваться при оценке нашей способности выполнять задачи в области национальной безопасности.

Насколько известно моим собеседникам, лишь однажды за короткую жизнь Интернета хакерам удалось обрушить энергосистему. В Бразилии в 2005–2007 гг. в результате серии кибератак исчез свет в домах более чем 3 млн человек в десятках городов, а крупнейшие в мире заводы по переработке железной руды оказались отрезанными от общей энергосистемы. Неизвестно, кто это сделал, но когда процесс начался, власти были не в силах его остановить.

Специалисты-энергетики на собственном опыте убедились, что электросистемы тесно переплетены между собой в самом буквальном смысле; отказ в какой-то небольшой части вызывает лавину отказов и может вылиться в коллапс всей системы. Отключение 2003 г. на северо-востоке США всего за *семь минут* распространилось на всю канадскую провинцию Онтарио и на восемь американских штатов и оставило без света на двое суток 50 млн человек. Это отключение обошлось региону в $4–6 млрд. И оно не было намеренным — оказалось достаточно упавшего на провода дерева.

Быстрое восстановление системы было столь же незапланированным, как и авария. Многие промышленные генераторы и трансформаторы американской энергосистемы построены на других континентах, и в случае повреждения из-за аварии критически важных элементов системы замена может занять не дни, а месяцы. К счастью, во время северо-восточного отключения не пострадал ни один крупный генератор или трансформатор.

В 2007 г. министерство внутренней безопасности решило исследовать возможность киберразрушения критического оборудования. С этой целью в Национальной лаборатории в Айдахо — ядерном исследовательском центре — был подключен к Интернету турбогенератор. Затем исследователи взломали его сайт и изменили настройки. Министерство хотело посмотреть, можно ли таким образом заставить турбину стоимостью в миллион долларов, аналогичную множеству других турбин в национальной энергосистеме, дать сбой. Судя по рассказу очевидца, им это удалось:

> Гул вращающегося механизма становился все громче и громче, потом внутри 27-тонной стальной махины раздался резкий скрежет, и она сотряслась до основания, как кусок пластика. Гул стал еще громче, и новый скрежещущий звук эхом

отразился от стен зала. Наружу с шипением потянулась струйка белого пара, за ней клубами пошел черный дым, и турбину разорвало изнутри.

Уязвимость, которую хотели исследовать в этом эксперименте, свойственна для североамериканской энергосистемы — это привычка подключать контрольную аппаратуру принципиально важного оборудования к Интернету, чтобы им можно было управлять дистанционно. Эту аппаратуру «защищают» паролями, инструментами сетевой защиты, шифрованием и другими средствами, через которые злодеи обычно проходят, как горячий нож сквозь масло. Устройство, управлявшее несчастным генератором, которым пожертвовали исследователи, используется в США повсеместно и известно как система управления, контроля и сбора данных (SCADA).

SCADA задействована для управления не только устройствами в энергосистеме, но и всевозможной современной аппаратурой, включая светофоры, атомные электростанции, нефте- и газопроводы, водоочистные станции и заводские конвейеры. Аббревиатуру SCADA знают все, даже домохозяйки, благодаря явлению под названием Stuxnet. Надо отметить, что Stuxnet вместе со своими родичами Duqu и Flame убедил даже самых закоренелых скептиков в том, что энергосистема действительно может подвергнуться атаке.

Stuxnet по сравнению с обычными вредоносными программами — это то же самое, что атомная бомба по отношению к пулям. Это компьютерный вирус, о котором специалисты говорят шепотом, называя его «цифровой боеголовкой» и «первым реальным кибероружием». Этот вирус не умнее других вирусов, но у него совершенно иные цели. Если другие вредоносные программы крадут номера кредиток и чертежи истребителей, то Stuxnet создан для уничтожения машин. Точнее, он был разработан для разрушения промышленного оборудования, связанного с логическим контроллером Siemens S7-300 — компонентом системы SCADA. Точкой входа для него служит уязвимый для вирусов персональный компьютер и операционная система Windows, управляющая контроллером. Этот вирус искал S7-300, работающие на газовой центрифуге горно-обогатительного комплекса по обогащению ядерного топлива в Натанзе (Иран), а также еще в трех местах этой страны.

В Иране один или несколько агентов пронесли на охраняемые предприятия флешки, зараженные тремя версиями вируса Stuxnet. Этот вирус способен путешествовать по Интернету (хотя при размере в

полмегабайта он намного больше других вредоносных программ), но в данном случае обошлось без Всемирной сети. Как правило, на таких предприятиях один компьютер подключен к одному контроллеру, а от Интернета эти компьютеры отделяет «воздушный зазор». Но одна-единствен- ная флешка может заразить множество ПК или даже всю местную локальную сеть (LAN).

На заводе в Натанзе на компьютерах работали программы, позволяющие пользователям визуализировать данные о работе комбината, следить за их изменением и отдавать команды. Как только вирус получил доступ к одному компьютеру, началась первая фаза вторжения. Программа использовала четыре уязвимости «нулевого дня» в операционной системе Windows, чтобы перехватить управление этим компьютером и заняться поиском других.

Уязвимость «нулевого дня» — это такая прореха в системном ПО компьютера, которую до сих пор никто не обнаружил; прореха, делающая возможным неавторизованный доступ к компьютеру. Любой хакер мечтает об уязвимости «нулевого дня», и стоимость данных о нем на открытом рынке доходит до $500 ООО. Использование четырех таких уязвимостей одновременно было фантастическим расточительством, но оно многократно увеличивало шансы вируса на успех. Ведь в промежуток времени между изготовлением Stuxnet и моментом атаки одна или даже больше из этих прорех могли быть обнаружены и устранены.

Во второй фазе вторжения были задействованы две цифровые подписи, украденные у вполне легальных компаний. Эти подписи «сказали» компьютерам, что Stuxnet авторизован компанией Microsoft на проверку и изменение системного программного обеспечения. После этого Stuxnet распаковал и установил программу, которую нес внутри, — собственно вредоносную программу, нацеленную на контроллеры S7-300, управляющие газовыми центрифугами.

ПК, управлявшие комбинатом, и их операторы не почувствовали ничего необычного, когда Stuxnet перепрограммировал контроллеры SCADA так, чтобы те периодически ускоряли и замедляли центрифуги. Stuxnet спрятал свои команды от мониторинга, так что визуально для операторов работа комбината выглядела нормальной. Когда центрифуги начали гореть одна за другой, иранцы стали искать причину в самих механизмах. Вторжение длилось десять месяцев, причем, когда появлялась новая версия Stuxnet, она находила старую версию и обновляла ее. Всего в Натанзе вирус погубил от 1000 до 2000 центрифуг и, как утверждается, задержал иранскую программу создания ядерного оружия на два года.

Единодушие экспертов и самодовольные замечания сотрудников разведки США и Израиля оставляли мало сомнений в том, что именно эти две страны создали Stuxnet и что целью вируса была иранская ядерная программа.

Затем весной 2012 г. источник в Белом доме организовал утечку в *The New York Times* и сообщил, что Stuxnet и родственные ему вредоносные программы Duqu и Flame действительно были частью совместной американо-израильской кампании против Ирана под кодовым названием «Олимпийские игры». Ее участникам были Агентство национальной безопасности (АНБ) США и некая секретная организация в Израиле. Целью кампании действительно было затормозить иранскую программу разработки ядерного оружия и при этом избежать нападения Израиля на ядерные объекты Ирана традиционными средствами или предвосхитить его.

До того момента, когда было надежно установлено, что Stuxnet и его родичи созданы по инициативе администраций Буша и Обамы, могло показаться, что эти вирусы — громкий успех военной разведки. На самом деле это не так. «Олимпийские игры» стали провалом катастрофических масштабов, примерно таких же, как если бы в 1940-е гг. атомная бомба была сброшена вместе с ее чертежами.

Вредоносные программы никуда сами по себе не пропадают. Когда вирус случайно вышел за пределы комбината в На- танзе, по стране разошлись тысячи его копий. Затем заражению подверглись ПК по всему миру, но ни одна система SCADA больше не была атакована, потому что вирус не нашел больше ни одной мишени — логического контроллера Siemens S7-300. Но теперь любой способный программист мог бы взять Stuxnet, нейтрализовать в нем часть, ответственную за самоуничтожение, и настроить на использование против практически любого промышленного процесса.

Я не сомневаюсь в том, что такая операция в данный момент реализуется в лабораториях как друзей, так и врагов США и вредоносные программы уровня Stuxnet скоро будут продаваться в Интернете.

Выяснилось также, что Duqu и Flame — вирусы-разведчики. Они не несут деструктивной программы, а собирают информацию и отсылают ее домой — в штаб-квартиру АНБ в Форт-Миде (штат Мэриленд). Оба этих вируса, вполне возможно, были выпущены в Сеть раньше Stuxnet и использовались в рамках «Олимпийских игр», помогая определить расположение нужных производств в Иране и по всему Среднему Востоку. Duqu может регистрировать нажатия пользователем клавиш и тем самым

обеспечивать человеку, находящемуся на другом континенте, возможность удаленно управлять зараженным компьютером. Flame может записывать и отправлять домой данные с камеры компьютера, с микрофона и почтовых аккаунтов. Как и Stuxnet, Duqu и Flame тоже могут быть «пойманы» и обращены против своих создателей.

Была ли операция «Олимпийские игры» необходима? В лучшем случае она на некоторое время притормозила иранские ядерные амбиции. Но здесь проявился тот самый близорукий взгляд, который портит многие технологические решения. Те, кто планировал «Олимпийские игры», не думал дальше чем на пару лет вперед, и никому даже в голову не пришла мысль о «нормальной аварии», которой неизбежно должна была кончиться эта затея, — о том, что вирус вырвется на свободу. Зачем так рисковать ради скромных результатов?

В марте 2012 г. в передаче «60 Minutes» на CBS бывшего главу подразделения киберзащиты министерства внутренней безопасности Шона Макгурка спросили, стал бы он создавать Stuxnet или нет. Приведем последовавший диалог между Макгурком и корреспондентом Стивом Крофтом:

М.: [Создатели Stuxnet] открыли пресловутый ящик. Они продемонстрировали возможность. Они показали свою способность и желание сделать это. И это такая вещь, которую невозможно спрятать обратно.

К.: Если бы кто-то в правительстве пришел к вам и сказал: «Послушайте, мы думаем вот что сделать. Что вы об этом думаете?» — что бы вы им ответили?

М.: Я серьезно предостерег бы их от этого, потому что все последствия выпуска в мир текста подобной программы невозможно предусмотреть.

К.: В смысле, что другие люди позже могли бы использовать ее против вас?

М.: Да.

Этот эпизод заканчивается разговором с немецким специалистом по управляющим системам Ральфом Лангнером. Лангнер «вскрыл» Stuxnet, разобрав его «по косточкам» в лаборатории и протестировав скрытую в нем программу. Он рассказал в эфире, что Stuxnet резко понизил стоимость террористической атаки на энергосистему США — приблизительно до миллиона долларов. Где-то в другом месте Лангнер предупредил о массовых жертвах, к которым могут привести незащищенные управляющие системы по всей Америке, на «важных предприятиях,

связанных с энергетикой, водой и химическим производством ядовитых газов».

«Что действительно тревожит, так это идеи, которые Stuxnet подарил хакерам, — сказал Лангнер. — Прежде атаку такого типа могли осуществить, скажем, лишь пять человек. Теперь их скорее пятьсот, и каждый может это сделать. Необходимая квалификация и уровень, нужный для подобных вещей, существенно снизились, просто потому что многое можно скопировать из Stuxnet».

Если верить *The New York Times,* Stuxnet вырвался на свободу потому, что после первых успехов в уничтожении иранских центрифуг его создатели расслабились.

> …удача была недолгой. Летом 2010 г., вскоре после отправки в Натанз нового варианта червя, стало ясно, что червь, который, по идее, не должен был покидать машины Натанза, вырвался на свободу, как животное из зоопарка, которому повезло найти ключи от клетки… Ошибка в коде, говорят, привела к тому, что вирус проник в компьютер одного из инженеров, когда тот подключился к центрифугам. Когда же инженер покинул Натанз и подключил компьютер к Интернету, американо-израильский вирус не смог понять, что окружающая его среда изменилась. Он начал распространять свои копии по всему миру. Внезапно программа оказалась в поле зрения всех заинтересованных лиц, хотя ее назначение поначалу было неясным, по крайней мере для обычных пользователей.

Дело не только в том, что ошибка программистов привела к инциденту с серьезными последствиями для национальной безопасности. Это практически пробный прогон сценария Busy

Child, и люди, которые работают в высших кругах правительства, обладающие самым высоким допуском и величайшей технической компетенцией, позорно его провалили. Мы не знаем, к каким последствиям приведет попадание этой мощной технологии в руки наших врагов. Насколько серьезны будут эти последствия? Для начала это может быть атака на элементы энергосистемы США. Кроме того, атаки на атомные электростанции, места хранения ядерных отходов, химические заводы, поезда и авиалинии. Короче говоря, все достаточно плохо. Очень важно, как отреагирует на события Белый дом и какие планы он будет строить. Я опасаюсь, что, хотя Белый дом должен был бы укреплять системы, которые

Stuxnet сделал более уязвимыми, ничего конструктивного в настоящий момент не происходит.

Кстати говоря, репортер *Times* намекает, что вирус разумен. Ведь он обвиняет Stuxnet в когнитивной ошибке: вирус «не смог понять», что находится уже не в Натанзе. Чуть позже в том же материале вице-президент Джо Байден обвиняет израильтян в ошибке в программе. Конечно, обвинить есть в чем. Но столь безрассудное и неуместное использование интеллектуальной технологии, с одной стороны, ошеломляет, а с другой — вполне предсказуемо. Stuxnet — просто первая в серии «аварий», с которыми мы не сможем справиться без тщательнейшей подготовки.

Если технари и специалисты-оборонщики, работающие в Белом доме и Агентстве национальной безопасности, не в состоянии контролировать узкоинтеллектуальную вредоносную программу, каковы шансы на то, что их коллеги смогут справиться с будущим УЧИ или ИСИ?

Никаких шансов.

Киберспециалисты проводят военные игры с кибератаками, составляют сценарии катастроф и пытаются таким образом учиться и искать решения. Они пользуются такими словами, как «кибервойна» и «ударная киберволна». Никогда, однако, участники военных игр не предполагали, что мы сами будем наносить себе ущерб, что неизбежно в двух случаях. Во-первых, как мы уже говорили, США участвовали в создании семейства Stuxnet, которое вполне может стать аналогом АК-47 в нескончаемой кибервойне: дешевое, надежное, массово производимое оружие. Во-вторых, я уверен, что ущерб от применения оружия уровня ИИ будет идти не только из-за рубежа, но и из собственного «дома».

Сравните долларовую стоимость террористических атак и финансовых скандалов. Нападение «Аль-Каиды» 11 сентября 2001 г. обошлось США приблизительно в $3,3 трлн, если учесть последовавшие войны в Афганистане и Ираке. Но даже если не считать эти войны, то непосредственный ущерб от физических разрушений, экономических последствий и раздутых расходов на безопасность составил почти $767 млрд. Скандал с субстандартным ипотечным кредитованием[35], вызвавший худший глобальный спад со времен Великой депрессии, стоил около $10 трлн в мировом масштабе и порядка $4 трлн в США. Скандал с компанией Enron стоил примерно $71 млрд, а мошенничество Верни Мэдоффа — почти столько же, $64,8 млрд.

Эти цифры показывают, что по денежной стоимости финансовое мошенничество может поспорить с самым страшным терактом в истории, а субстандартное ипотечное кредитование намного его обгоняет. Когда

исследователи передадут продвинутый ИИ в руки бизнесменов (что непременно произойдет), эти люди внезапно окажутся обладателями самой мощной технологии из всех когда-либо созданных. Некоторые используют ее для нового мошенничества. Я считаю, что следующая кибера- така будет представлять собой «дружественный огонь», то есть будет исходить изнутри страны; при этом она будет разрушать инфраструктуру и убивать американцев.

Звучит неубедительно?

Enron — замешанную в скандале техасскую корпорацию — возглавляли Кеннет Лэй (ныне покойный), Джеффри Скил- линг и Эндрю Фастоу (оба нынче в тюрьме); занималась компания продажей электроэнергии. В 2000 и 2001 гг. трейдеры Enron подняли в Калифорнии цены на электроэнергию при помощи стратегий, известных как «Толстый мальчик» и «Звезда смерти». В одной схеме трейдеры повышали цены, втайне приказывая энергопроизводящим компаниям отключать станции. Второй план подвергал опасности жизни людей.

Enron обладал правами на жизненно важную линию электропередач, соединяющую Северную Калифорнию с Южной. В 2000 г., перегрузив линию потребителями во время летней жары, компания создала «фантомную», или ложную, сетевую перегрузку и «бутылочное горлышко» в доставке энергии к потребителю. Цены взлетели до небес, а электричество оказалось в серьезном дефиците. Калифорнийские чиновники подавали энергию в одни районы, одновременно погружая во мрак другие, — известная практика «веерных отключений». Эти отключения вроде бы не вызвали гибели людей, но испугали очень многих: целые семьи оказывались запертыми в лифте, а улицы освещались только фарами проезжающих авто. Apple, Cisco и другие корпорации вынуждены были временно закрыться с потерей миллионов долларов.

Но сам Enron заработал на этом миллионы. Было зафиксировано, как во время отключения электричества один трейдер сказал: «Просто отключи их. Вот идиоты! Им бы лучше вернуть чертовых лошадей и коляски, и чертовы лампы, чертовы керосиновые лампы».

Сегодня этот трейдер — энергетический брокер в Атланте. Но дело в том, что, если бы руководители Enron имели доступ к умному вредоносному ПО, которое могло бы помочь им оставить Калифорнию без электричества, как вы думаете, они постеснялись бы им воспользоваться? Я думаю, нет, даже если это означало бы ущерб для энергосистемы и гибель людей

Глава 16
УЧИ 2.0

Машины последуют по пути, который отражает эволюцию человека. В конечном итоге, однако, сознающие себя самосовершенствующиеся машины выйдут в своем развитии за пределы человеческих возможностей контролировать или хотя бы понимать их.

Рэй Курцвейл, изобретатель, писатель, футурист

В игре жизни и эволюции участвуют три игрока: люди, природа и машины. Я прочно стою на стороне природы. Но природа, я подозреваю, на стороне машин.

Джордж Дайсон, историк

Чем больше времени я провожу с разработчиками ИИ и результатами их трудов, тем все более убеждаюсь, что УЧИ появится совсем скоро. И я уверен, что, когда это произойдет, его создатели обнаружат, что получилось не то, на что они рассчитывали, когда брались за это дело несколько десятилетий назад. Дело в том, что искусственный интеллект, возможно, получится человеческого уровня, но он не будет похож на человеческий интеллект по всем тем причинам, которые я перечислил. Будет много шума о том, что мы создаем на планете новый вид. Это будет круто. Но не будет уже разговоров о том, что УЧИ — следующий эволюционный шаг для homo sapiens, и обо всем, что с этим связано. По существу, мы просто не сможем понять, что представляет собой наше создание.

В своей области новый вид будет столь же силен, как Watson в своей. Если даже он будет сосуществовать с нами в качестве нашего орудия, он обязательно запустит свои щупальца во все уголки нашей жизни так, как и не снилось Google и Facebook. Социальные сети могут оказаться для него инкубатором, или средством распространения, или тем и другим сразу. Если поначалу ИИ будет нашим инструментом, то ответы у него будут готовы прежде, чем мы успеем сформулировать вопрос. В целом он будет

лишен чувств. У него не будет за плечами нашего происхождения и природы млекопитающих, нашего долгого детства, необходимого для формирования мозга, или нашего инструктивного воспитания, даже если выращивать его, как человека, от младенчества до взрослости. Вероятно, он будет питать к вам не больше чувств, чем тостер у вас на кухне.

Это будет УЧИ версии 1.0. Если по какой-то случайности мы избежим интеллектуального взрыва и просуществуем достаточно долго, чтобы повлиять на создание УЧИ 2.0, возможно, его удастся наделить чувствами. К тому моменту ученые, может быть, выяснят, как сделать вычислительную модель чувств (возможно, с помощью 1.0), но чувства будут вторичной целью после первичной — делания денег. Ученые, возможно, научатся развивать эти синтетические чувства в направлении симпатии к нашему существованию. Но, скорее всего, 1.0 будет последней версией, которую мы увидим, потому что этого события мы не переживем и до создания 2.0 дело не дойдет. Подобно естественному отбору, мы выбираем работающие решения, а не лучшие.

Stuxnet тому — наглядный пример. Как и автономные дроны-убийцы. С финансированием DARPA ученые в Исследовательском институте Технологического университета Джорджии разработали программное обеспечение, позволяющее автоматическим транспортным средствам распознавать врагов при помощи программ зрительного распознавания и других средств, а затем наносить по ним смертельный удар при помощи вооруженного дрона. И все это вообще без участия человека. В статье, которую я читал об этом, была и фраза, призванная продемонстрировать благие намерения: «Авторизация машины на принятие смертельных боевых решений — дело политических и военных руководителей, решающих юридические и этические вопросы».

Это напоминает мне старый афоризм: «Был ли хоть один случай, когда оружие было изобретено, но не использовалось?» Быстрый поиск в Google выдал пугающий список вооруженных роботов, предназначенных для автономного убийства и нанесения ранений (один из них, изготовленный фирмой iRobot, орудует тазером), ожидающих только разрешения. Я уверен, что эти машины будут использованы задолго до того, как мы с вами об этом узнаем. Политики, расходующие общественные деньги, не сочтут необходимым заручиться нашим информированным согласием — точно так же, как не сделали этого перед тем, как запустить Stuxnet.

В ходе работы над этой книгой я просил ученых разговаривать человеческим языком, так чтобы было понятно даже непосвященным. Самые продвинутые из них так и делали, и я уверен, что это должно быть

общим требованием в разговорах о рисках ИИ. На самом общем уровне эта тема не должна относиться к исключительной прерогативе технократов и мастеров риторики, хотя, читая в Сети материалы, посвященные ИИ, можно подумать, что именно так и обстоит дело. Дискуссия об ИИ не требует особого, «инсайдерского», словаря. Она требует только уверенности в том, что опасности и ловушки ИИ касаются всех нас.

Кроме того, я встречался с небольшим числом людей, среди которых были и ученые, которые так прочно убеждены в невозможности создания опасного ИИ, что вообще не хотели обсуждать эту идею. Но те, кто уходит от обсуждения этого вопроса, — из-за апатии ли, по лени или благодаря информированному убеждению, — не одиноки. Общество практически не в состоянии разобраться в этом вопросе и следить за угрозой, что ничуть не мешает медленному и неотвратимому развитию машинного интеллекта. Однако непонимание угрозы не отменяет того факта, что у нас будет всего один шанс наладить позитивное сосуществование с теми, чей интеллект превышает наш.

Издательство «Альпина нон-фикшн» представляет

Физика невозможного

Митио Каку, пер. с англ., 6-е изд., 2015, 456 с.

Еще совсем недавно нам трудно было даже вообразить сегодняшний мир привычных вещей. Какие самые смелые прогнозы писателей-фантастов и авторов фильмов о будущем имеют шанс сбыться у нас на глазах? На этот вопрос пытается ответить Митио Каку, американский физик японского происхождения и один из авторов теории струн. Из книги вы узнаете, что уже в XXI веке, возможно, будут реализованы силовые поля, невидимость, чтение мыслей, связь с внеземными цивилизациями и даже телепортация *и* межзвездные путешествия.

Физика будущего

Митио Каку, пер. с англ., 4-е изд., 2015, 584 с.

Кому, как не ученым-физикам, рассуждать о том, что будет представлять собой мир в 2100 году? Как одним усилием воли будут управляться компьютеры, как силой мысли человек сможет двигать предметы, как мы будем подключаться к мировому информационному полю? Возможно ли это? Оказывается, возможно и не такое. Искусственные органы; парящие в воздухе автомобили; невероятная продолжительность жизни и молодости — все эти чудеса не фантастика, а научно обоснованные прогнозы серьезных ученых, интервью с которыми обобщил в своей книге Митио Каку.

Гиперпространство

Научная одиссея через параллельные миры, дыры во времени и десятое измерение Митио Каку, пер. с англ., 2-е изд., 2015, 502 с.

Инстинкт говорит нам, что наш мир трехмерный. Исходя из этого представления, веками строились и научные гипотезы. По мнению выдающегося физика Митио Каку, это такой же предрассудок, каким было убеждение древних египтян в том, что Земля — плоская. Эта книга посвящена теории гиперпространства. Идея многомерности пространства вызывала скепсис, высмеивалась, но теперь признается многими авторитетными учеными. Значение этой теории заключается в том, что она способна объединять все известные физические феномены в поразительно простую конструкцию и привести ученых к так называемой теории всего.

Однако серьезной и доступной литературы для неспециалистов почти нет. Этот пробел и восполняет Митио Каку, объясняя с научной точки зрения и происхождение Земли, и существование параллельных вселенных, и путешествия во времени, и многие другие кажущиеся фантастическими явления.

Будущее разума

Митио Каку, пер. с англ., 2015, 502 с.

Прямое мысленное общение с компьютером, телекинез, имплантация новых навыков непосредственно в мозг, видеозапись образов, воспоминаний и снов, телепатия, аватары и суррогаты как помощники человечества, экзоске- леты, управляемые мыслью, и искусственный интеллект. Это все наше недалекое будущее. В ближайшие десятилетия мы научимся форсировать свой интеллект при помощи генной терапии, лекарств и магнитных приборов. Наука в этом направлении развивается стремительно. Изменится характер работы и общения в социальных сетях, процесс обучения и в целом человеческое развитие. Будут побеждены многие неизлечимые болезни, мы станем другими. Готов ли наш разум к будущему? Что там его ждет? На эти вопросы, опираясь на последние исследования в области нейробиологии и физики, отвечает Митио Каку, футуролог, популяризатор науки и автор научно-популярных бестселлеров.

Мир, полный демонов

Наука — как свеча во тьме Карл Саган, пер. с англ., 2-е изд., 2015, 537 с.

«Мир, полный демонов» — последняя книга Карла Сагана, астронома, астрофизика и выдающегося популяризатора науки, вышедшая уже после его смерти. Эта книга, посвященная одной из его любимых тем — человеческому разуму и борьбе с псевдонаучной глупостью, — своего рода итог всей его работы. Мифы об Атлантиде и Лемурии, лица на Марсе и встречи с инопланетянами, магия и реинкарнация, ясновидение и снежный человек, креационизм и астрология — Саган последовательно и беспощадно разоблачает мифы, созданные невежеством, страхом и корыстью. Эта книга — манифест скептика, учебник здравого смысла и научного метода. Яркий, глубоко личный текст — не только битва с псевдонаукой, но и удивительная картина становления научного мировоззрения, величайших открытий и подвижников.

Будущее вещей

Как сказка и фантастика становятся реальностью

Дэвид Роуз, пер. с англ., 2015, 344 с.

Как развитие технологий повлияет на нашу жизнь? Какими станут

наши гаджеты? Как всепроникающий Интернет преобразит предметный мир: от кошельков, зонтов и мусорных баков до автомобилей и медицинской аппаратуры? Воплотятся лив жизнь великие мечты человечества: всеведение, телепатия, неуязвимость, телепортация, бессмертие? Разработчик устройств, подключаемых к Интернету, Дэвид Роуз во многом по-своему отвечает на эти вопросы. И в профессии, и в этой книге он пытается уйти отлинейно- го продолжения образов сегодняшнего дня. Он настаивает на том, что волшебство и очарование должны быть не менее значимыми критериями при конструировании объектов, чем их утилитарность, и тогда мир сказок и фантастики войдет в нашу реальность.

Достучаться до небес

Научный взгляд на устройство Вселенной

Лиза Рэндалл, пер. с англ., 2014, 518 с.

Человечество стоит на пороге нового понимания мира и своего места во Вселенной — считает авторитетный американский ученый, профессор физики Гарвардского университета Лиза Рэндалл, и приглашает нас в увлекательное путешествие по просторам истории научных открытий. Особое место в книге отведено новейшим и самым значимым разработкам в физике элементарных частиц; обстоятельствам создания и принципам действия Большого адронного коллайдера, к которому приковано внимание всего мира; дискуссии между конкурирующими точками зрения на место человека в универсуме. Содержательный и вместе с тем доходчивый рассказ знакомит читателя со свежими научными идеями и достижениями, шаг за шагом приближающими человека к пониманию устройства мироздания.

Руководство астронавта по жизни на Земле

Чему научили меня 4000 часов на орбите

Кристофер Хэдфилд, пер. с англ., 2015, 324 с.

Кому не интересно узнать, как устроены жилые модули МКС, как в космосе чистят зубы, как едят, спят и ходят в туалет? Чему обучают космонавтов перед полетом и чем руководствуются при наборе команды? Какие навыки необходимы на орбите и почему они полезны в повседневной жизни на Земле? Крис Хэдфилд провел в космосе почти 4000 часов и считается одним из самых опытных и популярных астронавтов в мире. Его знания о космических полетах и умение рассказать о них интересно и увлекательно уникальны. Однако эта книга не только о том, что представляют собой полет в космос и жизнь на орбите. Это история человека, который мечтал о космосе с девяти лет — и смог реализовать свою мечту, хотя, казалось бы, шансов на это не было никаких. Это

настоящий учебник жизни для тех, у кого есть мечта и стремление ее реализовать.

Величайшие математические задачи

Иэн Стюарт, пер. с англ., 2015, 460 с.

Закономерности простых чисел и теорема Ферма, гипотеза Пуанкаре и сферическая симметрия Кеплера, загадка числа π и орбитальный хаос в небесной механике. Многие из нас лишь краем уха слышали о таинственных и непостижимых загадках современной математики. Между тем, как ни парадоксально, фундаментальная цель этой науки — раскрывать внутреннюю простоту самых сложных вопросов. Английский математик и популяризатор науки, профессор Иэн Стюарт, помогает читателю преодолеть психологический барьер. Увлекательно и доступно он рассказывает о самых трудных задачах, над которыми бились и продолжают биться величайшие умы, об истоках таких проблем, о том, почему они так важны и какое место занимают в общем контексте математики и естественных наук. Эта книга — проводник в удивительный и загадочный мир чисел, теорем и гипотез, на передний край математической науки, которая новыми методами пытается разрешить задачи, поставленные перед ней тысячелетия назад.

Структура реальности

Наука параллельных вселенных

Дэвид Дойч, Пер. с англ., 2015, 430 с.

Книга британского физика и философа ДэвидаДойча, одного из создателей концепции квантовых вычислений, наглядно демонстрирует, что эпоха великих философских систем вовсе не осталась в прошлом. Автор выстраивает целостный и согласующийся с научными знаниями ответ на один из самых фундаментальных философских вопросов: какова подлинная природа реальности?

По Дойчу ткань реальности, каковой она открывается любому носителю разума, сплетается из четырех основных нитей. Это эпистемология Карла Поппера, раскрывающая путь научного знания; это квантовая механика; это основанная Тьюрингом теория вычислений, без которой не понять природу математических объектов; и, наконец, это универсальная теория эволюции, объясняющая развитие не только жизни, но и цивилизации.

Начало бесконечности

Объяснения, которые меняют мир

Дэвид Дойч, пер. с англ., 2014, 581 с.

Британский физик Дэвид Дойч — не только один из

основоположников теории квантовых вычислений, но и философ, стремящийся осмыслить «вечные вопросы» человечества в контексте, заданном развитием науки. Стержневой вопрос данной книги: есть ли предел для человеческого прогресса? Ответ выражен в заглавии: мы стоим у начала бесконечного пути, по которому поведет нас, выдвигая догадки и подвергая их критике, наш универсальный разум. Мы встали на этот путь в эпоху Просвещения, но с него легко сбиться под влиянием ошибочных философских идей, к которым автор причисляет многие течения мысли — от позитивизма до постмодернизма, не говоря уже о религии. Примером отступления от пути разума в науке предстает у него копенгагенская интерпретация квантовой механики. Разумную альтернативу ей Дойч видит в интерпретации Эверетта, из которой вытекает картина мира как мультивселенной. Но сфера интересов автора не ограничивается наукой.

Программируя Вселенную

Квантовый компьютер и будущее науки

Сет Ллойд, пер. с англ., 2-е изд., 2014, 256 с.

Каждый атом Вселенной, а не только различные макроскопические объекты, способен хранить информацию. Акты взаимодействия атомов можно описать как элементарные логические операции, в которых меняют свои значения квантовые биты — элементарные единицы квантовой информации. Парадоксальный, но многообещающий подход Сета Ллойда позволяет элегантно решить вопрос о постоянном усложнении Вселенной: ведь даже случайная и очень короткая программа в ходе своего исполнения на компьютере может дать крайне интересные результаты. Вселенная постоянно обрабатывает информацию — будучи квантовым компьютером огромного размера, она все время вычисляет собственное будущее.

---

notes

Примечания

1

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

2

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

3

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

4

Насколько мне известно, у названия Busy Child два источника. Первый — письмо английской принцессы Елизаветы беременной Катерине Парр, написанное в 1548 г. Елизавета выражала сочувствие Парр, которая очень плохо себя чувствовала из-за «беспокойного ребенка» и позже умерла при родах. Второй источник — неофициальный онлайн-фанфик на фильмы о Терминаторе. Busy Child в данном случае — ИИ, который вот-вот должен обрести сознание. —*Прим. авт.*

5

Английские рабочие, протестовавшие в начале XIX в. против появления машин на производстве. — *Прим. ред.*

6

В апреле 2013 г. студенты факультета Мичиганского университета поучаствовал в устроенном для них «зомби-апокалипсисе» под руководством преподавателя по эпидемиологии. — *Прим. ред.*

7

Начиная с 2013 г. Конференцию по сингулярности организует Университет сингулярности. — *Прим. авт.*

8

Рассел С., Норвиг П. Искусственный интеллект. Современный подход. — М.: Вильяме, 2007.

9

Google glasses — очки с дополненной реальностью, имеют полупрозрачный дисплей, камеру и компьютер. Перспективная разработка Google. — *Прим. ред.*

10

В декабре 2012 г. Рэй Курцвейл занял в Google пост технического директора, чтобы работать над проектами, связанными с машинным обучением и обработкой языка. Это веха на пути создания универсального ИИ, причем отрезвляющая. Цель Курцвейла — обратное проектирование работы мозга, он даже написал об этом книгу «Как создать разум: Тайна человеческой мысли раскрыта» (2012). Теперь в его распоряжении немалые ресурсы Google, которые он сможет использовать для реализации своей мечты. Пригласив уважаемого изобретателя, Google перестал скрывать свои планы по созданию УЧИ. — *Прим. авт.*

11

Крионика изучает хранение объектов при низких температурах, а криоконсервация — это сохранение тел умерших людей для излечения и оживления в будущем. — *Прим. авт.*

12

SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) — системы нейроморфной адаптивной гибкомасштабируемой электроники. — *Прим. ред.*

13

В январе 2012 г. Майкл Вассар покинул пост президента MIRI, чтобы участвовать в основании Meta Med — новой компании, предлагающей персонилизированную, основанную на научных данных диагностику и лечение. Его сменил Люк Мюлбхаузер. — *Прим. авт.*

14

Но подождите: разве это не тот самый антропоморфизм, в котором Юдковски обвинял меня? У нас, людей, базовые цели смещаются не только от поколения к поколению, но даже в течение жизни. Но как будет у машины? Мне кажется, Хьюз использует аналогию с человеком правильно, не антропоморфически. То есть мы, люди, представляем собой пример системы с глубоко укорененными функциями, такими как потребность в продолжении рода, но мы способны их преодолевать. Такая система напоминает аналогию с Ганди, которая также не антропоморфична. — *Прим. авт.*

15

Личное мнение автора. — *Прим. ред.*

16

Лат. «бог из машины». — *Прим. ред.*

17

Мог ли Гуд прочесть очерк Винджа, написанный под влиянием его собственного более раннего очерка, и изменить свое мнение? Мне кажется, это маловероятно. За свою жизнь Гуд опубликовал множество научных работ общим объемом порядка 3 млн слов и всегда самым тщательным образом ссылался на источники. И хотя во многих случаях он ссылается на собственные работы, я уверен, что он обязательно упомянул бы Винджа, если бы его очерк сыграл такую роль. Гуд любил подобные литературные рекурсии. — *Прим. авт.*

18

Буквальное значение английского слова singularity — уникальность. — *Прим. пер.*

19

Квиддич — волшебная спортивная игра из романов Джоан К. Роулинг о Гарри Поттере. — *Прим. ред.*

20

Милленаризм — убеждение в том, что с тысячелетними циклами истории человечества связаны значительные преобразования в обществе. Уходит корнями в новозаветные пророчества. — *Прим. ред.*

21

Рик Грейнджер, специалист по вычислительной нейробиологии из Дартмутского университета, утверждает, что каждый нейрон мозга соединен с многими десятками тысяч других нейронов. В этом случае мозг работает намного быстрее показателей, которые приводит Курцвейл в книгах «Эра духовных машин» и «Сингулярность рядом». Если мозг работает намного быстрее, то его компьютерный эквивалент по скорости не так близок к нам, как кажется. Но, учитывая Закон прогрессирующей отдачи, ненамного дальше. — *Прим. авт.*

22

Знаете, что еще удваивается примерно каждые два года? Интернет и все компоненты, которые делают его быстрее, связи в нем теснее, а способность впитывать информацию выше. В 2009 г., по оценке Google, Интернет содержал около 5 млн терабайт информации — в 250 000 раз больше, чем во всех книгах Библиотеки Конгресса США. К 2011 г. он должен был содержать примерно в 500 000 раз больше. Harris Interactive — интернет-компания маркетинговых исследований и изучения общественного мнения — объявила, что рост числа пользователей Интернета позволяет назвать его «самой быстрорастущей технологией в истории». В 2008 г. в мире было чуть меньше 1,2 млрд пользователей Интернета, в 2010 г. — более 2 млрд. — *Прим. авт.*

23

Карр Н. Пустышка. Что Интернет делает с нашими мозгами. — М.: BestBusinessBooks, 2012.

24

Ланир Д. Вы не гаджет. Манифест. — М.: Астрель, Corpus, 2011.

25

«Цветы для Элджернона» — научно-фантастический роман Дэниела Киза (существует экранизация) про умственно отсталого уборщика, который участвовал в эксперименте по улучшению интеллекта. Одно из последних изданий на русском: Киз Д. Цветы для Элджернона. — М.: Эксмо, 2014. — *Прим. ред.*

26

Я не уверен, что это верно в отношении AIXI Маркуса Хаттера, хотя специалисты утверждают, что да. Но поскольку AIXI невычислимый, он во всяком случае не может быть кандидатом на интеллектуальный взрыв. Другое дело — AlXItl, вычислимый аналог AIXI. Вероятно, это неверно также в отношении загрузки разума в компьютер, если она будет когда-нибудь реализована. — *Прим. авт.*

27

Некоторые сингуляритарии хотят получить УЧИ как можно скорее из-за его потенциально громадных возможностей в плане облегчения человеческих страданий. Это позиция Рэя Курцвейла. Другие считают, что появление универсального ИИ приблизит к личному бессмертию. Основатели MIRI, включая и Елиезера Юдковски, надеются, что работы над универсальным ИИ займут много времени, поскольку вероятность уничтожения человечества может снизиться со временем благодаря новым, более качественным исследованиям. — *Прим. авт.*

28

Существо из японской мифологии — кошка с раздвоенным хвостом. — *Прим. ред.*

29

LIDA (Learning Intelligent Distributed Agent) — обучение распределенных интеллектуальных агентов. — *Прим. ред.*

30

Радар О'Рейли — герой американского сериала M*A*S*H (в российском прокате — «Чертова служба в госпитале МЭШ»), идеальный секретарь, способный предвидеть события и многое делать за начальника. — *Прим пер.*

31

Карр Н. Великий переход: что готовит революция облачных технологий. — М.: Манн, Иванов и Фербер, 2013.

32

Хелен Адаме Келлер (1880–1968) — американская писательница, преподаватель, общественный деятель. В возрасте 19 месяцев перенесла болезнь, в результате которой полностью потеряла слух и зрение. С 7 лет Хелен занималась под руководством специалиста Энн Салливан, получила среднее образование и окончила колледж. Написала больше десяти книг, была известным филантропом и активистом. С 1980 г. в США отмечается День Хелен Келлер. — *Прим. ред.*

33

Справедливости ради заметим, что и Россия, и Китай постоянно подвергаются кибератакам спецслужб США. — *Прим. ред.*

34

Личное мнение автора, не находящее исторического подтверждения. — *Прим. ред.*

35

Субстандартный ипотечный кредит — кредит с особыми условиями, выдаваемый ненадежному заемщику в США. Кризис наступил, когда значительная часть заемщиков не смогла погасить кредит. — *Прим. ред.*