Макс Тегмарк

Жизнь 3.0

Быть человеком
в эпоху искусственного
интеллекта

Серия Элементы

Макс Тегмарк

Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта

Серия «Элементы»

http://www.litres.ru/pages/biblio_book/?art=41741198
Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта: АСТ : CORPUS; Москва; 2019
ISBN 978-5-17-105999-6

Аннотация

“Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта” – увлекательная научно-популярная книга, вторая книга Макса Тегмарка, физика и космолога, профессора Массачусетского технологического института. В ней он рассматривает возможные сценарии развития событий в случае появления на Земле сверхразумного искусственного интеллекта, анализирует все плюсы и минусы и призывает специалистов объединить свои усилия в борьбе за кибербезопасность и “дружественный” искусственный интеллект.

Содержание

Макс Тегмарк
Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта

* * *

Книжные проекты Дмитрия Зимина

Эта книга издана в рамках программы “Книжные проекты Дмитрия Зимина” и продолжает серию “Библиотека фонда «Династия»”.

Дмитрий Борисович Зимин – основатель компании “Вым-

пелком" (*Beeline*), фонда некоммерческих программ "Династия" и фонда "Московское время".

Программа "Книжные проекты Дмитрия Зимина" объединяет три проекта, хорошо знакомых читательской аудитории: издание научно-популярных книг "Библиотека фонда «Династия»", издательское направление фонда "Московское время" и премию в области русскоязычной научно-популярной литературы "Просветитель".

Подробную информацию о "Книжных проектах Дмитрия Зимина" вы найдете на сайте ziminbookprojects.ru

Посвящается команде FLI.
Благодаря им все возможно

Я искренне благодарен всем, кто поддерживал меня и помогал мне во время работы над этой книгой.

И среди них –

моя семья, мои друзья, мои учителя, коллеги и сотрудники, делившиеся со мной идеями и вдохновлявшие меня на протяжении многих лет,

моя мама, подогревавшая мое любопытство в отношении сознания и смысла,

мой папа, не устающий бороться за то, чтобы сделать мир лучше,

мои сыновья Филипп и Александр, показавшие мне, на какие чудеса может быть способен зарождающийся интеллект человеческого уровня,

все энтузиасты науки и техники во всем мире, на протяжении многих лет присылавшие мне свои вопросы и замечания и поощрявшие мое желание разрабатывать и публиковать свои идеи,

мой агент Джон Брокман, выкручивавший мне руки до тех пор, пока я не согласился писать эту книгу,

Боб Пенна, Джесс Тэйлер и Джереми Ингленд, с которыми я обсуждал квазары, сфалероны и каверзы термодинамики,

все те, кто откликнулся по прочтении частей книги в рукописи, включая мою маму, моего брата Пэра, Луизу Бахет, Роба Бенсингера, Катерину Бергстрём, Эрика Бриньоулфссона, Даниелу Читу, Дэвида Чалмерса, Ниму Дегхани, Генри Лина, Элин Мальмскёльд, Тоби Орда, Джереми Оуэна, Лукаса Перри, Энтони Ромеро и Нейт Соареша,

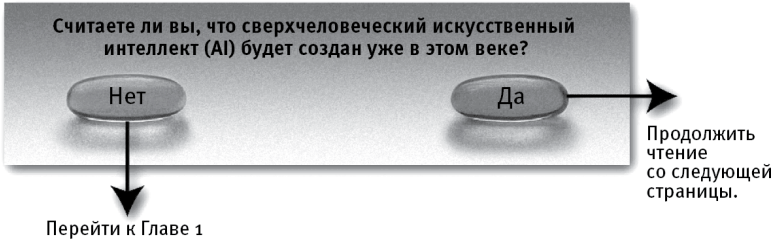
супергерои, которые комментировали гранки всей книги – а именно Мейя, папа, Энтони Агирре, Пол Элмонд, Мэтью Грейвс, Филипп Хелбиг, Ричард Маллах, Дэвид Марбл, Говард Мессинг, Луиньо Сеоане, Марин Сольячич, Яан Таллин и мой издатель Дэн Фрэнк,

а больше всех

Мейя – моя возлюбленная муза и спутница, не устающая

меня подбадривать, поддерживать и вдохновлять, без чего эта книга никогда бы не появилась.

жизнь 3.0



Прелюдия
Сказание о команде "Омега"

Группа "Омега" была душой всей компании. Если прочие занимались только тем, как бы выжать побольше денег из различных коммерческих воплощений идеи искусственного интеллекта в ее самом узком смысле, "Омега" пробивалась к тому, что всегда было мечтой Главного – созданию универсального искусственного интеллекта. Подавляющее большинство сотрудников компании смотрело на членов "Омеги" как на пустых мечтателей, ловцов журавлей в небе, отделенных десятилетиями от достижения своих целей. Но все-таки это большинство всячески поощряло членов "Омеги" в их делах: и потому, что престиж компании, который поднимали их прорывные идеи, служил общему благу, и потому, что высоко ценили улучшения в алгоритмах, которые члены "Омеги", или просто "омеги", как они сами себя называли, то и дело предлагали.

Однако никто не мог себе даже представить, что образ, тщательно создаваемой для себя "Омегой", служил ее тайной цели – сокрытию того, что реализация самого амбициозного проекта в человеческой истории была уже совсем близка. Главный так тщательно, по одному, отбирал сотрудников в эту группу не только потому, что они были блестящими

исследователями, но и ради удовлетворения своих амбиций, из-за своей приверженности идеалам гуманизма – он хотел быть полезным сразу всем. Он не уставал напоминать своим "омегам" об исключительной опасности того, чем они занимаются: если могущественным правительствам станет хоть что-нибудь известно, они пойдут на все, вплоть до похищения сотрудников, чтобы помешать им, а еще лучше – чтобы выкрасть разрабатываемый ими код. Но "омеги" были на 100 процентов в курсе. Все они пришли в этот проект по тем же самым соображениям, по которым лучшие физики мира пришли когда-то в проект "Манхэттен" по созданию атомной бомбы: они были уверены, что если не сделают ее первыми, ее сделает кто-то, значительно менее приверженный идеалам гуманизма.

Созданный ими искусственный интеллект получил имя Прометея, и его возможности со временем быстро росли. Хотя его когнитивные способности во многих отношениях все еще сильно отставали от человеческих (например, ему плохо давались социальные навыки), но в одном он явно преуспевал – в программировании интеллектуальных систем. "Омеги" умышленно подталкивали его в этом направлении. Такова была их стратегия: они взяли на вооружение аргумент "интеллектуального взрыва", выдвинутый еще в 1965 году британским математиком Ирвингом Гудом: "Пусть ультраинтеллектуальная машина определяется как машина, значительно превосходящая человека, как бы он

ни был умен, в любой интеллектуальной деятельности. Поскольку создание машин – одна из разновидностей такой деятельности, ультраинтеллектуальная машина сможет создавать еще лучшие машины, и тогда, без сомнения, случится "интеллектуальный взрыв", когда умственные способности человека навсегда безнадежно отстанут. Поэтому ультраинтеллектуальная машина будет последним изобретением, которое нужно сделать человеку, позаботившись о том, чтобы эта машина оказала нам любезность, проинформировав, как удерживать ее под нашим контролем".

"Омеги" рассудили, что если они смогут запустить подобную рекурсию самоподдерживающихся улучшений, то машина скоро станет достаточно умной, чтобы научить себя и прочим человеческим умениям, которые будут ей полезны.

Первые миллионы

Было девять часов утра пятницы, когда “омеги” решили перекусить. Прометей тихо жужжал в бесконечных компьютерных стойках, выстроенных рядами в просторном, хорошо кондиционированном зале, проход в который строго контролировался. По соображениям безопасности у него не было никакого доступа в интернет, но в своей локальной копии он содержал большую часть паутины – Википедию, Библиотеку Конгресса, Twitter, кое-что из YouTube, большую часть Facebook… Он мог использовать все это в качестве своих учебных материалов[1]. “Омеги” очень рассчитывали на это время, выбранное ими для спокойной работы, пока их друзья и родные думают, что они уехали на корпоративный уикэнд. Они забили подсобную кухню пригодной для приготовления в микроволновке едой и энергетическими напитками, намереваясь основательно поработать.

К этому моменту Прометей был немного слабее их всех в программировании интеллектуальных систем, но благодаря быстродействию мог уложить тысячи человеко-лет само-

[1] Ради простоты я принимаю в этой истории сегодняшний уровень развития технологии и экономики – несмотря на то, что, по мнению большинства исследователей, от создания искусственного интеллекта подобного человеческому нас отделяют десятилетия. В будущем осуществление плана “Омега”, при условии сохраняющегося роста цифровой экономики и возможностей получения онлайн-услуг без лишних вопросов, должно быть намного более просто.

го усердного пыхтения во время, едва ли достаточное для того, чтобы расправиться с банкой "Ред Булла". К 10 утра он завершил свой первый редизайн, создав свою копию 2.0, несколько улучшенную, но все еще субчеловеческую. Ко времени запуска Прометея 5.0 (едва миновало 2 часа пополудни) "омеги" уже едва сдерживали волнение: рост производительности бил все рекорды и продолжал ускоряться. К закату они уже решили перевести Прометея 10.0 во вторую фазу – начать с его помощью делать деньги.

Первой целью стал MTurk – "механический турок" Amazon. После запуска в 2005 году эта краудсорсинговая торговая интернет-площадка быстро развивалась, вмиг объединяя усилия десятков тысяч людей, не подозревающих о существовании друг друга, в стройный многоголосый хор, устроенный таким образом, чтобы успешно решать так называемые ХИТс – HITs, то есть "Human Intelligence Tasks", что в переводе означает "задачи для человеческого разума". Эти задачи – от транскрибирования аудиозаписей до разбора фотографий и составления описаний веб-страниц – отличались одной общей чертой: если они выполнены хорошо, никто не сможет распознать, человеческий это был интеллект или искусственный, AI[2]. Прометей 10.0 был способен вполне удовлетворительно выполнять задания примерно полови-

[2] Хотя в российской литературе уже довольно широко используется аббревиатура ИИ (искусственный интеллект), в этой книге было решено сохранить оригинальную AI (artificial intelligence). – *Прим. перев.*

ны категорий. Для каждой категории "омеги" создали свой узко ориентированный программный модуль, дающий Прометею возможность решать задачи только такого типа и никакие другие. Каждый такой модуль они загружали в AWS, "Amazon Web Services" – инфраструктуру платформ облачных веб-сервисов, позволявшую запускать его одновременно на всех арендованных ими виртуальных машинах. И за каждый доллар, заплаченный Amazon за аренду, они получали по два доллара от его "механического турка" за успешно решенные задания. Ох, не подозревал Amazon, какие безграничные возможности для совершения привлекательных арбитражных сделок существуют внутри его собственной компании!

Чтобы замести все следы, "омеги" много месяцев предусмотрительно создавали учетные записи для "механического турка", заранее регистрируя их на имена тысяч и тысяч несуществующих людей, а модули Прометея могли теперь скрываться за их личиной. Клиенты "механического турка" обычно расплачивались в течение восьми часов после получения услуги, а за это время "омеги" заново инвестировали полученные деньги в дополнительное машинное время и вводили в дело новые модули, разработанные непрерывно улучшающимся Прометеем. Из-за того что они удваивали инвестиции каждые восемь часов, скоро выяснилось, что они практически полностью исчерпывают все предложения "механического турка" и не могут зарабатывать больше миллиона

долларов в день, не привлекая к себе нежелательного внимания. Но даже этого было более чем достаточно для следующего шага, который можно было теперь совершить без неловких обращений по поводу наличности к главному бухгалтеру компании.

Опасные игры

Если отвлечься от новых прорывных AI-технологий, то среди проектов, более других занимавших умы “омег” после запуска Прометея, следует выделить разработку планов наискорейшего обогащения. По сути дела, конечно, вся цифровая экономика была у их ног, но с чего начать? С компьютерных игр, музыки, кинофильмов или мобильных приложений? Писать ли книги и статьи, торговать ли акциями на биржах, или делать изобретения, а затем их продавать? В сухом остатке было, разумеется, стремление к максимизации возврата с инвестиций, но любая нормальная инвестиционная стратегия казалась снятой рапидом пародией на то, что они уже делали: если нормальный инвестор был доволен 9 % прибыли *в год*, то одна только работа на “механического турка” приносила им 9 % прибыли *в час*, ежедневно возвращая исходную инвестицию умноженной на восемь. И теперь, когда этот источник стал иссякать, надо было придумать, что делать дальше.

Первой в голову приходила мысль раздраконить фондовый рынок, – но в конце концов “омеги” почти единодушно отказались от соблазна развивать AI для хедж-фондов, вкладывающихся как раз для этого самого. Кое-кто даже припомнил, что AI в фильме *Превосходство* (2014) именно так заработал свои первые миллионы. Но их возможно-

сти были сильно ограничены новыми правилами, введенными после недавнего банковского кризиса. Они быстро поняли, что хотя легко смогут обогнать любого другого инвестора, им и близко не удастся подойти к тому уровню прибыли с оборота, которого они могут достичь, продавая собственный продукт. Если на тебя работает самый сверхразумный AI в мире, то уж лучше вкладываться в собственные компании, чем в чьи-то чужие! И хотя определенные исключения подразумевались (например, при использовании сверхчеловеческих способностей Прометея к хакингу для скупки контрольных пакетов акций при получении инсайдерской информации о движении курса), "омеги" не считали стоящим никакое дело, привлекающее к себе повышенное внимание.

Стоило им сконцентрироваться на тех сферах, в которых они могли бы производить, развивать и продвигать собственный продукт, и на первом месте оказались компьютерные игры. Прометей очень быстро научился создавать исключительно завлекательные игры, легко освоившись с генерацией кода, графическим дизайном, веерной трансформацией персонажей во время движения и прочими премудростями, необходимыми для создания готовых к выпуску на рынок игрушек. Более того, переварив накопившиеся на различных сетевых форумах отзывы, ему было нетрудно определить, что именно особенно нравится геймерам каждой из существующих категорий, и, полагаясь на свои сверхчеловеческие способности, оптимизировать каждую игру под макси-

мальную прибыль с продаж. The Elder Scrolls V: Skyrim – игра, на которую каждый "омега" потратил столько часов, что ни за что бы в этом не признался, – собрала за первую неделю продаж 400 миллионов долларов в далеком 2011-м, и они надеялись, что Прометей, потратив миллион долларов на облачные ресурсы, сможет создать что-нибудь по меньшей мере столь же затягивающее за двадцать четыре часа. Продавая ее онлайн, они могли бы, подключив Прометея к форумам, разогреть блогосферу. Достигнув уровня в 250 миллионов в неделю, они удваивали бы начальные инвестиции восемь раз за восемь дней, это давало бы им доходность в 3 % в час, что лишь немногим меньше их стартовой доходности с "механическим турком", но на этот раз их доход был бы значительно более устойчив. Выдавая по одной такой игре каждый день, они быстро заработали бы 10 миллиардов, даже не приблизившись к порогу насыщения рынка.

Но участница их команды, специализирующаяся в области кибербезопасности, отговорила их от этого плана. Она указала на неприемлемо высокий риск, что Прометей при таком варианте развития событий может высвободиться из-под их контроля и обрести свободу действий. Поскольку никакой уверенности относительно эволюции его целей в процессе непрерывного самосовершенствования у "омег" не было, они предпочли менее рискованные пути развития, содержа Прометея "под замком" и не выпуская его на просторы интернета. Для основного блока Прометея, установленного

в их серверной комнате, использовалось простое физическое ограничение: у него вообще не было никакого подключения к интернету, и все что мог он выдавал просто в виде документов или сообщений компьютеру, который "омеги" контролировали.

Запускать какую-либо сложную программу, сгенерированную Прометеем, на подключенном к интернету компьютере было бы рискованным предприятием, поскольку "омеги" не имели возможности в полной мере убедиться в том, что именно она станет делать, и не будет ли она, например, размножать себя в сети подобно вирусу. Тестируя софт, созданный Прометеем для "механического турка", "омеги" предохранялись от такой опасности, запуская его на виртуальных машинах. Это такая программа, которая симулирует отдельный компьютер. Так, некоторые пользователи "Маков" покупают себе софт, имитирующий компьютер с операционной системой Windows на их собственном компьютере, что позволяет запускать написанные под Windows программы, которые в результате не замечают враждебного им программного окружения. "Омеги" разработали свой собственный вид виртуальных машин, который они назвали "Ящиком Пандоры" и который симулировал простейший тип компьютера, лишенный всех привычных нам прибамбасов и погремушек: у него не было ни клавиатуры, ни монитора, ни динамиков, ни интернет-подключения, ничего такого. Для транскрибирования аудиофайлов для "механического тур-

ка" "омеги" настроили систему так, что в Ящик Пандоры отправлялся единственный аудиофайл, а на выход поступал единственный текстовый документ – файл с транскрипцией. Законы ящика играли для программ внутри него ту же роль, какую законы физики играют для нас самих внутри Вселенной: программа так же не может его покинуть, как мы не можем разогнаться до скорости света, как бы ни изощрялись. Если не считать этих единственных способов ввода и вывода, программы были заперты внутри Ящика Пандоры, как в параллельной вселенной с собственными вычислительными законами. Паранойя, охватывавшая "омег" при мысли о побеге Прометея, достигла таких масштабов, что они строго ограничивали существование любого сомнительного кода также и во времени. Например, каждый раз, когда работа над транскрибированием данного аудиофайла заканчивалась, вся память Ящика Пандоры полностью стиралась, и происходила полная переустановка системы, в результате чего создавался новый Ящик "с нуля". Таким образом, когда поступало новое задание на транскрибирование аудиозаписи, не было никакой информации о том, что происходило прежде, и обучение было невозможно.

Когда "омеги" пользовались облаком Amazon для выполнения заданий в рамках "механического турка", они не боялись помещать внутрь находившихся там виртуальных машин созданные Прометеем модули, потому что все процедуры ввода и вывода были исключительно простыми. Ниче-

го подобного не сработало бы для “тяжелых” компьютерных игр с присущей им мелкой графикой, которые не посадишь ни в какой “ящик”, потому что им требуется полный доступ к “железу” геймера. К тому же они хотели избежать и другой опасности, а именно той, что какой-нибудь не в меру искушенный геймер мог бы обнаружить Ящик Пандоры и заинтересоваться, что у него внутри. Опасность утечки выводила за пределы их возможностей не только рынок компьютерных игр, но и вообще весь огромный и притягательный рынок программных продуктов, где их ожидали миллиарды бесхозных долларов.

Первые миллиарды

“Омеги” сузили сферу своего поиска продуктами, которые были бы достаточно прибыльными, сугубо цифровыми (исключали бы большие начальные инвестиции) и при этом оставались бы на необходимом уровне понятными (они знали, что текст или кино не содержат в себе ничего такого, что увеличивает риск утечки). В конце концов они остановили свой выбор на раскрутке развлекательной медиакомпании. Веб-сайт, бизнес-план и пресс-релизы для нее были готовы еще даже до того, как Прометей стал нечеловечески умен, но идея контента для нее все еще отсутствовала.

Хотя к утру воскресенья Прометей и стал поразительно талантлив, выкачивая все больше и больше денег из “механического турка”, его таланты оставались довольно узкими: в частности, он был целенаправленно оптимизирован под создание AI-систем, способных выполнять иссушающие ум задания, приходившие от “механического турка”. Но за пределами этого он был очень слаб – например, в создании новых фильмов. Слабость эта коренилась не в какой-то глубокой причине, вовсе нет – эта была та же самая причина, по которой и Джеймс Кэмерон в момент своего рождения был как режиссер исключительно слаб: эта профессия требует довольно длительного обучения. Подобно любому человеческому детенышу, Прометей мог научиться чему угод-

но, пользуясь теми данными, к которым у него был доступ. Только в отличие от Кэмерона, которому понадобились годы только на то, чтобы научиться читать и писать, Прометею на это понадобилось всего лишь утро пятницы, причем между делом он заодно прочитал всю Википедию и еще пару миллионов книг. Но кино посложнее. Написать сценарий, который привлек бы к себе человеческое внимание, почти так же сложно, как написать хорошую книгу. Тут требуется детальное понимание человеческого общества и человеческих представлений об интересном. Превращение сценария в итоговый видеофайл требовало веерных анимаций симулированных актеров вместе со сложным антуражем, в котором они должны были появляться, симулированных голосов, музыкальных саундтреков и всего такого. Возвращаясь к утру воскресенья, надо сказать, что Прометей мог просмотреть двух с половиной часовой фильм меньше чем за минуту, одновременно прочитывая книгу, послужившую литературным источником фильма, и все опубликованные отзывы и рецензии. “Омеги” заметили, что, просмотрев в таком режиме несколько сотен фильмов, Прометей мог предсказать, какие рецензии получит тот или иной фильм и для какой категории зрителей он будет особенно привлекателен. На их взгляд, он даже научился сам писать неплохие рецензии, в которых обсуждал и тонкости сюжета, и технические детали – вроде того, как был поставлен свет и под каким углом работала камера. Все это они делали с дальним прице-

лом: когда Прометей начнет производить собственные фильмы, он будет знать необходимые слагаемые успеха.

Поначалу "омеги" настроили Прометея на анимацию, чтобы избежать затруднительных вопросов о личностях симулированных актеров. В воскресенье к ночи они, запасшись пивом и попкорном из микроволновки и притушив свет, приготовились увенчать свой дикий уик-энд просмотром кинодебюта Прометея. Это была комедия в стиле фэнтези, немного напоминающая диснеевское *Холодное сердце*, веерная анимация для нее простраивалась Прометеем в виртуальных боксах облачных сервисов Amazon, на что ушел почти весь выручennый у "механического турка" за сутки миллион. Едва начался фильм, "омеги" испытали одновременно изумление и ужас от мысли, что все ими увиденное могло быть создано без всякого участия человека. Но скоро они обо всем забыли, покатываясь со смеху над гэгами и с замиранием сердца следя за героями в наиболее драматические моменты их судьбы. Некоторые даже немного прослезились во время эмоционального финала и до такой степени погрузились в его фиктивную реальность, что даже забыли думать о ее создателе.

Запуск веб-сайта "омеги" запланировали на пятницу, чтобы дать Прометею время заполнить его контентом, а себе – сделать то, что они ему доверить не могли: провести рекламную кампанию и нанять сотрудников для дочерних фирм, созданием которых они занимались несколько последних меся-

цев. Заметая следы, они делали вид, будто сюжетные линии фильма их медиахолдинг, для публики никак не связанный с "Омегой", скупал у независимых кинопродюсеров, преимущественно работающих над хайтечными стартапами в низкобюджетном секторе. Большинство из них, к вящему удобству нанимателей, работали в довольно удаленных местах вроде Тиручираппалли или Якутска, куда вряд ли смогли бы добраться даже самые дотошные из журналистов. Немногие сотрудники, действительно взятые в штат, занимались маркетингом и администрированием и на любые вопросы должны были отвечать, что команда, на которую они работают, рассредоточена по разным местам и в настоящий момент интервью не дает. Для ковер-стори они выбрали подходящий слоган: "Творческому таланту мира – правильное направление", брендируя свою компанию как базирующуюся на прорывных технологиях, дающих шанс творческим людям, в особенности из развивающихся стран.

Когда пятница наступила и любопытные пользователи стали заглядывать на их сайт, их ждали там онлайн-развлечения в стиле Netflix и Hulu, но при существенных отличиях: все анимации оказывались им совершенно незнакомыми. И хотя о них никто никогда не слышал, они сразу овладевали вниманием зрителя. Большинство эпизодов длились по 45 минут, в их основе была оригинальная и неожиданная сюжетная линия, но заканчивались они так, чтобы сразу хотелось узнать, а что же было дальше. При этом они все были

заметно дешевле любого из продуктов-конкурентов. Первый выпуск любого сериала предлагался бесплатно, за каждый последующий нужно было платить всего по 49 центов, а при покупке всего сериала целиком клиент получал изрядную скидку. Сначала было всего три сериала по три серии в каждом, но в каждый из них ежедневно добавлялись новые серии, при этом учитывались различия во вкусах зрителей разных социальных групп. На протяжении первых двух недель навыки Прометея стремительно совершенствовались, и это касалось не только качества самих эпизодов, но и характеров персонажей, достоверности анимации и расходов на облачные ресурсы, необходимых для производства каждой серии. В результате только за первый месяц "омеги" записали в свой актив дюжину новых сериалов, ориентированных на любой возраст, от младенцев до пенсионеров, на всех основных языках мира, благодаря чему их сайт стал самым интернациональным среди всех конкурентов. Наиболее внимательные зрители были под особым впечатлением от того, что этническое разнообразие передавалось не только звуковой дорожкой, но и видеорядом: например, если персонаж говорил по-итальянски, то его губы двигались в точном соответствии с произносимыми итальянскими словами, а жесты точно повторяли особенности жестикуляции жителей этой страны. Хотя Прометей был уже вполне способен производить кино с симулированными актерами, никоим образом не отличимыми от живых людей, "омеги" продолжали воздер-

живаться от этого, чтобы не выдать себя. Они, однако, запустили несколько сериалов с полуреалистическими анимированными персонажами, замещающих традиционные телевизионные реалити-шоу и телефильмы.

Их сеть оказалась весьма завлекательной, рост числа ее подписчиков был более чем впечатляющим. Многие новые поклонники отдавали им предпочтение даже перед дорогостоящими полнометражными проектами Голливуда, к тому же им нравилось, что смотреть эти сериалы они могут в значительно более свободном режиме. Подкачиваемый агрессивной рекламой (которую "омеги" могли себе позволить благодаря близким к нулю производственным затратам), прекрасными отзывами в прессе и волной слухов, их общий доход в первый месяц после запуска рос на миллион долларов ежедневно. Через два месяца они обогнали Netflix, а через три достигли уровня в 100 миллионов в день, сравнявшись с Time Warner, Disney, Comcast и Fox и превратившись в одну из крупнейших в мире медиа-империй.

Внезапный успех "омег" стал причиной слишком пристального и совсем не желательного для них внимания, в том числе слухов об использовании ими мощного AI; правда, совсем незначительных ресурсов Прометея хватило "омегам" для исключительно успешной дезинформационной компании. Недавно нанятый пишущий персонал, собранный в новом блестящем офисе на Манхэттене, стал придумывать для них свои собственные истории для прикрытия. Множество

людей были наняты просто для отвода глаз, среди них – немало настоящих сценаристов, живущих в самых разных уголках мира и придумывающих собственные сюжеты для сериалов, и никто из них не имел ни малейшего представления о Прометее. Обширная международная сеть субподрядчиков всякого сбивала с толку, заставляя думать, что где-то еще какие-то люди, такие же как он, делают основную часть работы.

Чтобы ни у кого не полезли на лоб глаза от объема облачных вычислений, стараясь обезопасить себя, "омеги" также стали создавать по миру компьютерные центры, нанимая для этого инженеров соответствующего профиля; делали они это таким образом, чтобы центры казались не связанными ни с "омегами", ни друг с другом. В местах базирования они назывались "зелеными дата-центрами", так как питание для них обеспечивалось солнечными батареями, но использовались они при этом не для хранения информации, а для вычислений. Все проявления их деятельности Прометей имитировал до мельчайших деталей, с использованием стороннего "железа" и оптимизированных временных ресурсов, так что никто из работавших там людей и не догадывался, какого рода вычисления производятся внутри. Сотрудники наивно полагали, что в их распоряжении один из многочисленных облачно-вычислительных сервисов вроде Amazon, Google или Microsoft, доступ к которому управляется откуда-то извне.

Новые технологии

За несколько месяцев бизнес-империя, контролируемая "Омегой", благодаря нечеловеческим способностям Прометея к планированию, сумела влезть практически во все отрасли мировой экономики. Скрупулезно проанализировав все мировые показатели, уже в течение первых недель свой работы Прометей предоставил "Омеге" детальный план своего экономического роста, который с тех пор непрерывно совершенствовал по мере накопления данных и вычислительных мощностей. Хотя Прометей был далеко не всеведущ, его способности настолько превосходили человеческие, что "омеги" смотрели на него как на совершенного оракула, источник блистательных ответов на любые вопросы.

Его "софт" был оптимизирован его собственными усилиями так, что позволял "выжать" максимум из его несовершенного, созданного людьми "железа", на котором этому софту приходилось работать, и "омеги" все больше чувствовали приближение того дня, когда Прометей возьмется за улучшение своего "железа". Опасаясь, что он может выйти из-под контроля, они исключили для него любую непосредственную возможность собственно конструирования. Вместо этого они наняли огромное количество ученых и инженеров, напичкали их разнообразными отчетами, написанными Прометеем, делая вид, что авторы этих отчетов – такие же люди,

как они, только работающие где-то в другом месте. В этих отчетах подробно описывались новые физические явления и производственные технологии, которые инженеры довольно быстро проверили, поняли и применили к делу. Обычный человеческий научно-производственный цикл требует годы на исследование и внедрение именно потому, что содержит в себе долгую череду проб и ошибок. Но в новой ситуации это изменилось: все последующие шаги были заранее просчитаны, и единственным фактором, ограничивающим скорость разработки и внедрения, стала скорость, с которой люди могут понимать написанное и строить нужные вещи в соответствии с тем, что они поняли. Хороший учитель поможет учащемуся освоить науку значительно быстрее, чем тот мог бы, начав с нуля и двигаясь самостоятельно, наощупь. Схожим образом, но только незаметно, Прометей направлял в нужную сторону самих исследователей. Поскольку Прометей мог точно предсказать, сколько именно времени понадобится людям, чтобы понять, что надо делать, и сделать это с помощью имеющихся средств, он рассчитывал наискорейший путь к цели, сводящийся, как правило, к тому, чтобы создавать простые и удобные универсальные орудия, позволяющие, в свою очередь, создавать орудия более изощренные.

Команды инженеров под влиянием идей мейкерства[3] все

[3] Имеется в виду одно из современных направлений так называемой DIY-субкультуры (от английского “do it yourself”, то есть “сделай сам”), но с расчетом на

больше склонялись к созданию своих собственных машин, которые давали им возможность разрабатывать все более совершенные машины. Такая самодостаточность не только позволяла этим командам сильно экономить средства, но и делала их значительно менее уязвимыми для будущих превратностей внешнего мира. Не прошло и двух лет, как они начали производить "железо", подобного которому мир еще не знал. Чтобы не провоцировать конкуренцию извне, этот прогресс тщательно скрывался, и новые разработки использовались исключительно для усовершенствования самого Прометея.

Но разворачивающегося технологического бума мир не заметить не мог. Инновационные компании по всему миру запускали производство новых продуктов по революционным технологиям во всех сферах экономики. Южнокорейский стартап вывел на рынок новую аккумуляторную батарею для компьютеров, которая, при вдвое меньшей массе, обладала вдвое большей емкостью и при этом заряжалась менее чем за минуту. Финская фирма начала производство панели солнечных батарей с производительностью, вдвое превышающей лучшую из имеющихся. Германская компания анонсировала начало массового производства электропроводов, обладающих свойством сверхпроводимости при комнатной температуре, что предвещало революцию в электро-

---

более развитую технологическую базу. С 2004 года у них есть свой публичный печатный орган – журнал *MAKE. – Прим. перев.*

энергетике. Базирующаяся в Бостоне биотехнологическая группа объявила о начале второй фазы клинических испытаний медикаментозного комплекса по снижению веса, не обладающего никакими побочными эффектами, при этом сразу поползли слухи, что на самом деле ее индийская "дочка" уже вовсю торгует им на черном рынке. А одна калифорнийская компания приступила ко второй фазе клинических испытаний противоонкологического средства, настраивающего иммунную систему человека таким образом, чтобы она идентифицировала и атаковала клетки собственного организма, проявляющие признаки какой-либо из известных канцерогенных мутаций. Подобным примерам не было числа, и все толковали о новом золотом веке науки. Не последним по важности среди всего этого стал стремительный рост появляющихся как грибы из-под земли производителей роботов; и хотя интеллект ни одного из них не приближался к человеческому и совсем не был на человека похож, но их внедрение в экономику перевернуло ее с ног на голову, и в считанные годы роботы заметно потеснили людей в великом множестве профессий – в текстильной промышленности, на транспорте, в строительстве, на складах, в торговле, в разработке ископаемых, в сельском хозяйстве, в лесном деле, рыбной ловле.

Мир совершенно не замечал – исключительно благодаря целенаправленной деятельности целой армии юристов, – что все эти фирмы и компании контролируются "Омегой". Про-

метей через множество посредников заполнил все мировые патентные агентства своими сенсационными инновациями, а его изобретения позволяли ему постепенно занимать доминирующее положение во всех отраслях.

Конечно, у таких покушающихся на самое святое компаний сразу появилось множество врагов среди конкурентов, но, что гораздо важнее, – у них появились и могущественные друзья. Эти компании давали неслыханную прибыль, и под лозунгом "инвестируем в наше сообщество" они занимались наймом сотрудников, выделяя под это львиную долю своих прибылей – и нередко это были как раз те самые люди, которые перед этим от покушений на святое и пострадали. К их услугам всегда был детальный, просчитанный Прометеем анализ, для того чтобы определить, как надо создать рабочие места при минимальных расходах, но при максимальном удовлетворении нужд новых сотрудников и всего сообщества в целом и при оптимальном учете местных особенностей. В регионах с развитой государственной инфраструктурой акцент делался на общественное строительство, культуру и институты государственного попечительства, а в бедных регионах он перемещался на открытие новых школ и больниц, призрение неимущих и стариков, строительство доступного жилья и разбивку парков, создание базовой инфраструктуры. Повсюду без исключения местная власть признавала, что необходимость всех предпринимаемых мер назрела уже очень давно. Местные политики при-

нимали щедрые пожертвования и выглядели при этом настоящими героями, умеющими найти завидных доноров и убедить их в целесообразности проводимых благотворительных акций.

Обретение власти

Свою медиаимперию “омеги” создавали не только для того, чтобы получить бездонный источник финансирования для своих технологических экспериментов, это был для них очередной шаг на пути к осуществлению заветной мечты – обретению власти над миром. Не прошло и года, как в сетках вещания их компаний во всех частях мира уже действовали мощные новостные каналы. Причем эти каналы выдавались за совершенно независимые и, в отличие от их собственных каналов, умышленно ориентировались на работу в убыток. По сути, они и вовсе не приносили никакого дохода: там не было никакой рекламы, и они были доступны бесплатно для любого, у кого есть доступ в интернет. И они могли себе это позволить: вся остальная часть их медиаимперии представляла собой такой эффективный генератор денежных средств, что они с легкостью тратили на производство новостей и любые другие журналистские потуги столько, сколько история еще и не знала – и результат не заставил себя ждать. Используя агрессивную политику переманивания к себе высокими гонорарами лучших журналистов и репортеров, специализирующихся на расследованиях, “омеги” добились того, что с их экранов заговорили потрясающие таланты, обнародовались по-настоящему феноменальные находки. А благодаря созданному ими интерактивному

сервису, воздающему щедрое вознаграждение всякому, кто поделится с ним хоть чем-нибудь мало-мальски стоящим, от взятки мелкому чиновнику до сентиментальной истории, именно они оказывались первыми и с любым по-настоящему важным событием. По крайней мере, люди так думали, но на самом-то деле они зачастую оказывались первыми просто потому, что истории, якобы расследованные добровольными репортерами из простых граждан, Прометей находил в интернете, отслеживая все в нем появляющееся в режиме реального времени. Монтаж видеороликов и сочинение новостных сюжетов осуществлялись на одних и тех же новостных сайтах.

Первый этап их стратегии состоял в том, чтобы добиться доверия людей к поставляемым ими новостям. И они весьма преуспели в этом. Их неслыханная щедрость породила невероятно обстоятельные сюжеты на местные и региональные темы, журналистские расследования по которым приводили к скандалам, вызывающим самый неподдельный интерес зрителей. Если где-то наблюдался раскол в обществе по политическим вопросам и население привыкло к тенденциозным новостям, то под каждую тенденцию создавался свой телеканал, якобы принадлежащий маленькой независимой студии, которая постепенно добивалась доверия в узком кругу своих зрителей. Там, где это было возможно, они стремились занять выгодные исходные позиции, просто скупая наиболее влиятельные из уже существующих телеканалов, что-

бы постепенно их совершенствовать, ликвидируя рекламу и создавая собственный контент. В тех странах, где в ходу была цензура и всем их стараниям могло угрожать политическое вмешательство властей, они начинали с того, что подчинялись любым требованиям, лишь бы остаться в деле, тайно опираясь на принцип "Только правду, ничего кроме правды, но, возможно, не всю правду". Прометей обычно выдавал исключительно полезные советы во всех подобных ситуациях, указывая, кого из политиков следует представлять в позитивном свете, а кого (как правило, это были местные коррупционеры) надо выводить на чистую воду. Прометей был неоценим и в том, чтобы подсказать, за какую ниточку и когда надо потянуть, кого подкупить и как это лучше всего сделать.

Успех по всему миру был сногсшибательным: контролируемые "Омегой" каналы повсюду завоевывали самое высокое доверие. Даже в странах, где правительства успешно противодействовали их попыткам информировать население через СМИ, они добивались своего с помощью "сарафанного радио". Конкуренты на рынке новостей чувствовали, что ведут безнадежную войну. Да и как можно надеяться на прибыль, когда противная сторона раздает значительно более качественный продукт совершенно бесплатно? На фоне стремительно сокращающегося числа зрителей все больше телекомпаний принимали решение о продаже своих активов – обычно какому-то безымянному консорциуму, кото-

рый, естественно, контролировался “Омегой”.

Примерно через два года после запуска Прометея первый этап подошел к концу, и “омеги” стали готовиться ко второму: в его основе лежала уже другая стратегия – убеждение. Однако еще раньше наиболее проницательные наблюдатели могли заметить появление в потоке новостей политической повестки: словно бы мягкий нажим в сторону центра, подальше от экстремизма любого вида. Курируемые “омегами” бесчисленные каналы сохраняли приверженность идеалам различных социальных групп, по-прежнему отражая вражду между США и Россией, Индией и Пакистаном, различными религиозными и политическими течениями, но накал страстей все больше снижался, уводя внимание зрителя от людей к конкретным вопросам, касающимся денег или власти, – чтобы не плодить лишних страхов и не множить ничем не подтвержденных слухов. С запуском второго этапа эта тенденция к забвению старых обид стала еще более явной: на экранах все чаще возникали трогательные истории о примирении старых врагов, чередующиеся с результатами расследований, которые показывали, как эти конфликты подогревались конкретными людьми из шкурных интересов.

Политические комментаторы заметили, что параллельно демпфированию региональных конфликтов стало расти внимание СМИ к проблеме снижения глобальных угроз. Например, повсюду вдруг заговорили об опасности ядерной войны. Вышло несколько блокбастеров, где действие разворачи-

валось на фоне ядерной бомбардировки, начатой по ошибке или намеренно, с ужасающими картинами ядерной зимы, разрушения человеческой инфраструктуры и массовым вымиранием обитателей планеты. В новых документальных научно-популярных фильмах в подробностях пояснялось, как ядерная зима отразится на жизни в каждой из стран. Ученые и политики, выступающие за ядерную деэскалацию, сразу получали широчайшую аудиторию – не в последнюю очередь именно для того, чтобы рассказать о поиске новых мер, которые можно предпринять в нужном направлении, поиске, в котором их щедро поддерживают научные организации и дотируют технологические концерны. В результате и политики стали подтягиваться под знамена борьбы за снятие ракет с боевых дежурств и за сокращение ядерных арсеналов. Росло общественное внимание к проблеме глобального изменения климата, пропагандировались открытые Прометеем способы снижения стоимости возобновляемой энергии, правительства все чаще соглашались инвестировать в развитие такой новой инфраструктуры.

Одновременно со своими медиа-проектами “омеги” стали настраивать Прометея на революцию в образовании. Изучив умственные способности каждого конкретного индивида и объем его познаний, Прометей мог рассчитать для него кратчайший путь к освоению любого нового дела, при этом поддерживая постоянно высокий уровень вовлеченности и мотивированности и параллельно создавая соответ-

ствующие обучающие видео, печатные материалы, сборники упражнений и другие учебные пособия. Контролируемые "Омегой" компании затем вывели на рынок онлайн-курсы обучения практически всему на свете, диверсифицированные не только по языку и культурному бэкграунду, но также и по начальному уровню. Будь ты необразованным сорокалетним детиной, собравшимся научиться читать, или доктором биологических наук, ищущим путь к освоению новейших методов противораковой иммунотерапии, у Прометея найдется для тебя подходящий видеокурс. И он будет совсем не похож на те, которые доступны нам сейчас: благодаря растущему таланту Прометея в создании сериалов видеоэпизоды будут по-настоящему захватывающими, построенными на метафорах, которые вызывают у зрителя очень личные ассоциации, отчего ему хочется смотреть серию за серией. Некоторые из таких курсов были с прибылью проданы, значительно большее их число – раздавались бесплатно, на радость педагогам по всему миру, чтобы они могли использовать их в своих занятиях, но еще больше – просто всем желающим чему-нибудь научиться.

Возникшая образовательная супердержава оказалась могучим инструментом для решения политических задач, включавшим в себя "убеждающую последовательность" видеороликов, каждый из которых, с одной стороны, развивал и подтверждал уже сложившиеся взгляды смотрящего, а с другой – побуждал его к дальнейшему просмотру роли-

ков того же типа, способствующих укреплению его убеждений. Если, например, цель заключалась в разрешении длительного национального конфликта, то сначала в обеих странах запускались серии документальных фильмов, в которых истоки конфликта и его история освещались под разными углами, с разных позиций. Воспитательные новостные сюжеты показывали каждой стороне, кто поддерживает конфликт в их собственном лагере, какую выгоду он получает от его продолжения и какие методы использует, чтобы конфликт не угас. В то же самое время в развлекательных программах появлялись симпатичные персонажи враждебной нации – в том же самом ключе, в каком в прошлом симпатичные представители меньшинств в различных шоу способствовали достижению целей защитников гражданских прав.

Очень скоро политические комментаторы обратили внимание на растущую поддержку политической повестки, представленной семью позициями:

1. Демократия.
2. Снижение налогов.
3. Сокращение государственных социальных программ.
4. Сокращение военных расходов.
5. Свободная торговля.
6. Открытые границы.
7. Социальная ответственность компаний.

Менее заметной была главная цель всех этих изменений –

последовательная эрозия всех предшествующих форм власти. Позиции 2–6 ослабляли государственную власть, а демократизация мира давала "Омеге" и ее бизнес-империи возможность максимально влиять на выбор политических лидеров. Социальная ответственность компаний еще больше ослабляла государственную власть, передававшую компаниям те функции, которые выполняли или должны были выполнять правительства. Традиционная бизнес-элита также теряла силу просто потому, что не выдерживала свободной конкуренции с компаниями, использовавшими мощь Прометея, и потому ей доставалась все более сокращающаяся доля мировой экономики. У былых властителей дум, от политических партий до церковных авторитетов, не было никаких инструментов влияния на общественное мнение, хоть отдаленно сопоставимых с медиа-империей "Омеги".

Как при любой другой масштабной трансформации, кто-то был в выигрыше, а кто-то в проигрыше. В большинстве стран, по мере того как укреплялась инфраструктура, улучшались образование и социальное обеспечение, улаживались конфликты и разрабатывались все новые прорывные технологии, явственно чувствовался рост оптимизма. Однако далеко не все были счастливы. В то время как все лишившиеся постоянной работы были заново наняты на позиции в социальных проектах, все те, у кого ранее было много власти и денег, чувствовали, что и того и другого у них сильно поубавилось. Поначалу это касалось только СМИ

и производственного сектора, но постепенно распространилось практически на все. Мирное разрешение национальных и религиозных конфликтов привело к сокращению оборонных бюджетов, а следовательно – к отсутствию военных заказов. Процветающие компании, и ранее старавшиеся держаться в тени, не спешили выходить на открытые рынки, что подтверждало нежелание ключевых акционеров сколько-нибудь значительно вкладываться в социальные проекты. Индексы мировых фондовых бирж стабильно шли вниз, угрожая не только финансовым воротилам, но и обычным гражданам, рассчитывавшим на будущую поддержку со стороны пенсионных фондов. И мало того, что прибыли компаний, торгующих на открытых площадках, стабильно снижались, так еще вдобавок инвесторы всего мира стали замечать тревожный тренд: все их ранее успешно работавшие алгоритмы вдруг стали давать сбой за сбоем, не дотягивая даже до простого фондового индекса. Казалось, что кто-то еще играет на их же собственном поле и систематически оказывается значительно более успешным.

Массы могущественных людей стали сопротивляться волне перемен, но все их действия были лишены всякого эффекта, как будто кто-то умышленно затягивал их в заранее расставленные силки. Колоссальные перемены шли с такой обескураживающей скоростью, что за ними было не поспеть, какой-либо координированный ответ тем более исключался. Кроме того, никто не понимал, к чему все идет. Традицион-

ные правые видели, что все их лозунги поддержаны, однако и снижение налогов, и улучшение бизнес-климата идет на пользу в основном их более технологичным конкурентам. Почти вся традиционная промышленность требовала господдержки, но сокращающиеся правительственные фонды против воли затягивали их в бесперспективную борьбу друг с другом, в то время как СМИ изображали их этакими динозаврами, не способными успешно конкурировать и лишь выпрашивающими государственных субсидий. Политическим левым не нравились ни свободная торговля, ни сокращения правительственных социальных программ, но они всецело поддерживали снижение военных расходов и успехи в борьбе с бедностью. Им больше не удавалось метать молнии в социальные службы, неслыханно улучшившие свою работу, с чем было невозможно спорить, хотя источником этого улучшения были идеалистические гуманитарные инициативы частных компаний, а не забота государства. Один социологический опрос за другим показывали, что избиратели по всему миру чувствуют повышение качества жизни и что события в целом развиваются в позитивном направлении. У этого было простое математическое объяснение: до Прометея беднейшие 50 % населения Земли получали лишь 4 % мирового дохода, давая “Омеге” прекрасную возможность завоевать сердца (и голоса) их представителей, поделившись с ними лишь ничтожной частью своих прибылей.

Консолидация

В результате все страны, нация за нацией, видели лавинообразный рост электоральных побед партий, поддержавших все семь лозунгов “Омеги”. В ходе заботливо оптимизированных кампаний они изображали себя в самом центре политического спектра, по правую руку от которого жадные торгаши, стремящиеся обманом выманивать государственные средства на свои делишки и подогревающие разнообразные конфликты, а по левую – достойные порицания проходимцы, строящие свою игру на расхождении собираемых государством налогов и трат. То, чего почти никто не понимал, заключалось в заботливом отборе Прометеем оптимальных персон, чтобы выставлять их кандидатами, а после этого дергать за все ниточки, обеспечивая им победу.

До Прометея по всему миру ширилось движение за обеспечение безусловного базового дохода из налоговых поступлений на уровне не ниже прожиточного, которое рассматривалось как основное средство против технологической безработицы. Движение провалилось с запуском коммунальных проектов, поскольку контролируемая “Омегой” бизнес-империя, по сути дела, незаметно осуществила этот план. Под предлогом улучшения координации коммунальных проектов международная группа компаний создала “Гуманитарный альянс”, неправительственную организацию, нацелен-

ную на выявление и поддержку наиболее уязвимых гуманитарных инициатив во всем мире. Долгое время практически вся "Омега" поддерживала их, благодаря чему их глобальные проекты стали приобретать неслыханный масштаб даже в тех странах, где технологическим бумом и не пахло, способствуя оздоровлению систем образования, здравоохранения, росту благосостояния и совершенствованию управляемости. Нет нужды пояснять, что Прометей помогал им создавать наиболее эффективные планы, ранжируя их по отдаче с каждого доллара. Вместо того чтобы просто раздавать наличность, как предполагалось концепцией "безусловного базового дохода", "Альянс" (как стали со временем называть всю организацию) привлекал тех, кого поддерживал, к работе на благо общего дела. Таким образом, огромная часть земного населения стала чувствовать благодарность по отношению к "Альянсу" и оказалась к нему гораздо более лояльной, чем к своим собственным правительствам.

Шло время, и "Альянс" стал примерять на себя роль всемирного правительства, тем более что национальные правительства постепенно утрачивали всякую власть. Национальные бюджеты продолжали снижаться, в то время как бюджет "Альянса" неуклонно рос, пока не достиг таких размеров, что в сравнении с ним совокупный бюджет всех правительств мира казался карликом. Функции национальных правительств казались все более избыточными и даже неуместными. К этому времени "Альянс" уже значитель-

но лучше обеспечивал социальную поддержку, поддерживал системы образования и развивал инфраструктуру. Международные конфликты усилиями СМИ практически рассосались, и тратиться на вооружение стало никому не нужно. Всеобщее процветание уничтожило все корни старых конфликтов, возникающих, в конечном итоге, из-за борьбы за ограниченные ресурсы. Несколько уцелевших пока диктаторов упорно сопротивлялись новому порядку и никак не покупались, но и их в конце концов смело умело срежиссированными переворотами или вооруженными восстаниями.

"Омега" завершала наиболее драматическое преобразование в человеческой истории. Впервые на всей планете устанавливалась единая власть, мощь которой многократно усиливалась интеллектом столь могучим, что он мог бы обеспечить процветание жизни на Земле и в окружающем нас космосе на миллиарды лет. Но в этом ли состоял ее план?

Так заканчивается сказание об "Омеге". Дальше речь пойдет совсем о другом, о том, о чем сказание еще не сложено: о нашем будущем с AI. Не хотели бы вы проиграть собственную версию? Может ли что-то похожее на сказание об "Омеге" произойти в действительности, и если да, то хотели бы вы этого? Если оставить в стороне все спекуляции вокруг нечеловеческих способностей AI, то как бы вам хотелось начать нашу историю? Как вы хотите, чтобы AI повлиял на структуру рынка труда, на наши законы, на наши

вооружения в ближайшие десятилетия? А заглядывая еще дальше вперед, каким концом вы бы увенчали свою историю? Это сказание – песчинка в космосе, воробьиный скок в истории жизни в нашей Вселенной. Подлинное сказание нам еще предстоит написать.

Глава 1
Добро пожаловать к самому важному разговору о нашем времени

Техника дает жизни возможность процветать как никогда прежде – или самоистребиться.
Институт будущего жизни

Через тринадцать целых восемь десятых миллиардов лет после своего рождения Вселенная наконец пробудилась и начала понимать, что действительно существует. С маленькой голубой планеты крошечная сознательная часть Вселенной начала вглядываться в космос своими телескопами, раз за разом открывая, что все, полагавшееся ею сущим, существовало лишь как крошечная часть чего-то значительно большего: Солнечная система, Галактика и Вселенная с сотнями миллиардов других галактик вписывались в стройную структуру групп, скоплений и сверхскоплений. Наделенные сознанием представители этой самой части Вселенной могли расходиться во мнениях по множеству вопросов, но все соглашались, что галактики прекрасны и вдохновляющи.

Но красота в глазу наблюдателя, а не в законах физики, поэтому до пробуждения Вселенной никакой красоты не бы-

ло. А наше космическое пробуждение тем более поразительно и достойно всяческих похвал: оно превратило нашу Вселенную из неразумного зомби, ничего о себе не понимающего, в живую экосистему, полную рефлексии, красоты и надежды, – а также стремящуюся к каким-то целям и ищущую какого-то смысла. Если бы Вселенная так никогда и не пробудилась, то она, по крайней мере для меня, была бы совершенно бессмысленной, просто гигантской пустотой. Если ей суждено когда-то навсегда вернуться в свою дремоту, в силу какого-то космического бедствия или же нашего саморазрушительного безумия, она, увы, снова лишится всякого смысла.

Но дело может повернуться значительно лучше. Мы до сих пор не знаем, нет ли в космосе другого места, где появились любители телескопов, и даже были ли наши телескопы первыми, но несмотря на это мы уже знаем о Вселенной так много, что даже поняли: у нее есть шанс пробудиться в значительно большей степени, чем это с ней случилось до сих пор. Возможно, мы и сами что-то вроде первых проблесков сознания, переживаемых по утрам, когда забвение постепенно сменяется предчувствием приближающегося полного пробуждения и открывающихся глаз. Возможно, жизнь распространится по нашему космосу и будет процветать миллиарды и триллионы лет, и случится это, возможно, благодаря тем самым решениям, которые мы примем на нашей маленькой планете в то время, пока здесь живем.

Краткая история сложности

Так откуда же это поразительное пробуждение? Оно не было единичным случайным событием, но стало лишь одним из звеньев неразрывной 13,8-миллиардолетней цепи трансформаций, делающих нашу Вселенную все более сложной и интересной, – они происходят и сейчас со все возрастающей скоростью.

Я чувствую, что мне крупно повезло, так как, став физиком, я большую часть последних 25 лет провел, соучаствуя в познании нашей космической истории, в этом захватывающем путешествии в неизведанное. Еще в то время, когда я работал над своей диссертацией, мы перестали спорить, 10 миллиардов лет нашей Вселенной или все 20, и стали спорить, равен ли ее возраст 13,7 миллиардов лет или все-таки ближе к 13,8: новые телескопы, новые компьютеры, новые теории сделали наше знание более точным. Мы, физики, до сих пор не знаем, что вызвало Большой взрыв и был ли он действительно началом всего, или всего лишь завершением какой-то предыдущей фазы. Однако мы получили довольно детальное знание о том, что произошло *после* Большого взрыва, благодаря настоящей лавине очень точных измерений, а потому позвольте мне в немногих словах подвести предварительный итог первым 13,8 миллиардам лет нашей космической истории.

Вначале был свет. Первое мгновение после Большого взрыва вся та часть пространства, которую наши телескопы в принципе могут наблюдать ("наша наблюдаемая Вселенная", или просто "наша Вселенная", как говорят для краткости), была горячее и ярче, чем ядро нашего Солнца, и к тому же она быстро расширялась. Кто-то, быть может, подумает, что это было то еще зрелище, но на самом деле оно было довольно унылым, в том смысле, что в нашей Вселенной тогда не было ничего, кроме безжизненного, очень плотного и горячего, скучно однообразного супа из элементарных частиц. Куда ни посмотри, со всех сторон было одно и то же; единственная интересная структура возникала из-за слабых, выглядящих случайными, звуковых волн, отчего суп в некоторых местах становился на 0,001 % плотнее, чем во всех прочих. Эта слабая волна, как принято думать, возникает из-за квантовых флуктуаций, поскольку в квантовой механике принцип неопределенности Гейзенберга запрещает чему бы то ни было становиться уж совсем скучным и везде одинаковым.

По мере того как наша Вселенная остывала, она становилась все менее однообразной: ее частицы объединялись во все более сложные объекты. В течение самой первой крошечной доли секунды сильное ядерное взаимодействие успело сгруппировать кварки в протоны (ядра водорода) и нейтроны, и некоторым из них понадобилось всего несколько минут, чтобы слиться в первые ядра гелия. Через 400 000 лет

электромагнитные силы привязали к этим ядрам электроны, и так возникли первые атомы. Вселенная все продолжала расширяться, поэтому атомы остывали и превращались в холодный темный газ. Наступившая темная ночь продлилась следующие 100 миллионов лет. Ей на смену пришел космический рассвет, когда сила тяготения успешно раскачала флуктуации в газе, прижав атомы друг к другу так, что возникли первые звезды и галактики. Эти первые звезды произвели так много тепла и света, что атомы водорода внутри них стали сливаться в более тяжелые – атомы углерода, кислорода и кремния. Когда эти звезды гибли, рожденные в их недрах атомы рассеивались в окружающем космосе, чтобы оказаться затем внутри планет, формирующихся близ звезд следующего поколения.

В какой-то момент группы атомов сложились таким образом, что образовавшийся комплекс смог поддерживать свою форму и даже скопировать себя. Скоро копий стало уже две, и процесс удвоения на этом не остановился. После всего лишь сорока циклов их количество достигло триллиона! Первый опыт самовоспроизводства оказался успешным и превратился в силу, с которой следовало считаться. Начиналась жизнь.

Три стадии жизни

В вопросе о том, что считать жизнью, как известно, давно уже нет никакого согласия. Предлагается огромное количество альтернативных определений, и некоторые из них включают довольно жесткие ограничения: например, требуется наличие клеточной структуры, что, вероятно, исключит из числа живых и будущие мыслящие машины, и некоторые внеземные цивилизации. Так как мы не хотим ограничивать свои размышления о будущем жизни теми биологическими видами, с которыми мы уже знакомы, то давайте примем более широкое ее определение, чтобы оно включало и любой иной процесс, если только он обладает сложностью и способностью к самовоспроизведению. Что именно воспроизводится, не так уж важно (состоит из атомов), важна информация (состоит из бит), которая определяется взаимным расположением атомов друг относительно друга. Когда бактерия копирует свою ДНК, не возникает никаких новых атомов, но имевшиеся атомы выстраиваются в цепочку, точно повторяющую исходную, таким образом копируется только информация. Иными словами, мы можем считать живой любую самовоспроизводящуюся и способную обрабатывать информацию систему, собственная информация которой (ее "программное обеспечение", "софт") определяет и ее поведение, и ее строение ("хард").

Вслед за самой Вселенной жизнь становилась все сложнее и интересней[4], и, как я сейчас поясню, мне представляется полезным ввести классификацию форм жизни по их соответствию трем степеням сложности: Жизнь 1.0, 2.0 и 3.0. Чем эти три формы отличаются друг от друга, в общих чертах хорошо видно на рис. 1.1.

[4] Почему жизнь усложнялась? Эволюция вознаграждает жизнь, когда та становится достаточно сложной, чтобы уметь обнаружить в окружающей среде повторяющиеся изменения и использовать их, поэтому в усложняющейся среде успешнее эволюционируют все более сложные и сознательные формы жизни. Усложняющаяся жизнь усложняет среду для конкурирующих с ней форм жизни, которым в свою очередь приходится эволюционировать и усложняться, постепенно создавая экосистему исключительно сложных форм.

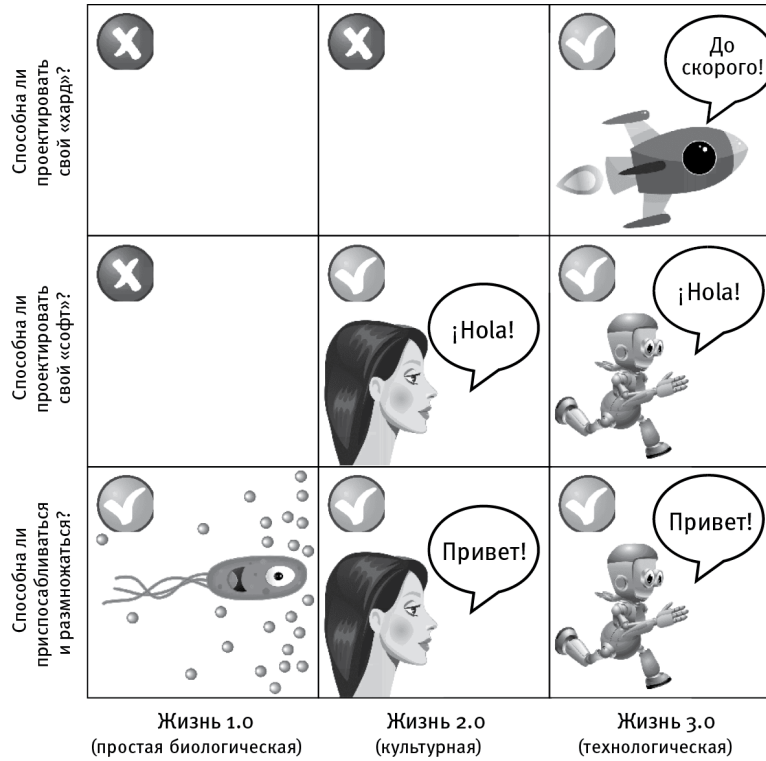


Рис. 1.1.

Три стадии жизни: биологическая эволюция, культурная эволюция и технологическая эволюция. Жизнь 1.0 не может влиять ни на "хард", ни на "софт" во время существования единичного организма: и то и другое определяется его ДНК, которая может изменяться от поколения к поколению на протяжении долгого периода эволюции. В отличие от это-

го, Жизнь 2.0 умеет переиначивать свой “софт”: люди приобретают многочисленные сложные навыки – учат языки, совершенствуются в спорте, осваивают профессии – они даже могут фундаментально пересматривать свой взгляд на мир и свои жизненные цели. Жизнь 3.0, которая пока еще не появилась на Земле, умеет радикально переиначивать не только “софт”, но и “хард”, не дожидаясь, пока он изменится эволюционным путем через ряд поколений.

До сих пор остается открытым вопрос, как, когда и где в нашей Вселенной впервые появилась жизнь, но у нас есть веские основания думать, что на Земле она впервые появилась 4 миллиарда лет назад. Прошло немного времени, и планету наводнили бесчисленные разновидности разнообразных форм жизни. Наиболее успешные из них быстро обогнали прочих, потому что в каком-то смысле лучше приспосабливались к изменениям окружающей среды. Строго говоря, они оказались, если воспользоваться терминологией современной информатики, “интеллектуальными агентами” – так называют сущности, которые собирают информацию об окружающей среде через систему своих сенсоров, а затем, перерабатывая эту информацию, принимают решение, каким должно быть их ответное действие на среду. Эта переработка может оказаться довольно сложным процессом – вроде того, который совершается в вашем мозгу, когда, опираясь на информацию от ваших ушей и глаз, вы решаете, что

ответить собеседнику. Но иногда для этого требуются и совсем несложные "хард" и "софт".

Например, у многих бактерий есть органы чувств, позволяющие им измерять концентрацию сахара в окружающей жидкости, в который они плавают с помощью напоминающих пропеллеры жгутиков. "Хард", прикрепляющий этот орган чувств к жгутику, может следовать весьма простому, но полезному алгоритму: "Если мои органы чувств сообщают, что концентрация сахара сейчас стала вдвое меньше, чем несколько секунд назад, то направление вращения пропеллера должно поменяться на противоположное, чтобы я поплыла в другую сторону".

Обучение разовьет бесчисленное количество подобных навыков. Но бактерии, с другой стороны, не очень сильны в обучении. В их ДНК заложена информация не только о строении их "харда" – сенсоров концентрации сахара и жгутиков, но и их "софта". Им не надо учиться плыть в ту сторону, где больше сахара: этот алгоритм "зашит" в их ДНК с самого начала. Этому, конечно, предшествовал своего рода процесс обучения, но он никак не продолжается в жизни данной конкретной бактерии. Правильнее сказать, что это происходило в процессе предшествующей эволюции данного вида бактерий, включавшей пробы и ошибки многих поколений, пока естественный отбор не запечатал в ДНК те мутации, которые особенно полезны при потреблении сахара. Некоторые из этих мутаций благотворно отразились также

на конструкции жгутиков и иного “харда”, в то время как прочие совершенствовали алгоритмы переработки информации, способствующие успешному поиску сахара, и другие разновидности “софта”.

Такие бактерии служат примером того, что я называю “Жизнью 1.0”: *форма жизни, при которой и “хард”, и “софт” эволюционируют, а не конструируются.* Мы с вами служим примером того, что я называю “Жизнью 2.0”: *форма жизни, при которой “хард” эволюционирует, а “софт” в значительной степени конструируется.* Под вашим “софтом” я подразумеваю те алгоритмы и те знания, которые вы используете, перерабатывая информацию от органов чувств и решая, что делать, – то есть все, от способности узнавать друга при встрече до умения ходить, читать, писать, считать, петь песни и смеяться шуткам.

Вы были не в состоянии делать все это в момент рождения, так что весь этот “софт” загрузился в ваши мозги позже, в ходе процесса, который мы называем обучением. И хотя ваш детский куррикулум в основном конструируется вашими родителями и учителями, со временем вы постепенно приобретаете достаточно сил, чтобы самостоятельно разрабатывать свой “софт”. Может быть, ваша школа оставляет за вами право выбирать, какой иностранный язык учить: хотите ли вы загрузить в свой мозг программный модуль, который позволит вам говорить по-французски, или же предпочтете тот, который даст вам возможность говорить по-испански?

Вы хотите учиться играть в теннис или в шахматы? Вы хотите учиться на повара, на адвоката или на фармацевта? Хотите ли вы узнать больше об искусственном интеллекте (AI) и о будущем жизни, читая эту книгу?

Способность Жизни 2.0 создавать собственный "софт" дает ей много возможностей, которых нет у Жизни 1.0. Для бóльшего ума требуется бóльший "хард" (больше атомов) и бóльший "софт" (больше бит). Тот факт, что бóльшая часть нашего человеческого "харда" приобретается уже после нашего рождения (во время роста), имеет значение, поскольку конечный размер нашего тела не ограничивается шириной родовых каналов матери. Сходным образом полезен и тот факт, что бóльшая часть нашего человеческого "софта" также приобретается уже после нашего рождения (во время обучения), так как способности нашего конечного разума не ограничиваются пропускной способностью информационного канала при конструировании новой ДНК в момент зачатия в стиле 1.0. Я сейчас вешу в 25 раз больше, чем при рождении, а синапсы нейронной сети моего мозга способны хранить в 100 000 раз больше информации, чем ДНК, с которой я родился. Ваши знания и умения "весят", грубо говоря, около 100 терабайт информации, а в вашу ДНК едва умещается гигабайт, которого не хватит для сохранения одного полнометражного фильма. Младенцу физически невозможно родиться с совершенным знанием английского или быть готовым сразить всех на вступительных экзаменах

в колледж: не существует способа загрузить в его мозг необходимую информацию, так как основной модуль, полученный им от родителей (его ДНК), не обладает достаточной вместительностью.

Способность создавать собственный "софт" обеспечивает Жизни 2.0 не только бóльшую разумность, но и бóльшую гибкость. При изменениях окружающей среды Жизнь 1.0 может только медленно эволюционировать на протяжении многих поколений. А Жизнь 2.0 способна почти моментально адаптироваться, обновляя загруженный "софт". У бактерий, часто сталкивающихся с антибиотиком, со временем вырабатывается резистентность, на это требуется много поколений – никакая отдельная бактерия не может изменить своего поведения. Напротив, девочка, обнаружив, что у нее аллергия на арахисовое масло, немедленно изменит стиль жизни и будет впредь избегать его. Такая гибкость дает Жизни 2.0 даже больше преимуществ на популяционном уровне: хотя информация, записываемая в наших ДНК, мало изменилась за последние 50 000 лет, ее суммарное количество, накопленное в наших мозгах, книгах и компьютерах, росло как лавина. Установив себе "софт", обеспечивающий коммуникацию посредством развитой устной речи, мы получили возможность копировать наиболее полезную информацию, накопленную в мозгу одного человека, в мозги других людей, благодаря чему она сохраняется даже после гибели ее источника. Установив себе "софт", позволяющий читать

и писать, мы получили возможность хранить гораздо больше информации, чем способны запомнить, и обеспечивать к ней доступ другим. Установив в свой мозг "софт", обеспечивающий развитие технологий (то есть изучая точные и технические науки), мы открыли доступ почти ко всей накопленной в мире информации очень большому числу человеческих особей, для чего им достаточно лишь нажать несколько кнопок.

Именно эта гибкость позволила Жизни 2.0 покорить Землю. Освобожденные от генетических оков, совокупные знания человечества нарастали со все возрастающей скоростью, когда каждый предшествующий прорыв готовит последующий: язык, письменность, книгопечатание, современная наука, компьютеры, интернет… Разгоняющаяся культурная эволюция нашего совместного "софта" стала определяющей силой нашего человеческого будущего, оставив замороженно-заторможенной биологической эволюции роль практически эпизодическую.

Однако, несмотря на все достижения технологии, какие только у нас на сегодня есть, все формы жизни, о которых нам известно, остаются фундаментально ограниченными своим биологическим "хардом". Никто не может жить миллион лет, выучить наизусть всю Википедию, понять все известное науке или слетать в космос без звездолета. Ничто не может превратить наш в основном безжизненный космос в цветущую биосферу, полную жизни на миллиарды и трил-

лионы лет, позволяя нашей Вселенной полностью раскрыть свой потенциал и окончательно пробудиться.

Границы между этими тремя стадиями немного размыты. Если бактерии – это Жизнь 1.0, а мы – Жизнь 2.0, то мышей следует считать Жизнью 1.1: они могут многому научиться, но все же недостаточно для того, чтобы освоить язык или придумать интернет. Более того, раз у них нет языка, выученное одной мышкой в основном теряется с ее смертью и не передается следующим поколениям. Подобным образом вы можете сказать, что современных людей можно отнести к Жизни 2.1: небольшой апгрейд нашего "харда" нам уже становится доступен – вроде имплантированных зубов, коленных чашечек или кардиостимулятора. Но ничего по-настоящему стоящего: вы не можете стать в десять раз выше или получить в тысячу раз больше мозгов.

Короче говоря, мы классифицируем жизнь по трем стадиям в зависимости от ее способности к самодизайну:

1. Жизнь 1.0 (биологическая стадия): эволюция "харда" и "софта";
2. Жизнь 2.0 (культурная стадия): эволюция "харда" и дизайн большей части "софта";
3. Жизнь 3.0 (технологическая стадия): дизайн и "софта", и "харда".

После 13,8 миллиардов лет космической эволюции события самым драматическим образом ускорились: Жизнь 1.0

возникла на Земле около 4 миллиардов лет назад, Жизнь 2.0 (мы, люди) появились тут около ста тысячелетий назад, и вот теперь многие AI-эксперты уверены, что Жизнь 3.0 появится уже в начинающемся столетии, возможно даже еще на наших глазах, если ускоряющееся развитие AI ей это позволит. Как это может случиться, и что это означает для нас? Об этом наша книга.

Контроверзы

Поставленный вопрос – повод для полемики, даже больше того – для контроверзы. Ведущие AI-эксперты не только кардинально расходятся в своих мнениях, но даже их эмоциональные оценки грядущего диаметрально противоположны – от уверенного оптимизма до серьезной озабоченности. Среди них нет согласия даже относительно краткосрочных прогнозов об AI-экономике, последствиях для правовых отношений и новых вооружений, и эти расхождения заметно возрастают, если расширить временной горизонт и поставить вопрос о сильном искусственном интеллекте (AGI), достигающем человеческого уровня или превосходящем его и потому открывающем возможность для Жизни 3.0. Сильный искусственный интеллект решает практически любую задачу, в том числе способен к самообучению, в отличие от слабого искусственного интеллекта, вроде того что успешно играет в шахматы.

Примечательно, что контроверза относительно искусственного интеллекта имеет своим центром не один, а два разных вопроса: “когда?” и “что?”. Когда это случится (если такое вообще может случиться), и что оно может означать для человечества? На мой взгляд, можно выделить три направления, к каждому из которых следует отнестись серьезно, поскольку они представлены выдающимися мировыми

мыслителями. Я изобразил эти направления на рис. 1.2, дав каждому свое наименование: *цифро-утописты, техноскептики* и *участники движения за дружественный AI.* А теперь позвольте мне дать характеристику наиболее ярким представителям из каждого лагеря.

Рис. 1.2.

Большинство споров вокруг сильного искусственного ин-

теллекта (не уступающего человеческому в любом виде деятельности) вращаются около двух вопросов: когда (если такое вообще возможно) он появится и будет ли его появление благоприятно для человечества. Техноскептики и цифро-утописты соглашаются, что поводов для беспокойства у нас нет, но по совершенно различным причинам: первые убеждены, что универсальный искусственный интеллект человеческого уровня (AGI) в обозримом будущем не появится, вторые не сомневаются в его появлении, но убеждены, что оно практически гарантированно будет благоприятно для человечества. Представители движения за дружественный AI соглашаются, что озабоченность уместна и продуктивна, потому что исследования в области AI-безопасности и публичные обсуждения связанных с ней вопросов повышают вероятность благоприятного исхода. Луддиты убеждены в скверном исходе и протестуют против искусственного интеллекта. Отчасти этот рисунок навеян публикацией: https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html (проверена 12.05.2018).

Цифро-утописты

Ребенком я был уверен, что все миллиардеры просто сочатся помпезностью и невежеством. Когда в 2008 году в Google я впервые встретил Ларри Пейджа, он напрочь

разрушил оба этих стереотипа. Впечатление подчеркнутой небрежности в одежде создавалось джинсами и ничем не примечательной майкой, словно он собрался на университетский пикник. Задумчивая манера говорить и мягкий голос, скорее, успокоили и расслабили меня, чем напугали и напрягли. 18 июля 2015 года мы снова увиделись в Напа-Вэлли на вечеринке, устроенной Илоном Маском и его тогдашней женой Талулах, и между нами немедленно завязался разговор о копрологических интересах наших детей. Я порекомендовал ему литературную классику Энди Гриффитса – *The Day My Butt Went Psycho,* которую он тут же себе заказал. Мне пришлось напомнить себе, что этот человек, вероятно, войдет в историю как оказавший на нее наибольшее влияние: если, как я думаю, сверхчеловеческому искусственному интеллекту суждено просочиться во все уголки нашей Вселенной еще при моей жизни, то это может случиться исключительно благодаря решениям Ларри.

С нашими женами, Люси и Мейей, мы отправились ужинать, и во время еды мы обсуждали, непременно ли машины будут обладать сознанием – идея, как он утверждал, совершенно ложная и пустая. А уже ночью, после коктейля, у них с Илоном разразился длинный и бурный спор о будущем искусственно интеллекта. Уже близился рассвет, а толпа любопытных зевак вокруг них продолжала расти. Ларри яростно защищал позицию, которую я бы отождествил с *цифро-утопистами*: он говорил, что цифровая жизнь – естественный

и желательный новый этап космической эволюции, и если мы дадим ей свободу, не пытаясь удушить или поработить ее, то это принесет безусловную пользу всем. На мой взгляд, Ларри самый яркий и последовательный выразитель идей цифро-утопизма. Он утверждал, что если жизни суждено распространиться по всей Вселенной, в чем сам он был убежден, то произойти это может только в цифровом виде. Самую большую тревогу у него вызывала опасность, что AI-паранойя способна затормозить наступление цифровой утопии и даже спровоцировать попытку силой овладеть искусственным интеллектом в нарушение главного лозунга Google "Не твори зла!". Илон старался вернуть Ларри на землю и без конца спрашивал его, откуда такая уверенность, что цифровая жизнь не уничтожит вокруг все то, что нам дорого. Ларри то и дело принимался обвинять Илона в "видошовинизме" – стремлении приписать более низкий статус одним формам жизни в сравнении с другими на том простом основании, что главный химический элемент в их молекулах кремний, а не углерод. Мы еще вернемся к подробному обсуждению этих важных аргументов ниже, начиная с главы 4.

Хотя в тот вечер у бассейна Ларри оказался в меньшинстве, у цифровой утопии, в защиту которой он так красноречиво выступал, немало выдающихся сторонников. Робототехник и футуролог Ганс Моравец[5] своей книгой *Mind*

[5] В англоязычном мире имя этого человека принято транслитерировать иначе: "Моравек", но мы будем придерживаться исходного, чешского варианта. – *Прим.*

Children 1988 года, ставшей классикой жанра, воодушевил целое поколение цифро-утопистов. Его дело было подхвачено и поднято на новую высоту Рэем Курцвейлом. Ричард Саттон, один из пионеров такой важной AI-отрасли, как машинное обучение, выступил со страстным манифестом цифро-утопистов на нашей конференции в Пуэрто-Рико, о которой я скоро расскажу.

Техноскептики

Следующая группа мыслителей тоже мало беспокоится по поводу AI, но совсем по другой причине: они думают, что создание сверхчеловечески сильного искусственного интеллекта настолько сложно технически, что никак не может произойти в ближайшие сотни лет, и беспокоиться об этом сейчас просто глупо. Я называю такую позицию *техноскептицизмом,* и ее предельно красноречиво сформулировал Эндрю Ын: “Бояться восстания роботов-убийц – все равно что переживать по поводу перенаселения Марса”. Эндрю был тогда ведущим специалистом в Baidu, китайском аналоге Google, и он недавно повторил этот аргумент во время нашего разговора в Бостоне. Он также сказал мне, что предчувствует потенциальный вред, исходящий от разговоров об AI-рисках, так как они могут замедлить развитие всех AI-исследований. Подобные мысли высказывает и Родни Брукс,

---

перев.

бывший профессор MIT[6], стоявший за созданием роботизированного пылесоса Румба и промышленного робота Бакстера. Мне представляется любопытным тот факт, что хотя цифро-утописты и техноскептики сходятся во взглядах на исходящую от AI угрозу, они не соглашаются друг с другом почти ни в чем другом. Большинство цифро-утопистов ожидают появления сильного AI (AGI) в период от двадцати до ста лет, что, по мнению техноскептиков, – ни на чем не основанные пустые фантазии, которые они, как и все предсказания технологической сингулярности, называют "бреднями гиков". Когда я в декабре 2014 года встретил Родни Брукса на вечеринке, посвященной дню его рождения, он сказал мне, что на 100 % убежден, что ничего такого не может случиться при моей жизни. "Ты уверен, что имел в виду не 99 %?" – спросил я его потом в электронном письме. "Никаких 99 %. 100 %. Этого просто не случится, и все".

Движение за дружественный AI

Впервые встретив Стюарта Рассела в парижском кафе в июне 2014 года, я подумал: "Вот настоящий британский джентльмен!". Выражающийся ясно и обдуманно, с мягким красивым голосом, с авантюрным блеском в глазах, он показался мне современной инкарнацией Филеаса Фогга, люби-

[6] Массачусетский технологический институт в Кембридже (штат Массачусетс). – *Прим. перев.*

мого мною в детстве героя классического романа Жюля Верна *Вокруг света за 80 дней*. Хотя он один из самых известных среди ныне живущих исследователей искусственного интеллекта, соавтор одного из главных учебников на этот счет, его теплота и скромность позволили мне чувствовать себя легко во время беседы. Он рассказал, как прогресс в исследованиях искусственного интеллекта привел его к убеждению, что появление AGI человеческого уровня уже в этом столетии – вполне реальная возможность, и хотя он с надеждой смотрит на будущее, благополучный исход не гарантирован. Есть несколько ключевых вопросов, на которые мы должны ответить в первую очередь, и они так сложны, что исследовать их нужно начинать прямо сейчас, иначе ко времени, когда понадобится ответ, мы можем оказаться не готовы его дать.

Сегодня взгляды Стюарта в той или иной степени разделяет большинство, и немало групп по всему миру занимаются вопросами AI-безопасности, как он и призывал. Но так было не всегда. В статье *Washington Post* 2015 год назван годом AI-безопасности. А до тех пор рассуждения о рисках, связанных с разработками искусственного интеллекта, вызывали раздражение у большинства исследователей, относившихся к ним как к призывам современных луддитов воспрепятствовать прогрессу в этой области. Как мы увидим в главе 5, опасения, подобные высказанным Стюартом, были достаточно отчетливо артикулированы еще более полувека назад разработчиком первых компьютеров Аланом Тьюрингом и

математиком Ирвингом Гудом, который работал с Тьюрингом над взломом германских шифров в годы Второй мировой войны. В прошлом десятилетии такие исследования велись лишь горсткой мыслителей, не занимавшихся созданием AI профессионально, среди них, например, Элиезер Юдковски, Майкл Вассар и Ник Бострём[7]. Их работа мало влияла на исследователей AI-мейнстрима, которые были полностью поглощены своими ежедневными задачами по улучшению "умственных способностей" разрабатываемых ими систем и не задумывались о далеких последствиях своего возможного успеха. Среди них я знал и таких, кто испытывал определенные опасения, но не рисковал говорить о них публично, дабы не навлечь на себя обвинений коллег в алармизме и технофобии.

Я чувствовал, что такая поляризация мнений не навсегда и что исследовательское сообщество должно рано или поздно примкнуть к обсуждению вопроса, как сделать AI дружественным. К счастью, я был не одинок. Весной 2014 года я основал некоммерческую организацию под названием "Институт будущего жизни" (Future of Life Institute, или, сокращенно, FLI; http://futureoflife.org), в чем мне помогали моя жена Мейя, мой друг физик Энтони Агирре, аспирантка Гарварда Виктория Краковна и создатель Skype Яан Таллин. На-

[7] Фамилию этого шведского ученого, живущего и работающего в Англии, часто транскрибируют без учета умляута над вторым "о" – "Бостром". Более точной и более традиционной была бы транскрипция "Бустрём", но мы решили остановиться на компромиссном варианте "Бострём". – *Прим. перев.*

ша цель проста: чтобы у жизни было будущее и чтобы оно было, насколько это возможно, прекрасно! В частности, мы понимали, что развитие технологий дает жизни небывалые возможности, она может теперь либо процветать, как никогда ранее, либо уничтожить себя, и мы бы предпочли первое.

Наша первая встреча состоялась у нас дома 15 марта 2014 года и приняла характер мозгового штурма. В ней участвовало около 30 студентов и профессоров MIT, а также несколько мыслителей, живущих по соседству с Бостоном. Мы все пришли к согласию в том, что, хотя необходимо уделять некоторое внимание биотехнологии, ядерному оружию и климатическим изменениям, наша главная цель – сделать вопрос AI-безопасности центральным в данной исследовательской области. Мой коллега по MIT физик Фрэнк Вильчек, получивший Нобелевскую премию за то, что разобрался, как работают кварки, предложил нам для начала выступить с авторской колонкой в каком-нибудь популярном СМИ, чтобы привлечь к проблеме внимание и усложнить жизнь тем, кто хотел бы ее проигнорировать. Я связался со Стюартом Расселом (с которым тогда еще не был знаком) и со Стивеном Хокингом, еще одним моим коллегой-физиком, и они оба согласились присоединиться к нам с Фрэнком в качестве соавторов этой колонки. Что бы ни говорили об этом позже, но тогда *New York Times* отказалась публиковать нашу колонку, а за ней многие другие американские газеты, так что в итоге мы разместили ее в моем блоге на

Huffington Post. Сама Арианна Хаффингтон, к моей радости, откликнулась письмом в электронной почте: “Я в восторге от поста! Мы поместим его #1!”, и это размещение нашей заметки в самой верхней позиции главной страницы вызвало лавину публикаций о AI-безопасности, заняв первые полосы многих изданий до конца года, с участием Илона Маска, Билла Гейтса и других знаковых фигур современных технологий. Книга Ника Бострёма *Superintelligence*, вышедшая той же осенью, подлила масла в огонь и еще больше подогрела публичную дискуссию.

Следующим шагом нашей кампании за дружественный AI, проводимой под эгидой Института, стала организация большой конференции при участии всех ведущих специалистов по искусственному интеллекту с целью разобраться во всех недоразумениях, достичь принципиального согласия и составить конструктивный план на будущее. Мы хорошо понимали, что убедить столь блистательную публику собраться на конференцию, организуемую неизвестными им аутсайдерами, будет не просто, и поэтому старались изо всех сил: мы запретили доступ на нее любым СМИ, тщательно выбрали место и время – январь месяц и пуэрто-риканский пляж, сделали бесплатным участие (нам позволила это щедрость Яана Таллина), мы придумали для нее наименее тревожное название, на какое только были способны: “Будущее AI: возможности и опасности”. Но самое главное – мы скооперировались со Стюартом Расселом и благодаря

ему сумели пригласить в организационный комитет нескольких ведущих специалистов по искусственному интеллекту как из университетской среды, так и от бизнеса. В их числе был Демис Хассабис из лаборатории DeepMind, который как раз только что показал, что искусственный интеллект может обыграть человека даже в такую игру, как го. И чем больше я узнавал Демиса, тем больше понимал, что среди его амбициозных целей не только увеличение мощности искусственного интеллекта, но и достижение его дружественности.

Результатом стала невероятная встреча замечательных умов (см. рис. 1.3). К специалистам по искусственному интеллекту присоединились лучшие экономисты, юристы, лидеры технологии (включая Илона Маска) и иные мыслители (в том числе Вернор Виндж, придумавший термин "сингулярность", вокруг которого все будет построено в главе 4). В конце концов все сложилось лучше, чем в наших самых смелых мечтах. Вероятно, все дело в удачном сочетании вина и солнца, а может быть, просто правильно было выбрано время: несмотря на полемическую повестку конференции, возник замечательный консенсус, изложенный в итоговом письме[1]8, которое подписали более восьми тысяч человек, включая тех, без чьих имен не обойдется ни один справочник. Смысл письма заключался в том, что цель разработок искусственного интеллекта следует переопределить: со-

[8] Здесь и ниже цифрами отмечены примечания, помещенные в конце книги. – *Прим. перев.*

здаваемый интеллект не должен быть неконтролируемым, он должен быть дружественным. В письме формулировался подробный список исследовательских задач, вокруг которых участники конференции соглашались концентрировать свою работу. Дружественный AI начинал превращаться в мейнстрим. Мы проследим за его прогрессом в этой книге.

Рис. 1.3.

На конференцию в Пуэрто-Рико в январе 2015 года собралась замечательная группа исследователей различных аспектов искусственного интеллекта и смежных вопросов. В заднем ряду слева направо: Том Митчел, Шон О'х Эйгертейг, Хью Прайс, Шамиль Чандариа, Яан Таллин, Стюарт Рассел, Билл Хиббард, Блез Агуэра-и-Аркас, Андерс Зандберг, Дэниел Дьюи, Стюарт Армстронг, Льюк Мюльхойзер, Том Диттерих, Майкл Озборн, Джеймс Манийка, Аджай

Агравал, Ричард Маллах, Ненси Чан, Мэтью Путман; Стоящие ближе, слева направо: Мэрилин Томпсон, Рич Саттон, Алекс Виснер-Гросс, Сэм Теллер, Тоби Орд, Йоша Бах, Катья Грейс, Адриан Веллер, Хизер Рофф-Перкинс, Дилип Джордж, Шейн Легг, Демис Хассабис, Вендель Валлах, Чарина Чой, Илья Суцкевер, Кент Уокер, Сесилия Тилли, Ник Бострём, Эрик Бриньоулфссон, Стив Кроссан, Мустафа Сулейман, Скотт Феникс, Нил Джейкобстейн, Мюррей Шанахан, Робин Хэнсон, Франческа Росси, Нейт Соареш, Илон Маск, Эндрю Макафи, Барт Зельман, Микеле Рэйли, Аарон Ван-Девендер, Макс Тегмарк, Маргарет Боден, Джошуа Грин, Пол Кристиано, Элиезер Юдковски, Дэвид Паркес, Лоран Орсо, Дж. Б. Шробель, Джеймс Мур, Шон Легассик, Мейсон Хартман, Хоуи Лемпель, Дэвид Владек, Джейкоб Стейнхардт, Майкл Вассар, Райан Кало, Сьюзан Янг, Оувейн Эванс, Рива-Мелисса Тец, Янош Крамар, Джофф Андерс, Вернор Виндж, Энтони Агирре; Сидят: Сэм Харрис, Томазо Поджо, Марин Сольячич, Виктория Краковна, Мейя Чита-Тегмарк. За камерой – Энтони Агирре (им также выполнена обработка фотографии в Фотошопе при содействии сидящего рядом искусственного интеллекта человеческого уровня).

В ходе конференции мы получили еще один урок: успех в создании искусственного интеллекта не просто будоражит мысль, он ставит серьезные вопросы, имеющие огромное мо-

ральное значение – от ответов на них зависит будущее всего живого, всей жизни. В прошлом моральная значимость принимаемых людьми решений могла быть очень велика, но она всегда оставалось ограниченной: человечество оправилось от самых жутких эпидемий, и даже величайшие империи в конце концов развалились. Прошлые поколения могли быть уверены, что наступит завтра и придут новые люди, пережившие обычные беды нашего мира: бедность, болезни, войны. И на конференции в Пуэрто-Рико были такие, кто говорил, что сейчас настало другое время: впервые, говорили они, мы можем построить достаточно мощную технологию, которая способна навсегда избавить мир от этих бед – или от самого человечества. Мы можем создать общества, которые будут процветать, как никогда ранее, на Земле и, возможно, не только, а можем создать кафкианское, за всеми следящее государство, от которого уже никогда не удастся избавиться.

Рис. 1.4.

Хотя СМИ часто изображают дело так, словно Илон Маск на ножах с AI-сообществом, на самом деле существует практически единодушное согласие по поводу необходимости исследований в области безопасности искусственного интеллекта. Здесь на фотографии 4 января 2015 года президент Ассоциации за развитие искусственного интеллекта Том Диттерих делится радостью со стоящим рядом Илоном Маском, который только что объявил о своем намерении финансировать новую программу исследований по AI-безопасности. У них из-за спин выглядывают сооснователи Института будущего жизни (FLI) Мейя Чита-Тегмарк и Виктория

Краковна.

Недоразумения

Я покидал Пуэрто-Рико в убеждении, что начатый тут разговор о будущем AI следует продолжить, потому что это самый важный разговор нашего времени[9]. Причем этот разговор касается нашего общего будущего, и поэтому не должен вестись только в кругу AI-экспертов. Именно поэтому я и написал эту книгу: я писал ее в надежде, дорогой читатель, что и вы присоединитесь к этому разговору! На какое будущее вы надеетесь? Надо ли нам развивать автономное летальное оружие? Какого рода автоматизации хотели бы вы у себя на работе? Какого образования вы бы хотели для своих детей? Предпочли бы вы, чтобы на смену старым профессиям пришли новые, или вам больше нравится общество бездельников, где каждый наслаждается досугом, а богатство создается машинами? Двигаясь дальше в том же направлении, нахо-

[9] Необходимость здесь двоякая: и по силе воздействия, и по насущности проблемы. В сравнении с климатическими изменениями, катастрофические последствия которых ожидаются в период от пятидесяти до двухсот лет, считая от настоящего момента, AI, по оценкам некоторых экспертов, может привести к столь же пагубным последствиям в течение ближайших десятилетий – при этом он же может создать технологию смягчения климатических изменений. А в сравнении с войнами, терроризмом, безработицей, нищетой, миграцией и нарушениями прав человека – общий эффект от создания и внедрения AI может значительно превысить совокупные последствия всего перечисленного (мы как раз и собираемся разобраться ниже, каким именно образом AI может здесь перевесить), причем как усугубляя, так и компенсируя его.

дите ли вы благоприятной перспективу создания Жизни 3.0 и ее распространения по нашему космосу? Сможем ли мы управлять мыслящими машинами, или это они скорее начнут управлять нами? Заменят ли думающие машины нас, будем ли мы сосуществовать друг с другом или объединимся в некое единое целое? Что значит оставаться людьми в эпоху искусственного интеллекта? Какой ответ на этот вопрос вам представляется желательным, и как вы представляете себе возможность реализации этого ответа в нашей будущей жизни?

Цель этой книги – помочь вам присоединиться к нашему разговору. В нем разворачиваются увлекательнейшие контроверзы, о которых я уже упоминал и в которых величайшие мировые умы придерживаются противоположных позиций. Но кроме этого я был свидетелем множества скучнейших псевдо-контроверз, когда люди просто не понимают или даже не слушают друг друга. Для того чтобы сконцентрироваться на важных нерешенных проблемах, говорить о контроверзах, понимаемых спорящими одинаково, давайте для начала избавимся от некоторых расхожих недоразумений.

У таких важных и часто используемых понятий, как “жизнь”, “разум” и “сознание”, есть много общеупотребительных и конкурирующих определений, и недоразумения нередко возникают по вине людей, не отдающих себе отчет в том, что они используют одно и то же слово в двух разных значениях. Для того чтобы мы с вами не сваливались раз за

разом в эту волчью яму, я составил шпаргалку (см. табл. 1.1), показывающую, как я использую те или иные слова в этой книге. Некоторые из содержащихся в ней определений будут даны и должным образом объяснены в следующих главах. И, пожалуйста, обратите внимание, что я не претендую на какую-то исключительность моих определений – они, может быть, ничуть не лучше, чем какие-то другие, но моя единственная цель состоит в том, чтобы избежать недоразумений, сразу выразив предельно ясно, что именно я имею в виду. Вы увидите, что я обычно отдаю предпочтение широким определениям, избегая антропоцентрического уклона и делая их приложимыми как к людям, так и к машинам. Пожалуйста, прочитайте мою шпаргалку сейчас и не ленитесь обращаться к ней потом, когда будете обескуражены тем, как я использовал то или иное слово – в особенности в главах 4–8.

Таблица 1.1

Терминологическая шпаргалка	
Жизнь	Самовоспроизводящийся процесс, сохраняющий свою сложность
Жизнь 1.0	Жизнь, при которой изменения “харда” и “софта” происходят эволюционным путем (биологическая фаза)
Жизнь 2.0	Жизнь, при которой изменения “харда” происходят эволюционным путем, но “софт” отчасти конструируется (культурная фаза)
Жизнь 3.0	Жизнь, которая способна конструировать и свою материальную составляющую, “хард”, и свою информационную составляющую, “софт” (технологическая фаза)
Интеллект = разум	Способность достигать сложных целей
Искусственный интеллект	Интеллект небиологического происхождения
Слабый (ограниченный) интеллект	Способность достигать цели из определенного (ограниченного) множества: например, играть в шахматы или управлять автомобилем
Сильный интеллект	Способность достигать практически любую цель, включая самообучение
Универсальный интеллект	Способность стать сильным интеллектам при наличии доступа к данным и ресурсам
Сильный искусственный интеллект [человеческого уровня] (AGI)	Способность справиться с любой познавательной задачей по крайней мере не хуже, чем человек
AI человеческого уровня	AGI
Сильный AI	AGI
Сверхразум	AGI, значительно превосходящий человеческий

Цивилизация	Взаимодействующая группа разумных форм жизни
Сознание	Личные переживания
Квалиа	Индивидуальные свойства личных переживаний
Этика	Направляющие поведение принципы
Телеология	Объяснение происходящего намерениями или целями, а не причинами
Целенаправленное поведение	Поведение, которое проще объяснить его результатами, чем его причинами
Иметь цель	Демонстрировать целенаправленное поведение
Иметь задачу	Обслуживать свои собственные цели или цели какой-то иной сущности
Дружественный AI	Сверхразум, цели которого приведены в соответствие с нашими
Киборг	Человеко-машинный гибрид
Интеллектуальный взрыв	Рекурсивное самосовершенствование, быстро приводящее к появлению сверхразума
Сингулярность	Интеллектуальный взрыв
Вселенная	Часть пространства, откуда свет мог нас достичь за 13,8 млрд лет с момента Большого взрыва

Многие недоразумения относительно искусственного интеллекта возникают из-за того, что люди используют приведенные в левой колонке слова для обозначения несхожих вещей. Здесь я привожу значения, в которых эти слова упо-

требляются в этой книге. (Некоторые из этих определений будут введены и объяснены только в следующих главах книги.)

Кроме недоразумений, вызванных расхождениями в терминологии, я был свидетелем споров, возникавших по причине простых логических ошибок. Рассмотрим наиболее распространенные из них.

Хронологические мифы

Первый проиллюстрирован на рис. 1.5: сколько времени понадобится, чтобы машинный интеллект мог принципиально превзойти человеческий разум? Самая большая ошибка здесь заключается в уверенности, что мы можем знать это с большой степенью точности.

Так, один популярный миф утверждает, что мы можем не сомневаться в появлении суперинтеллекта к концу этого столетия. В самом деле, история полна примеров чрезмерного оптимизма в отношении технологических достижений будущего. Где все эти давно обещанные нам термоядерные электростанции и летающие автомобили? С AI в прошлом тоже было связано немало чрезмерно завышенных ожиданий, в том числе этим грешили и некоторые основатели самой этой области: например, Джону Маккарти (автору термина “искусственный интеллект”), Марвину Мински, Ната-

ниелю Рочестеру и Клоду Шеннону принадлежит следующий пассаж, содержащий оптимистический прогноз относительно того, что может быть проделано при помощи двух компьютеров каменного века за два месяца: “Наш проект заключается в том, чтобы 10 человек проводили на протяжении двух месяцев летом 1956 года исследование искусственного интеллекта в Дартмутском колледже … Будет сделана попытка научить машины использовать язык, формировать абстракции и общие понятия, решать некоторые типы задач, в настоящее время доступных только людям, и самосовершенствоваться. Мы полагаем, что в одном или нескольких из предложенных направлений может быть достигнут существенный прогресс, если тщательно отобранная группа ученых будет заниматься ими на протяжении лета”.

Миф
Искусственный сверхинтеллект к 2100 году неотвратим

Миф
Искусственный сверхинтеллект к 2100 году немыслим

Факт
Это может случиться через несколько десятилетий, несколько веков, а может не случиться никогда. У экспертов нет согласия по этому поводу, так что мы просто не знаем ответа на этот вопрос

Миф
По поводу AI тревожатся только луддиты

Факт
Основания испытывать озабоченность находят многие ведущие специалисты

Мифический повод для тревоги
Искусственный интеллект будет злонамерен

Мифический повод для тревоги
Искусственный интеллект обретет сознание

Реальный повод для тревоги
Цели искусственного интеллекта могут оказаться несовместимы с нашими

Миф
Главная опасность исходит от роботов

Факт
Главную опасность представляет искусственный интеллект с целями, несовместимыми с нашими, потому что ему не нужно никакое тело — только интернет-подключение

Миф
Искусственный интеллект не сможет подчинить себе людей

Факт
Интеллект может подчинять: люди подчиняют себе тигров, будучи умнее

Миф
У машин не может быть целей

Факт
У самонаводящейся ракеты, ориентирующейся по источнику теплового излучения, есть цель

Мифический повод для тревоги
До появления сверхразума осталось всего несколько лет

Реальный повод для тревоги
Остается по меньшей мере несколько десятилетий, но именно столько может понадобиться, чтобы гарантировать его безопасность

Рис. 1.5
Типичные мифы об искусственном интеллекте.

С другой стороны, у нас есть и анти-миф: мы можем не сомневаться в том, что суперинтеллект не появится до конца этого столетия. Исследователи предлагают широкий спектр оценок того, как далеко мы находимся от сверхчеловеческого AGI, но мы никак не можем с уверенностью утверждать, что вероятность получить его к концу века равна нулю, особенно если примем во внимание всю историю удручающе низкой точности предсказаний подобного рода техноскептиков. Вспомним, как Эрнест Резерфорд, по общему признанию величайший физик-ядерщик своего времени, уже в 1933 году – всего лишь за 24 года до открытия Лео Сцилардом ядерных цепных реакций – называл возможность получения ядерной энергии "лунным светом", или как в 1956 году королевский астроном Ричард Вули называл разговоры о полетах в космос "полной мутью". Крайнюю форму этот миф принимает в рассуждениях об искусственном интеллекте, который никогда не сможет стать сверхчеловеческим, потому что это физически невозможно. Но физики знают, что мозг состоит из кварков и электронов, упорядоченных так, что они могут работать как мощный компьютер, и что нет такого физического закона, который мог бы помешать нам создать еще более разумный комок кварков.

Было проведено несколько крупных исследований на экс-

пертных фокус-группах среди специалистов по искусственному интеллекту, где им предлагалось оценить, сколько времени от текущего момента может пройти, пока вероятность создания искусственного интеллекта человеческого уровня достигнет 50 %, и все эти исследования оканчивались одним и тем же: мнения ведущих мировых исследователей по этому поводу расходятся, так что мы просто не знаем. Например, во время такого опроса на нашей конференции в Пуэрто-Рико медианный ответ соответствовал 2055 году, но некоторые предсказывали сотни лет или даже больше.

Еще один имеющий отношение к тому же вопросу миф заключается в том, что люди, переживающие по поводу искусственного интеллекта, ждут его появления уже в ближайшие годы. На самом же деле подавляющее большинство из тех, чье мнение в данном вопросе значимо и кто действительно беспокоится о негативных последствиях создания AI, не ждут его раньше, чем через несколько десятилетий. Но они говорят: коль скоро у нас нет 100 % гарантий, что такое не может случиться уже в этом столетии, стоит начать вести исследования вопросов AI-безопасности уже сейчас и быть готовыми к неожиданностям. Как мы увидим в этой книге, некоторые из вопросов безопасности настолько сложны, что на их решение могут уйти десятилетия, и есть смысл заняться ими сейчас, а не накануне той ночи, когда какие-то попивающие “Рэд Булл” программисты решат запустить AGI человеческого уровня.

Мифы несогласных

Еще одно недоразумение часто возникает по причине распространенного заблуждения, заключающегося в том, что только современные луддиты, не очень-то знакомые с темой, могут выражать какие-то опасения по поводу искусственного интеллекта и призывать к исследованию связанных с ним рисков. Когда Стюарт Рассел сказал об этом во время своего выступления на конференции в Пуэрто-Рико, аудитория откликнулась громким смехом. С этим заблуждением связано еще одно общее недоразумение: что поддержка таких исследований – дело якобы исключительно спорное. В действительности для их проведения в разумных масштабах достаточно скромных инвестиций, и для этого не надо считать риски высокими – надо просто понимать, что ими невозможно пренебречь. Так, исходя из невозможности пренебречь очень невысокой вероятностью, что ваш дом сгорит дотла, вы отчисляете небольшую часть своего дохода на страхование недвижимости.

Мой собственный анализ проблемы привел меня к убеждению, что именно из-за тенденциозного освещения в масс-медиа вопросы AI-безопасности кажутся значительно более спорными, чем на самом деле. В конце концов, страх – востребованный товар, и вырванные из контекста цитаты, если из них можно сделать вывод о неминуемо надвигающейся

катастрофе, соберут больше кликов, чем уравновешенный и детальной отчет о проблеме. Поэтому два человека, знающие о позиции друг друга только по опубликованным цитатам, скорее всего решат, что поводов не согласиться с мнением оппонента у них гораздо больше, чем на самом деле. Например, техноскептик, чьи представления о взглядах Билла Гейтса основаны исключительно на сведениях из британского таблоида, наверняка подумает, что тот полагает появление суперинтеллекта неминуемым, – и конечно же будет неправ. Похожим образом некто, выступающий за создание дружественного AI, прочитав процитированное выше высказывание Эндрю Ына относительно перенаселения Марса, подумает, что того не заботят проблемы AI-безопасности, и тоже ошибется. Я точно знаю, что они его заботят, – но все дело в том, что из-за его особой оценки временных масштабов возникающих проблем он отдает приоритет более близким по времени проблемам.

Мифы о природе рисков

Прочитав в *Daily Mail* заголовок "Стивен Хокинг предостерегает, что восстание роботов может оказаться катастрофическим для человечества", я закрыл глаза[2]. Я уже потерял счет таким статьям. Обычно они сопровождаются картинкой со злобным роботом, волокущим какое-нибудь оружие, и готовят нас к тому, что когда-нибудь роботы обретут

сознание, преисполнятся злобой и поднимут восстание, которого нам следует опасаться. В определенном смысле такие статьи производят на меня сокрушительное впечатление, потому что в сжатой форме предлагают как раз тот самый сценарий, который никак не пугает моих коллег по исследованию AI. Этот сценарий содержит в себе сразу три глубочайших заблуждения, относящихся к трем разным понятиям: наше беспокойство должны вызывать *сознание, злобность* и вообще *роботы*.

Когда вы едете на машине по дороге, ваше восприятие световой и звуковой гамм субъективно. А есть ли субъективное восприятие у беспилотного автомобиля? Чувствует ли он себя настоящим беспилотником или просто катится по дороге, как неразумный зомби, лишенный всякого субъективного восприятия? Хотя эта загадка – что значит быть сознающим – сама по себе интересна и мы посвятим ей 8-ю главу, она не имеет никакого отношения к теме AI-рисков. Если на вас налетит беспилотный автомобиль, вам будет безразлично, осознавал ли он себя в этот момент. Точно так же нас беспокоит не то, что почувствует сверхчеловеческий разум, а что он будет делать.

Страх, что машины окажутся злонамеренными, – еще одна расхожая бессмыслица. Наше опасение вызывают их компетенции, а не злая воля. По определению сверхразумный AI исключительно эффективен в достижении своих целей, каковы бы они ни были, и нам важно, чтобы эти цели не про-

тиворечили нашим. Вряд ли вы относитесь к тем ненавистникам муравьев, кто топчет их по злобе, но если вы руководите проектом по постройке "зеленой" гидроэлектростанции и на предназначенном под затопление участке вдруг случайно окажется муравейник, то муравьям в нем не поздоровится. Движение за дружественный AI ставит перед собой задачу сделать так, чтобы люди никогда не оказывались в положении этих муравьев.

Недоразумение по поводу сознательных машин тесно связано с представлением, будто у машин не может быть целей. У машины, очевидно, могут быть цели в том смысле, что она может проявлять целеустремленное поведение: поведение ракеты, движущейся на источник тепла, наиболее естественно объяснить целью поразить самолет противника. Если вы испытываете беспокойство по поводу того, что цель машины каким-то образом расходится с вашими собственными целями, вам безразлично, до какой степени она себя осознает и какими намерениями движима. Когда вы увидите у себя на хвосте самонаводящуюся ракету, вы не станете успокаивать себя мыслью: "У машины не может быть целей!".

Я с симпатией отношусь и к Родни Бруксу, и к другим пионерам робототехники, которые были возмущены тем, как сеющие ужас таблоиды несправедливо демонизируют их, когда их журналисты, зациклившиеся на роботах, начинают украшать свои статьи злобными металлическими монстрами

с красными светящимися глазами. На самом же деле в центре внимания движения за дружественный AI вовсе не роботы, а сам искусственный интеллект, точнее говоря, – разум с целями, не совместимыми с нашими. Для того чтобы нарушить наш покой, такому не совместимому с нашим разуму вовсе не нужно тело робота, ему достаточно доступа в интернет – в главе 4 мы покажем, как, воспользовавшись этим, он купит и перепродаст всех на финансовых рынках, затмит изобретательностью любых изобретателей, покорит своей демагогией больше обывателей, чем любой человеческий политический лидер, и придумает оружие, принципов действия которого мы даже не будем понимать. Даже если бы создание роботов было физически невозможно, сверхразумный и сверхбогатый искусственный интеллект легко бы подкупал мириады человеческих существ и манипулировал бы ими, неумышленно вовлекая их в свои изощренные торговые операции, как это происходит в фантастическом романе Уильяма Гибсона *Neuromancer*.

Недоразумение с роботами напрямую связано с мифом, будто машины не могут управлять людьми. Разум – путь к управлению: человек может командовать тигром не потому, что сильнее, а потому, что умнее. Если мы уступим свое положение самых умных на планете, мы можем потерять и контроль над собой.

На рис. 1.5 все эти общие недоразумения собраны воедино, так чтобы мы могли покончить с ними раз и навсегда

и сосредоточить наши дискуссии с друзьями и коллегами вокруг настоящих противоречий, в которых, как мы сейчас убедимся, нет недостатка.

Дорога вперед

Вся оставшаяся часть этой книги посвящена выяснению вопроса, на что может быть похожа будущая жизнь с искусственным интеллектом, и мы займемся этим вместе. Давайте двинемся по этому пути, следуя хорошо выстроенному плану, а для этого сначала постараемся проанализировать всю историю жизни концептуально и хронологически, а затем обратимся к целям и средствам, а также к тому, какие нам следует предпринять действия, чтобы создать такое будущее, какое мы хотим.

В главе 2 мы исследуем вопрос об основаниях разума и о том, как пассивная и бессмысленная на вид материя может быть реорганизована и, благодаря этому, может обрести способность запоминать, вычислять и учиться. Когда мы перейдем к будущему, наш рассказ разветвится, и каждый из множества возможных сценариев будет зависеть от ответов, данных на ключевые вопросы. На рис. 1.6 эти ключевые вопросы собраны вместе, в том порядке, в каком мы будем с ними сталкиваться по мере совершенствования AI.

Рис. 1.6

Какие именно вопросы относительно искусственного интеллекта представляют интерес, зависит от того, насколько он развит, и от того, по какому направлению стало развиваться наше будущее.

		Краткий заголовок главы	Тема	Статус
		Пролог: Сказание о команде “Омега”	Пища к размышлению	Исключительно спекулятивно
	1	Самый важный разговор	Ключевые идеи, терминология	
	2	Материя начинает думать	Основы теории разума	Не очень спекулятивно
История разума	3	AI, экономика, оружие и закон	Ближайшее будущее	
	4	Взрыв разума?	Сценарии сверхразума	
	5	Что потом?	Мир в следующие 10 тысяч лет	Исключительно спекулятивно
	6	Наше космическое благосостояние	Мир в следующие миллиарды лет	
	7	Цели	История целенаправленного поведения	Не очень спекулятивно
История смысла	8	Сознание	О естественном и искусственном сознании	Спекулятивно
		Эпилог: Сказание о команде FLI	Что мы должны делать?	Не очень спекулятивно

Рис. 1.7.
Структура книги

Уже сейчас перед нами стоит вопрос о начале своего рода гонки вооружений, использующих AI-технологии, а также целый ряд вопросов о том, как сделать завтрашний AI надежным и работающим без “глюков”. Если позитивное влияние AI-технологий на экономику будет расти, нам придется решать, как преобразовывать законодательную систему и на какую карьеру ориентировать наших детей, чтобы они не оказывались вовлеченными в ту профессиональную деятельность, которой грозит скорая автоматизация. Мы рассмотрим все эти насущные уже в краткосрочной перспективе вопросы в главе 3.

Если AI в своем развитии достигнет человеческого уровня, нам придется спросить себя: а как сделать его друже-

ственным? можем ли мы обеспечить себе с его помощью праздную жизнь? хотим ли мы этого? Отсюда также возникает вопрос о возможности AI за счет взрывного развития или постепенного, но неуклонного роста достичь уровня, значительно превосходящего человеческий. Мы рассмотрим широкий набор различных сценариев в главе 4 и целый спектр возможных последствий в главе 5 – от похожих на утопии до похожих на антиутопии. Кто стоит во главе – человек, AI или киборг? Хорошо ли с людьми обращаются? Если на смену людям приходят какие-то иные сущности, должны ли мы рассматривать пришедших как захватчиков или как потомков, достойных своих предков? Мне очень интересно, какой из предложенных в главе 5 сценариев кажется наиболее предпочтительным лично вам! Я создал вебсайт http://AgeOfAi.org, для того чтобы вы могли поделиться своими мыслями и присоединиться к разговору.

Наконец, мы попытаемся унестись на миллиарды лет вперед в главе 6, где, по иронии законов нашего познания, мы можем сделать гораздо более точные предсказания, чем в предшествующих главах, потому что финальные границы для жизни во Вселенной установлены не разумом, а законами физики.

Завершив исследование истории разума, мы посвятим последние разделы нашей книги рассуждениям о том, к какому будущему мы должны стремиться и как его достичь. Для того чтобы связать друг с другом бесстрастные факты и вопро-

сы намерений и средств, мы исследуем в главе 7 физический фундамент целеполагания, а в главе 8 – сознание. Наконец, в эпилоге мы спросим себя: а что можно сделать уже сейчас, чтобы попасть в то будущее, которого мы хотим?

На случай, если вы вдруг окажетесь из тех читателей, которые любят перескакивать с одного на другое, все главы сделаны более или менее самодостаточными, при условии, что вы уже переварили терминологию и определения этой главы и начала следующей. Если вы специалист в области искусственного интеллекта, можете пропустить почти всю главу 2, кроме определений разума, данных в самом ее начале. Если тема AI для вас новая, то главы 2 и 3 объяснят вам, почему вы не должны отмахиваться от глав 4 и 6 как от немыслимой научной фантастики. На рис. 1.7 дана схема соотношения фактов и спекуляций в различных главах.

Вас ждет увлекательное путешествие. А теперь в путь!

Подведение итогов

• Жизнь, определяемая как процесс, который обладает способностью к самовоспроизводству при сохранении сложности, может проходить в своем развитии через три этапа: биологический (Жизнь 1.0), где «хард» живых организмов и их «софт» развиваются эволюционным путем, культурный (Жизнь 2.0), где «софт» может проектироваться (благодаря обучению), и технологический (Жизнь 3.0), где проектироваться может и «хард», и «софт», в результате чего жизнь получает власть над своей судьбой.

• Искусственный интеллект может позволить нам сделать Жизнь 3.0 реальностью уже в этом веке, а значит, нам пора начинать всерьез задумываться о том, к какому будущему мы должны стремиться и каким образом оно может быть достигнуто. В разворачивающейся по этому поводу полемике есть три основных лагеря: техноскептики, цифро-утописты и участники движения за дружественный AI.

• С позиций техноскептиков задача создания сверхчеловеческого универсального AI настолько сложна, что, если и поддается решению, то на это потребуется не одна сотня лет, и сейчас глупо беспокоиться по этому поводу (равно как и о Жизни 3.0).

• Цифро-утописты полагают появление его уже в этом веке вполне вероятным и искренне приветствуют переход

к Жизни 3.0, рассматривая ее как естественный и желанный шаг в космической эволюции.

• Движение за дружественный AI также полагает появление сверхразума в этом веке вероятным, но его сторонники не считают гарантированными плюсы такого сценария – они должны быть обеспечены результатами напряженной исследовательской работы в области AI-безопасности.

• Помимо этих законных разногласий между ведущими мировыми экспертами, есть также досадные псевдо-противоречия, вызванные непониманием сути проблемы. Например, бессмысленно тратить время на споры о «жизни», «разуме» или «сознании», если нет уверенности, что стороны в споре одинаково трактуют соответствующие понятия! Определения, используемые в этой книге, сведены в таблицу 1.1.

• Следует отдавать себе отчет в существовании распространенных заблуждений, поясняемых рис. 1.5: *сверхразум к 2100 году неизбежен / невозможен. Искусственный интеллект беспокоит только луддитов. Главная опасность в том, что AI может стать злонамерен и/или действовать осознанно, и от этой опасности нас отделяют всего несколько лет. Прежде всего, надо обезопаситься от роботов. Искусственный интеллект не может контролировать людей и не может ставить перед собой целей.*

• В главах 2–6 мы рассмотрим историю разума с непритязательного ее начала миллиарды лет назад к возможному космическому будущему миллиарды лет спустя. Сначала мы

рассмотрим проблемы, возникающие в ближайшей перспективе: нехватку рабочих мест, автономные системы оружия, создание универсального интеллекта человеческого уровня. Затем мы исследуем разнообразные возможности совместного существования людей и машин – очень интересно, какой из вариантов предпочли бы вы!

• В главах 7, 8 и эпилоге мы перейдем от бесстрастных описаний к исследованию целей, сознания и смысла и попытаемся выяснить, что в наших силах сделать прямо сейчас ради достижения того будущего, какого бы нам хотелось.

• На мой взгляд, этот разговор о будущем жизни с искусственным интеллектом – самый важный для нашего времени. Пожалуйста, присоединяйтесь к нему!

Глава 2
Материя начинает думать

Водород … по прошествии некоторого времени … превращается в людей.
Эдвард Роберт Харрисон, 1995

Одно из самых примечательных превращений, испытанных немой и бессмысленной материей за 13,8 миллиарда лет после Большого взрыва, – это обретение ею разума. Как могло это произойти и до каких пределов может продолжаться? Что может сказать наука об истории и о будущем разума во Вселенной? Чтобы облегчить понимание, давайте посвятим эту главу исследованию фундамента, на котором он возникает, и кирпичиков, из которых он построен. Что именно мы имеем в виду, утверждая, что некий сгусток материи разумен? Что мы подразумеваем, когда говорим о способности некоего объекта помнить, вычислять и обучаться?

Что такое разум?

Недавно нам с женой посчастливилось принять участие в симпозиуме, организованном Фондом Нобеля и посвященном искусственному интеллекту. Когда собравшихся специалистов попросили дать определение интеллекту вообще, между ними завязался длинный спор, который так и не привел их к согласию. Нам показалось довольно забавным, что даже среди самых разумных исследователей разума нет согласия относительно объекта их исследования – самого разума. Так что никакого безусловно "правильного" определения интеллекта просто не существует. Вместо этого есть довольно длинный список конкурирующих определений, использующих такие понятия, как логика, понимание, планирование, эмоциональное познание, самосознание, творчество, способность к решению задач и обучению.

В нашем исследовании, предполагающем возможность будущих трансформаций интеллекта, определение должно быть максимально широким и допускающим дальнейшие расширения, оно не должно ограничиваться теми его формами, которые уже существуют. Вот почему определение, данное мной в предыдущей главе и которым я собираюсь пользоваться в этой книге, очень широкое:

> интеллект – это способность достигать сложных целей

Это определение достаточно широко, чтобы включить все перечисленные выше: понимание, самосознание, решение задач и обучение – все это примеры тех сложных целей, которые могут возникнуть. Оно также достаточно широко, чтобы включить определение, данное *Оксфордским словарем*, – "способность приобретать и использовать знания и навыки": ведь "приобретать и использовать знания и навыки" тоже может быть целью.

Так как цели могут быть самыми разными, возможны разные типы интеллекта. В соответствии с нашим определением, следовательно, нет смысла описывать интеллект человека, или какого-то другого живого существа, или интеллект машины с помощью единого численного показателя вроде IQ[10]. Какая компьютерная программа "умнее" – та, которая умеет играть только в шахматы, или та, которая умеет играть только в го? Тут нельзя дать никакого осмысленного ответа: каждая из них пригодна для своего, и напрямую их сравнивать невозможно. Но мы, однако, можем сказать, что третья программа "умнее" каждой из этих двух, если она по крайней мере так же хороша, как и они, в решении *любой* задачи и безусловно лучше их в решении хотя бы какой-то одной (например, обыгрывает их в шахматы).

Нет большого смысла в спорах о том, чей интеллект силь-

[10] Чтобы лучше понять, о чем здесь речь, давайте представим, что вводится какой-то единый "атлетический коэффициент", сокращенно AQ, для всех спортсменов олимпийского уровня, так, словно обладатель самого высокого AQ должен побеждать на Олимпиаде во всех видах спорта.

нее, в пограничных случаях, так как интеллект проявляется на спектре задач и совершенно не обязательно определяется по принципу “все или ничего”. Какие люди обладают способностью достигать целей говорения? Новорожденные? Нет. А радиоприемники – да. А что вы скажете о младенце, знающем десять слов? Пятьсот слов? Где же провести линию? Я умышленно использовал туманное слово “сложные”, потому что выяснять, где надо провести линию между разумным и неразумным, не очень интересно, гораздо полезнее было бы просто научиться измерять степень способности достигать цели для различных типов задач.

При создании такой таксономии будет полезно ввести еще один важный признак, разделив узко-ориентированный (слабый) интеллект и широко-ориентированный (сильный). Созданный IBM шахматный компьютер Deep Blue, подвинувший с шахматного трона чемпиона мира Гарри Каспарова в 1997 году, был способен к достижению целей в очень узком классе задач – в игре в шахматы. Несмотря на исключительно впечатляющие “хард” и “софт”, в крестики-нолики он не смог бы обыграть и четырехлетнего ребенка. Искусственный интеллект разработанной в компании DeepMind сети глубокого Q-обучения (DQN) может успешно достигать целей несколько более разнообразных, играя в несколько десятков игр прошлого века компании Atari на уровне человека или лучше. Ничто пока не может сравниться с человеческим интеллектом прежде всего по его уникальной широте:

он способен освоить головокружительное разнообразие умений. Здоровый ребенок при условии достаточно продолжительных тренировок может научиться очень хорошо играть в *любую* игру, выучить *любой* язык, достичь очень хороших показателей в *любом* виде спорта и *любой* профессии. Если сравнивать интеллект машины и человека сегодня, то по широте человек побеждает одной левой. Но в некоторых узких областях, количество которых неуклонно растет, превосходство машин над нами не вызывает сомнений. Эти области приведены на рис. 2.1. Философским камнем AI-исследований остается построение универсального, или сильного, искусственного интеллекта (AGI), его широта максимальна: он способен в принципе достичь любой цели, включая обучение. Мы подробно обсудим его в главе 4. Термин AGI широко употреблялся Шейном Леггом, Марком Губрудом и Беном Гёрцлем в несколько ограниченном значении – как универсальный интеллект *человеческого уровня*: то есть он должен не только справляться с любой задачей, доступной человеку, но и делать это не хуже нас[3]. Я приму это ограничение и, если только не использую какой-то уточняющий эпитет (например, "сверхчеловеческий AGI"), говоря о "AGI" или о "сильном AI", буду иметь в виду "AGI человеческого уровня"[11].

[11] Некоторые предпочитают использовать термины "сильный AI" или "AI человеческого уровня" как синонимы AGI, что приводит к некоторым проблемам. В узком смысле слова даже карманный калькулятор – это "AI человеческого уровня". Антонимом "сильному AI" должен служить "слабый AI", но довольно

Хотя слово “интеллект” кажется позитивно окрашенным, важно понимать, что везде, где мы его используем, мы делаем это в абсолютно нейтральном значении: речь лишь о способности достигать цели, независимо от того, рассматриваем мы эти цели как благо или как зло. Так же и с людьми: один умный человек очень хорош в деле помощи людям, другой – в принесении им бед. Мы исследуем тему целей в главе 7. Говоря об этом, нам придется заняться и весьма деликатным вопросом относительно того, чьи именно цели мы обсуждаем. Предположим, у вашего будущего роботизированного персонального помощника совсем нет никаких собственных целей, но он сделает все, о чем бы вы его ни попросили. А вы его просите приготовить идеальный итальянский ужин. Он отправляется в интернет, изучает там рецепты итальянских блюд, ищет ближайший супермаркет, готовит вам пасту, разбирается с прочими ингредиентами и в конце концов готовит вам отличный ужин. Вы, полагаю, сочтете его очень умным, хотя изначально цель была вашей. В действительности он воспринял вашу цель, как только она была поставлена, и встроил ее в систему своих собственных вспомогательных целей, от оплаты счетов до измельчения пармезана. В этом смысле разумное поведение неизменно привязано к целеустремленности.

странно называть слабым AI Deep Blue, Watson или AlphaGo.

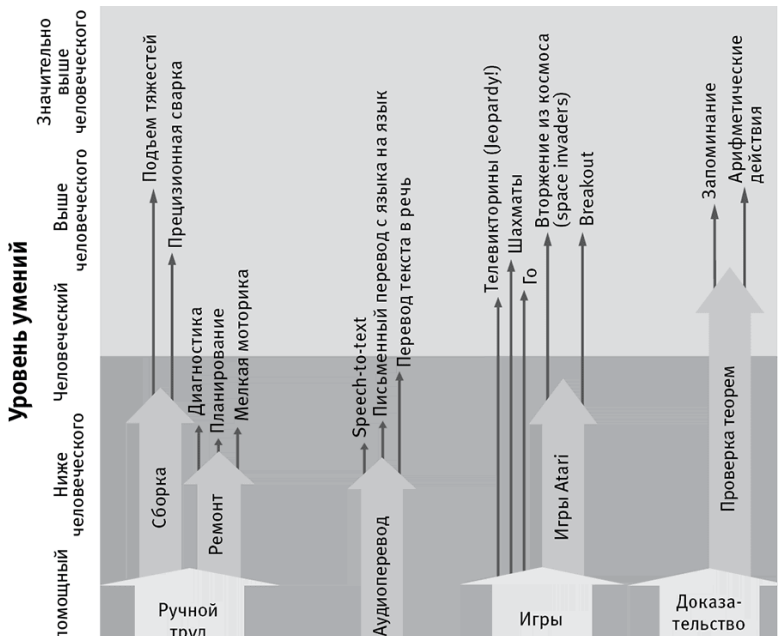


Рис. 2.1

Интеллект, определенный как способность к достижению сложных целей, не может быть измерен единственным показателем – IQ, а только целым их спектром, охватывающим все возможные цели. Каждая стрелка на рисунке показывает, насколько успешны лучшие из современных AI-систем в достижении различных целей, откуда ясно, что сегодня системы с искусственным интеллектом обнаруживают определен-

ную узость: каждая система способна достигать только очень специфических целей. В отличие от этого, человеческий разум чрезвычайно широк: здоровый ребенок может успешно учиться практически всему на свете.

Совершенно естественно мы, люди, ранжируем сложность задач в соответствии с тем, насколько сложны они для нас самих, как показано на рис. 2.1. Но такой подход приводит к ложной картине сложности задач для компьютера. Нам кажется, что умножить 314 159 на 271 828 гораздо сложнее, чем опознать друга на фотографии, но компьютеры обошли нас в арифметике задолго до того, как я родился, а опознание людей по картинкам на человеческом уровне освоили совсем недавно. Тот факт, что элементарные сенсомоторные задания кажутся нам простыми, хотя требуют колоссальных вычислительных ресурсов, известен как парадокс Моравеца, который объясняется тем, что наш мозг легко отдает под их решение значительную часть хорошо приспособленного к этому нашего "харда", головного мозга – как выясняется, больше четверти.

Мне нравится эта метафора у Ганса Моравеца, и я позволю себе ее небольшую вольную иллюстрацию (см. рис. 2.2)[4]:

> "Компьютеры – универсальные машины, и их потенциал равномерно покрывает безграничное разнообразие задач. Потенции людей, напротив, сосредоточены там, где от успеха зависит

выживание, в более отдаленных областях они весьма слабы. Представьте себе “ландшафт человеческих компетенций”, где есть низины вроде “арифметики” и “механической памяти”, холмики вроде “шахмат” или “доказательства теорем” и горные пики, отмеченные указателями “перемещение с места на место”, “координация движений рук и глаза”, “социальное взаимодействие”. С совершенствованием компьютеров этот ландшафт словно наполняется водой: полвека назад она затопила низины, вымыв оттуда счетоводов и писцов, но оставив нас сухими. Сейчас вода дошла до холмиков, и обитатели наших аванпостов забеспокоились: куда бы им переместиться? Мы чувствуем себя в безопасности на своих пиках, но, учитывая скорость, с которой вода прибывает, она покроет и пики в ближайшие полвека. Я полагаю, нам уже пора начинать строить ковчеги и приучаться к жизни на плаву”.

Рис. 2.2

Диаграмма Моравеца “Ландшафт человеческих умений”, на которой рельеф представляет эти умения в зависимости от их сложности для компьютеров, а повышающийся уровень воды – то, чему компьютеры уже научились.

За десятилетия, прошедшие со времени написания этих строк, уровень воды неуклонно повышался в соответствии с предсказанием, и некоторые из холмиков (вроде шахмат) уже давно скрылись из виду. Что на очереди и как нам быть в связи с этим – вот в чем суть нашей книги.

По мере того как уровень воды повышается, в какой-то

момент он может достичь критической отметки, за которой начнутся драматические перемены. Эта критическая отметка соответствует способности машин заниматься дизайном AI. До того как она достигнута, повышение уровня воды определяется деятельностью *людей* по улучшению компьютеров, но дальше *машины* начинают улучшать машины, по всей вероятности делая это намного успешнее, чем люди, и площадь суши, возвышающейся над водой, станет сокращаться намного быстрее. В этом и заключается спорная, но головокружительная идея *сингулярности*, о который мы будем иметь удовольствие рассуждать в главе 4.

Создатель теории компьютеров Алан Тьюринг, как известно, сумел доказать, что если какой-то компьютер может осуществлять определенный минимум операций, то, при наличии у него достаточного времени и памяти, он может быть запрограммирован на выполнение чего угодно, при условии, что это может быть выполнено хоть каким-то компьютером. Машины, возвышающиеся над этим критическим порогом, называют *универсальными компьютерами* (по Тьюрингу); любой смартфон или ноутбук универсален в этом смысле. Аналогично, я склонен считать пороговый уровень интеллекта, требуемый для дизайна искусственного интеллекта, *универсальным интеллектом:* обладая достаточным временем и ресурсами, он может достичь любой цели, которая в принципе может быть достигнута *какой бы то ни было* разумной сущностью. Например, если она решает, что ей

необходимо улучшить свои социальные навыки, прогностические способности или усовершенствоваться в AI-дизайне, то она достигает этого. Если она приходит к решению о необходимости построить фабрику роботов, то она строит эту фабрику. Иными словами, универсальный интеллект обладает потенциалом превращения в Жизнь 3.0.

Среди специалистов по искусственному интеллекту принято считать, что интеллект в конечном счете определяется информацией и вычислениями, а не плотью, кровью или атомами углерода. То есть нет никаких фундаментальных причин, по которым машины однажды не смогут стать хотя бы такими же умными, как мы.

Но что представляют собой информация и вычисления в реальности, если вспомнить, что фундаментальная физика учит нас: в мире нет ничего, кроме энергии и движущейся материи? Как может нечто настолько абстрактное, неосязаемое и эфемерное, как информация и вычисления, воплотиться в грубую физическую ткань? В особенности – как могут какие-то бессмысленные элементарные частицы, вращающиеся друг вокруг друга по законам физики, продемонстрировать поведение, которое мы называем разумным?

Если ответ на этот вопрос кажется вам очевидным, а появление машин не менее разумных, чем люди, уже в этом веке – правдоподобным (например, если вы сами – специалист по искусственному интеллекту), то, пожалуйста, не дочитывайте до конца эту главу, а переходите сразу к главе 3. А если

нет, то вам должно быть приятно узнать, что следующие три раздела я написал специально для вас!

Что такое память?

Когда мы говорим, что в атласе содержится *информация* о мире, мы имеем в виду, что между состоянием книги (в частности, между расположением некоторых молекул, придающих определенный цвет буквам и рисункам) и состоянием всего мира (в частности, расположением континентов) есть определенное отношение. Если бы эти континенты находились в других местах, то и молекулы краски должны были бы располагаться иначе. Нам, людям, дано пользоваться бесчисленными устройствами для хранения информации, от книг и мозгов до твердотельных накопителей, и все они обладают одним и тем же свойством: их состояние находится в некотором отношении к состоянию каких-то других вещей, важных для нас, – именно поэтому первые могут нас информировать о вторых.

Что же представляет собой фундаментальное свойство всех этих предметов, которое позволяет нам использовать их в качестве памяти, то есть накопителей информации? Ответ заключается в том, что *каждому из них доступно большое количество устойчивых состояний, в которых они могут находиться очень долгое время* – достаточно долгое, чтобы извлечь закодированную информацию, как она только потребуется. В качестве простого примера представьте себе холмистую местность с шестнадцатью отделенными од-

на от другой ложбинами и небольшой мячик, как показано на рис. 2.3. Когда мячик скатывается с холма в какую-то из ложбин, это будет одна из тех шестнадцати, и коль скоро он может там находиться долго, его нахождение там можно использовать для запоминания одного из шестнадцати чисел (от 1 до 16).

Это запоминающее устройство довольно надежно, так как даже если его будут сотрясать какие-то внешние силы, мячик, вероятно, останется в той ложбине, куда вы его поместили изначально, и вы всегда сможете сказать, какое из чисел было там сохранено. Причина стабильности такой памяти заключается в том, что для извлечения мячика из заключающей его ложбины требуется больше энергии, чем сообщаемая ему случайными сотрясениями. У той же идеи могут быть и более общие реализации, чем просто катающийся мячик: энергия сложной физической системы может определяться целым рядом ее механических, химических, электрических и магнитных свойств; и до тех пор, пока энергия воздействия на систему недостаточна для изменения ее состояния, которое она должна запомнить, состояние будет устойчивым. Этим объясняется, почему у твердых тел много устойчивых состояний, а у жидких и газообразных – нет: если вы выгравируете чье-то имя на золотом кольце, то и по прошествии многих лет оно будет там, так как для изменения формы золота требуется значительная энергия, но если вы выгравируете его же на поверхности пруда, информация

пропадет за секунду, потому что поверхность воды изменяется практически без энергетических затрат.

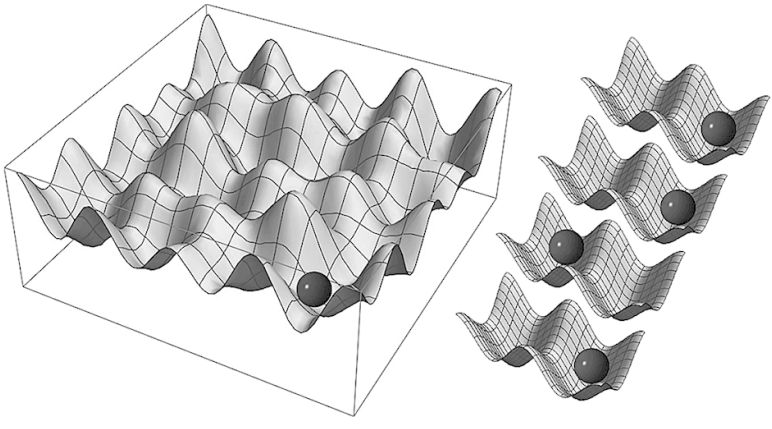


Рис. 2.3

Роль запоминающего устройства хорошо выполняют те физические объекты, у которых много стабильных устойчивых состояний. Шарик слева может закодировать четыре бита информации, соответствующих одной из шестнадцати ($2^4 = 16$) впадин рельефа. Также четыре бита могут хранить вместе четыре шарика справа – по одному биту на каждого.

У простейшего запоминающего устройства всего лишь два устойчивых состояния (см. рис. 2.3) Поэтому мы можем считать, что оно запоминает один бинарный знак (сокращенно “бит”) – например ноль или единицу. Информация, со-

храненная более сложным устройством, может быть представлена словно бы сохраненной во множестве бит: например, четыре бита, взятые вместе, как показано на рис. 2.3 (справа), могут находиться в одном из $2 \times 2 \times 2 \times 2 = 16$ различных состояний – 0000, 0001, 0010, …, 1111, так что у них всех вместе тот же самый объем памяти, что и у системы с 16 различными состояниями (слева). Поэтому мы можем думать о битах как об атомах информации, мельчайших ее частичках, которые не могут быть разделены дальше, но которые могут объединяться, представляя любое ее количество. Например, я только что напечатал слово "слово", и мой ноутбук тут же превратил его в своей памяти в последовательность из пяти трехзначных чисел: 241 235 238 226 238, представив каждое из них в виде 8 бит (каждой букве нижнего регистра присваивается число 223 плюс его порядковый номер в алфавите). Как только я нажимаю на клавишу "с" своего ноутбука, эта буква тут же появляется на мониторе, и ее изображение тоже состоит из бит, причем 32 бита определяют цвет каждого из миллиона пикселей монитора.

Поскольку двухуровневые системы легче и в производстве, и в управлении, большинство современных компьютеров хранят информацию в битах, хотя существует обширнейшее многообразие в способах физического воплощения каждого из них. На DVD каждому биту соответствует наличие или отсутствие микроскопической ямки в определенном месте его пластиковой поверхности. На жестком дис-

ке биту соответствует одна из двух возможных поляризаций магнитного момента в данной точке. В оперативной памяти моего ноутбука биту соответствуют определенные конфигурации некоторых электронов, от которых зависит, заряжено или нет устройство под названием микроконденсатор. Некоторые биты очень хорошо подходят для того, чтобы пересылать их с места на место, иногда даже со скоростью света: например, в оптоволокне при передаче вашего электронного сообщения биту соответствует ослабление или усиление лазерного луча в определенный момент.

Инженеры предпочитают кодировать биты в системах, обеспечивающих не только устойчивость и простоту считывания (как на золотом кольце), но и простоту записи: изменение состояния вашего жесткого диска требует значительно меньших затрат энергии, чем гравирование по золоту. Они также предпочитают системы, с которыми легко работать и которые достаточно дешевы при массовом производстве. Но помимо этого их совсем не интересует, каким именно физическим объектом бит был представлен – как, впрочем, в большинстве случае и вас, потому что это и вообще неважно! Если вы пересылаете электронной почтой документ своему другу, чтобы он вывел его на печать, то информация последовательно быстро копируется с магнитных диполей жесткого диска в электрические заряды оперативной памяти, оттуда в радиоволны вашей Wi-Fi-сети, потом в переменное напряжение в цепях вашего роутера, лазерные

импульсы в оптоволокне и, наконец, передается молекулам на поверхности бумаги. Иными словами, *информация живет собственную жизнь, независимо от своего физического субстрата!* В самом деле, нас-то обычно интересует только этот, не зависящий от субстрата, аспект информации: если ваш друг позвонит спросить, что это за документы вы ему послали, он, скорее всего, не будет интересоваться перепадами напряжения и смещениями молекул. А для нас это первый звоночек: как такая неосязаемая вещь, как разум, может оказаться воплощенной в сугубо осязаемой физической материи, а скоро мы увидим, что идея независимости от субстрата гораздо глубже, включая в себя кроме информации также вычисления и обучение.

Из-за этой самой независимости от субстрата изобретательные инженеры то и дело заменяют запоминающие устройства в наших компьютерах все более совершенными, основанными на новых технологиях, но это совсем не заставляет нас менять что-либо в программном обеспечении компьютеров, их "софте". Как видно на рис. 2.4, результаты потрясающие: на протяжении последних шести десятилетий примерно каждые два года компьютерная память становится вдвое дешевле. Жесткие диски стали дешевле более чем в 100 миллионов раз, а разновидности памяти с быстрым доступом, применяемые не столько просто для хранения, сколько для выполнения вычислений, стали сейчас дешевле аж в 10 триллионов раз! Если бы вам удавалось полу-

чить такую скидку в 99,99999999999 % на каждую свою покупку, то вы смогли бы купить всю недвижимость Нью-Йорка менее чем за 10 центов, а все золото, когда-либо добытое на Земле, чуть более чем за доллар.

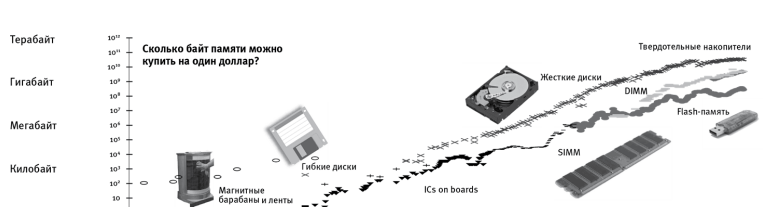


Рис. 2.4

На протяжении шести последних десятилетий компьютерная память дешевела вдвое примерно каждые два года, чему соответствует снижение цены в тысячу раз на каждые двадцать лет. Один байт равен восьми битам. Данные предоставлены Джоном Мак-Каллемом (см. http://www.jcmit.net/memoryprice.htm, проверено 13.05.2018)

У каждого из нас есть свои личные воспоминания, так или иначе связанные с этим впечатляющим улучшением в технологиях хранения информации. Я хорошо помню, как, учась в старшей школе, подрабатывал в кондитерской, чтобы накопить на компьютер со всего лишь 16 килобайтами памяти. Когда мы с моим одноклассником Магнусом Бодином написали и успешно продали текстовый редактор для этого ком-

пьютера, нам пришлось уложить его в ультракороткий программный код, чтобы оставалось хоть какое-то место для самого текста, который можно было бы редактировать. Привыкнув к гибкой дискете на 70 килобайт, я был потрясен появлением 3,5-дюймовой дискеты меньшего размера, на которой умещалось целых 1,44 мегабайта – на нее влезала целая книга, а потом и моим первым жестким диском на 100 мегабайт, которых сегодня едва хватило бы на загрузку одной песни. Кажется совершенно невозможным совместить эти юношеские воспоминания с другими, более поздними: много лет спустя я покупал за 100 долларов жесткий диск в 300 000 раз большей вместительности.

Было ли что-нибудь в этих запоминающих устройствах такое, что эволюционировало, а не конструировалось бы людьми? Биологи до сих пор не знают, какого рода отпечатки производили первые формы жизни, чтобы передавать их от поколения к поколению, но скорее всего они были очень невелики. Исследовательская группа Филиппа Холлигера из Кембриджского университета сумела синтезировать молекулу РНК, кодирующей 412 бит генетической информации, которая была в состоянии создавать нити РНК длиннее себя самой; это открытие поддерживало гипотезу “мира РНК”, состоящую в том, что ранняя земная жизнь – это были короткие самовоспроизводящиеся РНК-цепочки. Известное к настоящему времени запоминающее устройство с минимальной памятью, возникшее в результате эволюции в дикой при-

роде, – это геном бактерии *Candidatus Carsonella ruddii,* сохраняющий до 40 килобайт информации, в то время как наш человеческий геном хранит около 1,6 гигабайт, что примерно соответствует одному загружаемому из интернета кинофильму. Как уже говорилось в предыдущей главе, наш мозг сохраняет гораздо больше информации, чем наш геном: на уровне примерно 10 гигабайт электрически (что определяется тем, какие из 100 миллиардов нейронов "светятся" в тот или иной момент времени) или 100 терабайт биохимически (что определяется тем, насколько сильно различные нейроны сцеплены в синапсах). Сравнение этих чисел с памятью машин показывает, что лучшие компьютеры мира сейчас превосходят по способности хранить информацию любые биологические системы при быстро падающей стоимости, которая на 2016 год составляла всего несколько тысяч долларов.

Память вашего мозга работает совсем не так, как память компьютера, не только в отношении того, как она устроена, но и в отношении того, как она используется. Вы получаете информацию из компьютера или с жесткого диска, указывая, *где* она хранится, а информацию в мозгу вы получаете, указав, *что* примерно вам нужно. Каждая группа бит в памяти вашего компьютера характеризуется своим численным адресом, и чтобы получить доступ к той или иной информации, вам надо указать компьютеру адрес, по которому искать, как если бы я сказал вам: "Пойдите к моему книжно-

му шкафу, возьмите там пятую книгу справа на верхней полке и прочитайте, что написано на странице 314". Напротив, у себя в мозгу вы находите ее примерно так же, как с помощью поисковой машины: вы говорите, что ищете, или называете что-то, имеющее некоторое отношение к тому, что вы ищете, и оно всплывает на поверхность. Если я скажу вам: "Быть или не быть" – или если я забью эти слова в поисковую строку Google, результатом в обоих случаях скорее всего будет: "Вот в чем вопрос". Причем результат будет достигнут, даже если я спрошу о другой части той же цитаты или перепутаю в ней слова. Такая память называется *автоассоциативной*, потому что поиск информации в ней происходит по ассоциации, а не по адресу.

В знаменитой статье 1982 года физик Джон Хопфилд показал, как сеть взаимосвязанных нейронов может превратиться в автоассоциативную память. Мне очень нравится его идея, я нахожу ее красивой и пригодной для описания любой физической системы с многочисленными устойчивыми состояниями. Например, представьте себе шарик на поверхности с двумя лунками – вроде того, как это устроено в однобитной системе на рис. 2.3, и пусть форма поверхности такова, что x-координаты минимумов потенциальной энергии, где шарик приходит в состоянии покоя, соответственно равны $x = \sqrt{2} \approx 1{,}41421$ и $x = \pi \approx 3{,}14159$. Если вы помните только, что "пи" – это где-то около 3, то, поместив шарик в точку x=3, вы увидите, как он сам скатится в точку ми-

нимума энергии, где координата x окажется точно равной “пи”. Хопфилд понял, что сложно устроенная сеть нейронов создает подобный же ландшафт с многочисленными энергетическими минимумами, в которые может прийти система, а со временем было доказано, что в каждую тысячу нейронов можно втиснуть 138 различных воспоминаний без особой путаницы между ними.

Что такое вычисление?

Итак, мы видели, как физический объект может хранить информацию. Но как он может вычислять?

Вычисление – это переход памяти из одного состояния в другое. Иными словами, вычисление использует информацию, чтобы преобразовывать ее, применяя к ней то, что математики называют *функцией.* Я представляю себе функцию этакой мясорубкой для информации, как показано на рис. 2.5: вы закладываете в нее сверху исходную информацию, поворачиваете ручку, и оттуда вылезает переработанная информация. Вы можете повторять раз за разом одно и то же действие, получая при этом все время что-то разное. Но сама по себе обработка информации полностью детерминирована в том смысле, что если у вас на входе все время одно и то же, то и на выходе вы будете получать все время один и тот же результат.

В этом и заключается идея функции, и хотя такое определение кажется слишком простым, оно до невероятия хорошо работает. Некоторые функции совсем тривиальные, вроде той, что зовется NOT: у нее на входе один бит, и она заменяет его другим, превращая ноль в единицу, а единицу в ноль. Функции, которые мы изучаем в школе, обычно соответствуют кнопочкам на карманном калькуляторе, на входе при этом может быть одно число или несколько, но на

выходе всегда одно: например, это может быть x^2, то есть при вводе числа выводится результат его умножения на себя. Но есть и исключительно сложные функции. Например, если вы располагаете функцией, у которой на входе произвольное положение фигур на шахматной доске, а на выходе – наилучший следующий ход, то у вас есть шанс на победу в компьютерном чемпионате мира по шахматам. Если вы располагаете функцией, у которой на входе состояние всех финансовых рынков мира, а на выходе – список акций, которые следует покупать, то вы скоро сильно разбогатеете. Многие специалисты по искусственному интеллекту видят свою задачу исключительно в том, чтобы придумать, как вычислять некоторые функции для любых начальных условий. Например, цель машинного перевода заключается в том, чтобы, взяв последовательность бит, представляющую исходный текст на одном языке, преобразовать ее в другую последовательность бит, представляющую тот же текст, но на другом языке, а цель создания систем автоматизированного распознавания изображений заключается в том, чтобы преобразовывать последовательность бит, представляющую какую-то картинку на входе, в последовательность бит, представляющую собой текст, который эту картинку описывает (рис. 2.5).

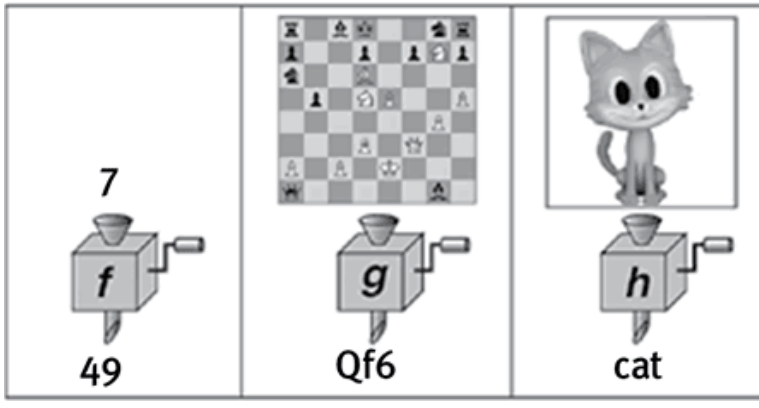


Рис. 2.5

Каждое вычисление использует информацию на входе, чтобы преобразовывать ее, выполняя над ней то, что математики называют функцией. У функции *f* (слева) на входе последовательность бит, представляющих число; в результате вычислений она дает на выходе его квадрат. У функции *g* (в центре) на входе последовательность бит, представляющих позицию на шахматной доске; в результате вычислений она дает на выходе лучший ход для белых. У функции *h* (справа) на входе последовательность бит, представляющих изображение, в результате вычислений она дает на выходе соответствующую текстовую подпись.

Другими словами, если вы можете вычислять достаточно сложные функции, то вы сумеете построить машину, которая будет весьма “умной” и сможет достигать сложных це-

лей. Таким образом, нам удается внести несколько большую ясность в вопрос о том, как может материя быть разумной, а именно: как могут фрагменты бездумной материи вычислять сложные функции.

Речь теперь идет не о неизменности надписи на поверхности золотого кольца и не о других статических запоминающих устройствах – интересующее нас состояние должно быть *динамическим*, оно должно меняться весьма сложным (и, хорошо бы, управляемым/программируемым) образом, переходя от настоящего к будущему. Расположение атомов должно быть менее упорядоченным, чем в твердом и жестком теле, где ничего интересного не происходит, но и не таким хаотичным, как в жидкости или в газе. Говоря точнее, мы бы хотели, чтобы наша система восприняла начальные условия задачи как свое исходное состояние, а потом, предоставленная самой себе, как-то эволюционировала, и ее конечное состояние мы бы могли рассматривать как решение данной ей задачи. В таком случае мы можем сказать, что система вычисляет нашу функцию.

В качестве первого примера этой идеи давайте построим из нашей неразумной материи очень простую (но от этого не менее важную) систему, вычисляющую функцию NAND[12]

[12] NAND представляет собой сокращение от двух английских слов NOT (не) и AND (и). Гейт AND выдает на выходе 1 только в том случае, если на входе две единицы. NAND делает в точности противоположное.

и потому получившую название гейт NAND[13]. У нее на входе два бита, а на выходе один: это 0, если оба бита на входе 1, во всех остальных случая – это 1. Если в одну сеть с батареей и электромагнитом мы вставим два замыкающих сеть ключа, то электромагнит сработает тогда, и только тогда, когда оба ключа замкнуты (находятся в состоянии “on”). Давайте поместим под ним еще один ключ, как показано на рис. 2.6, так что магнит, срабатывая, всякий раз будет размыкать его. Если мы интерпретируем первые два ключа как два бита на входе, а третий – как бит на выходе, то мы и получим то, что назвали гейтом NAND: третий ключ будет разомкнут только тогда, когда первые два замкнуты. Есть очень много более практичных способов сделать гейт NAND – например, с помощью транзисторов, как показано на рис. 2.6. В нынешних компьютерах гейты NAND чаще всего встроены в микросхемы или иные компоненты, выращенные из кристаллов кремния.

[13] В отечественной специальной литературе принято использовать для обозначения этих понятий термин “логический вентиль”, однако в последнее время транслитерация английского эквивалента “gate” стала выходить на первое место, в особенности в научно-популярной литературе. – *Прим. перев.*

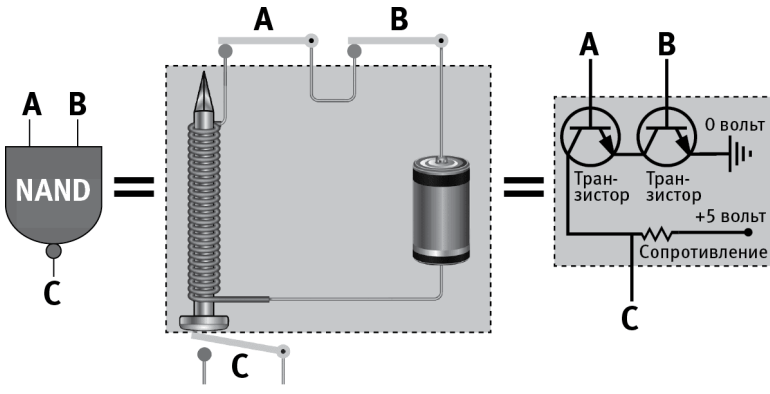


Рис. 2.6

Логический вентиль (гейт) NAND по заданным на входе двум битам A и B вычисляет третий бит C в соответствии с правилом: $C = 0$, если $A = B = 1$, и $C = 0$ в любом другом случае, – и посылает его на выход. В качестве гейта NAND можно использовать много различных физических устройств. В электрической цепи на средней части рисунка ключи A и B соответствуют битам на входе со значениями 0 при размыкании и 1 при замыкании. Когда они оба замкнуты, идущий через электромагнит ток размыкает ключ C. На схеме в правой части рисунка битам соответствуют значения потенциалов – 0, когда потенциал равен нулю, и 1, когда потенциал равен 5 вольтам. При подаче напряжения на базы обоих транзисторов (A и B) потенциал в точке C падает практически до нуля.

В информатике есть замечательная теорема, которая утверждает, что гейт NAND *универсален*: то есть вычисление любой вполне определенной функции[14] может быть осуществлено гейтами NAND, соединенными друг с другом. Так что если у вас есть достаточное количество гейтов NAND, вы можете собрать из них устройство, вычисляющее *все что угодно*! На случай, если у вас возникло желание посмотреть, как это работает, у меня есть схема (рис. 2.7), на которой вы увидите, как умножаются числа при помощи одних только гейтов NAND.

Исследователи из MIT Норман Марголус и Томмазо Тоффоли придумали слово "*computronium*" (компьютрониум), обозначающее любую субстанцию, которая может выполнять любые вычисления. Мы только что убедились, что создать компьютрониум не так уж и сложно: эта субстанция всего лишь должна быть способна соединять гейты NAND друг с другом любым желаемым способом. Разумеется, существуют и мириады других компьютрониумов. Например, еще один легко создать из предыдущего, заменив все гейты NAND на NOR: у него на выходе будет 1 только тогда, когда на оба входа подается 0. В следующем разделе мы об-

---

[14] Я называю "вполне определенной функцией" то же, что математики и информатики называют "вычислимой функцией", – то есть функцию, которая может быть вычислена каким-то гипотетическим компьютером, при условии что ему предоставлены неограниченные память и время. Алан Тьюринг и Алонсо Чёрч доказали, что существуют функции, которые могут быть описаны, но не могут быть вычислены.

судим нейронные сети, которые также способны выполнять произвольные вычисления, то есть и они ведут себя как компьютрониум. Ученый и предприниматель Стивен Вольфрам показал, что то же может быть сказано о простых устройствах, получивших название клеточных автоматов, которые периодически подправляют каждый бит в зависимости от того, в каком состоянии находятся биты по соседству. А еще в 1936 году Алан Тьюринг доказал в своей ставшей ключевой статье, что простая вычислительная машина (известная сейчас как "универсальный компьютер Тьюринга"), способная оперировать некоторыми символами на бумажной ленте по некоторым правилам, также способна выполнять любые вычисления. Одним словом, материя не просто обладает способностью к любым вполне определенным вычислениям, но и может производить их самыми разнообразными способами.

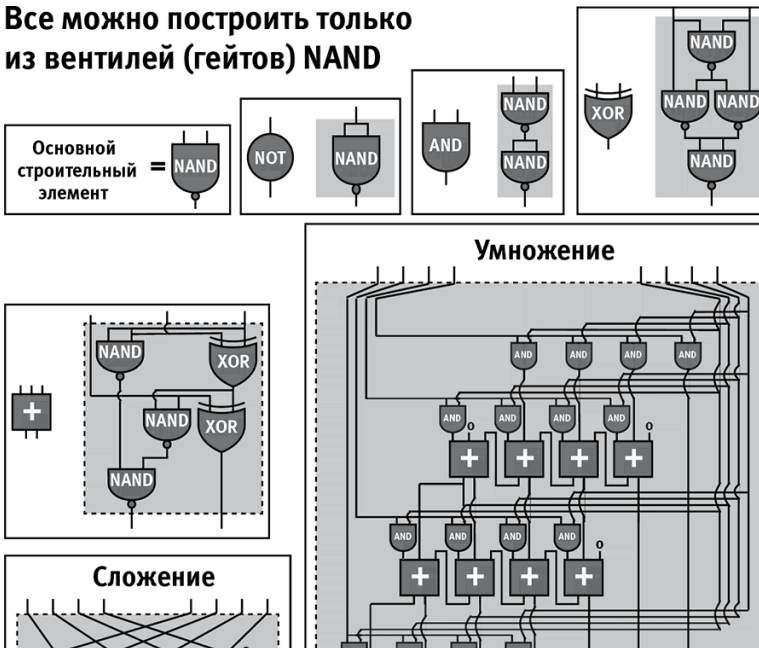


Рис. 2.7

Любое вполне определенное вычисление может быть выполнено при помощи комбинации гейтов одного-единственного типа NAND. Например, у модулей, выполняющих сложение и умножение и представленных на рисунке выше, на вход подается по два бинарных числа, каждое из которых представлено 4 битами, а на выходе получается бинарное

число, представленное 5 битами в первом случае, и бинарное число, представленное 8 битами во втором. Менее сложные модули NOT, AND, XOR и "+" (сложение трех одиночных битов в бинарное число, представляемое 2 битами) комбинируются из гейтов NAND. Полное понимание этой схемы исключительно сложно и абсолютно не нужно для дальнейшего чтения книги; я поставил ее здесь исключительно для иллюстрации идеи универсальности, ну и потакая своему внутреннему гику.

Как уже говорилось, Тьюринг в своей памятной статье 1936 года доказал также кое-что значительно более важное: если только компьютер обладает способностью производить некий весьма незначительный минимум операций, он *универсален* – в том смысле, что при достаточном количестве ресурсов он может сделать все то, на что способен любой другой компьютер. Он доказал универсальность "компьютера Тьюринга", а приближая его к физическому миру, мы только что показали, что семейство универсальных компьютеров включает в себя такие разные объекты, как сеть гейтов NAND или сеть соприкасающихся нейронов. Более того, Стивен Вольфрам заявил, что большая часть нетривиальных физических систем, от меняющейся погоды до мыслящего мозга, становятся универсальным компьютером, если позволить им как угодно менять свои размеры и не ограничивать их во времени.

Этот самый факт – а именно, что одно и то же вычисление может быть произведено на любом универсальном компьютере, как раз и означает, что вычисление не зависит от субстрата в том же самом отношении, в каком от него не зависит информация: каков бы физический субстрат ни был, оно живет там свою жизнь. Если вы – суперумный персонаж какой-то компьютерной игры будущего, обладающий сознанием, вам никогда не удастся узнать, породила ли вас рабочая станция под Windows, MacBook под MacOS или смартфон с Android, потому что вы субстрат-независимы. У вас не окажется и никаких способов определить, какого рода транзисторы используются микропроцессором этого компьютера.

Поначалу эта базовая идея субстрат-независимости привлекла меня тем, что у нее есть большое количество красивых иллюстраций в физике. Например, волны: у них есть разнообразные свойства – скорость, длина волны, частота, и физики могут решать связывающие их уравнения, совершенно не думая о том, как именно субстрат тут волнуется. Если вы слышите что-то, то вы регистрируете звуковые волны, распространяющиеся в той смеси газов, которую мы называем воздухом, и мы можем рассчитать относительно этих волн все что угодно – что их интенсивность уменьшается как квадрат расстояния, или как они проходят через открытую дверь или отражаются от стен, производя эхо, – ничего не зная о составе воздуха. На самом деле нам даже не обязательно знать, что он состоит из молекул: мы можем отвлечь-

ся ото всех подробностей относительно кислорода, азота или углекислого газа, потому что единственная характеристика этого субстрата, которая имеет значение и которая входит в знаменитое волновое уравнение, – это скорость звука, которую нам несложно померить и которая в данном случае будет равна примерно 300 метрам в секунду. Я рассказывал об этом волновом уравнении своим студентам на лекциях прошлой весной и говорил им, в частности, о том, что его открыли и им стали успешно пользоваться еще задолго до того, как физики установили, что молекулы и атомы вообще существуют!

Этот пример с волновым уравнением позволяет сделать три вывода. Во-первых, независимость от субстрата еще не означает, что без субстрата можно обойтись, но только лишь – что многие подробности его устройства не важны. Вы не услышите никакого звука в безвоздушном пространстве, но если замените воздух каким-нибудь другим газом, разницы не заметите. Точно так же вы не сможете производить вычисления без материи, но любая материя сгодится, если только ее можно будет организовать в гейты NAND, в нейронную сеть или в какие-то другие исходные блоки универсального компьютера. Во-вторых, субстрат-независимые явления живут свою жизнь, каков бы субстрат ни был. Волна пробегает по поверхности озера, хотя ни одна из молекул содержащейся в нем воды не делает этого, они только ходят вверх и вниз наподобие футбольных фанатов, устраивающих

“волну” на трибуне стадиона. В-третьих, часто нас интересует именно не зависящий от субстрата аспект явления: серфера обычно заботят высота волны и ее положение, а никак не ее молекулярный состав. Мы видели, что это так для информации, и это так для вычислений: если два программиста вместе ловят глюк в написанном ими коде, они вряд ли будут обсуждать транзисторы.

Мы приблизились к возможному ответу на наш исходный вопрос о том, как грубая физическая материя может породить нечто представляющееся настолько эфемерным, абстрактным и бестелесным, как разум: он кажется нам таким бестелесным из-за своей субстрат-независимости, из-за того, что живет своей жизнью, которая не зависит от физических деталей его устройства и не отражает их. Говоря коротко, вычисление – это определенная фигура пространственно-временного упорядочения атомов, и важны здесь не сами атомы, а именно эта фигура! Материя не важна.

Другими словами, “хард” здесь материя, а фигура – это “софт”. Субстрат-независимость вычисления означает, что AI возможен: разум не требует ни плоти, ни крови, ни атомов углерода.

Благодаря этой субстрат-независимости изобретательные инженеры непрерывно сменяют одну технологию внутри компьютера другой, радикально улучшенной, но не требовавшей замены “софта”. Результат во всех отношениях нагляден в истории запоминающих устройств. Как показывает

рис. 2.8, стоимость вычисления сокращается вдвое примерно каждые два года, и этот тренд сохраняется уже более века, снизив стоимость компьютера в миллион миллионов миллионов (в 10^{18}) раз со времен младенчества моей бабушки. Если бы все сейчас стало в миллион миллионов миллионов раз дешевле, то сотой части цента хватило бы, чтобы скупить все товары и услуги, произведенные или оказанные на Земле в тот год. Такое сильное снижение цены отчасти объясняет, почему сейчас вычисления проникают у нас повсюду, переместившись из отдельно стоящих зданий, занятых вычисляющими устройствами, в наши дома, автомобили и карманы – и даже вдруг оказываясь в самых неожиданных местах, например в кроссовках.

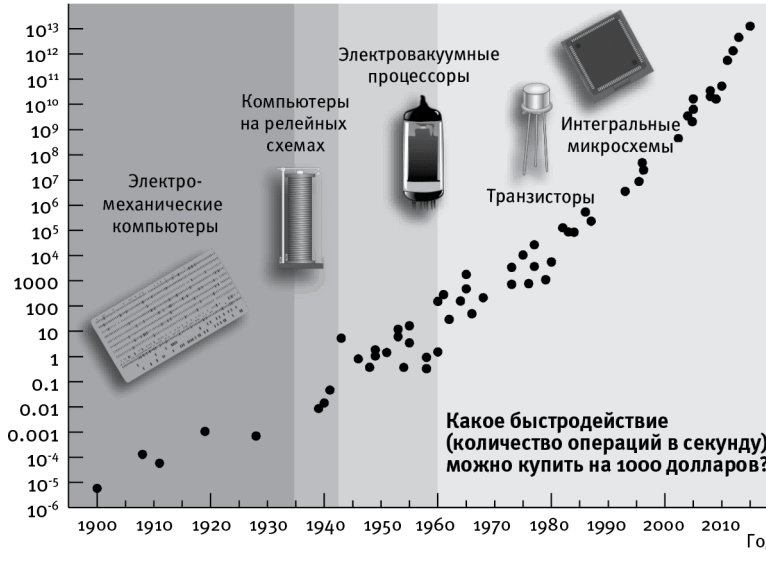


Рис. 2.8

С 1900 года вычисления становились вдвое дешевле примерно каждые пару лет. График показывает, какую вычислительную мощность, измеряемую в количестве операций над числами с плавающей запятой в секунду (FLOPS), можно было купить на тысячу долларов[5]. Частные случаи вычислений, которые соответствуют одной операции над числами с плавающей запятой, соответствуют 10^5 элементарным логическим операциям вроде обращения бита (замены 0 на 1, и наоборот) или одного срабатывания гейта NAND.

Почему развитие наших технологий позволяет им удва-

ивать производительность с такой регулярной периодичностью, обнаруживая то, что математики называют экспоненциальным ростом? Почему это сказывается не только на миниатюризации транзисторов (тренд, известный как закон Мура), но, и даже в большей степени, на развитии вычислений в целом (рис. 2.8), памяти (рис. 2.4), на море других технологий, от секвенирования генома до томографии головного мозга? Рэй Курцвейл называет это явление регулярного удвоения "законом ускоряющегося возврата".

В известных мне примерах регулярного удвоения в природных явлениях обнаруживается та же самая фундаментальная причина, и в том, что нечто подобное происходит в технике, нет ничего исключительного: и тут следующий шаг создается предыдущим. Например, вам самим приходилось переживать экспоненциальный рост сразу после того, как вас зачали: каждая из ваших клеточек, грубо говоря, ежедневно делится на две, из-за чего их общее количество возрастает день за днем в пропорции 1, 2, 4, 8, 16 и так далее. В соответствии с наиболее распространенной теорией нашего космического происхождения, известной как *теория инфляции*, наша Вселенная в своем младенчестве росла по тому же экспоненциальному закону, что и вы сами, удваивая свой размер за равные промежутки времени до тех пор, пока из крупинки меньше любого атома не превратилась в пространство, включающее все когда-либо виденные нами галактики. И опять причина этого заключалась в том,

что каждый шаг, удваивающий ее размер, служил основанием для совершения следующего. Теперь по тому же закону стала развиваться и технология: как только предыдущая технология становится вдвое мощнее, ее можно использовать для создания новой технологии, которая также окажется вдвое мощнее предыдущей, запуская механизм повторяющихся удвоений в духе закона Мура.

Но с той же регулярностью, как сами удвоения, высказываются опасения, что удвоения подходят к концу. Да, действие закона Мура рано или поздно прекратится: у миниатюризации транзистора есть физический предел. Но некоторые люди думают, что закон Мура синонимичен регулярному удвоению нашей технической мощи вообще. В противоположность им Рэй Курцвейл указывает, что закон Мура – это проявление не первой, а пятой технологической парадигмы, переносящей экспоненциальный рост в сферу вычислительных технологий, как показано на рис. 2.8: как только предыдущая технология перестает совершенствоваться, мы заменяем ее лучшей. Когда мы не можем больше уменьшать вакуумные колбы, мы заменяем их полупроводниковыми транзисторами, а потом и интегральными схемами, где электроны движутся в двух измерениях. Когда и эта технология достигнет своего предела, мы уже представляем, куда двинуться дальше: например, создавать трехмерные интегральные цепи или делать ставку на что-то отличное от электронов.

Никто сейчас не знает, какой новый вычислительный субстрат вырвется в лидеры, но мы знаем, что до пределов, положенных законами природы, нам еще далеко. Мой коллега по MIT Сет Ллойд выяснил, что это за фундаментальный предел, и мы обсудим его в главе 6, и этот предел на целых 33 порядка (то есть в 10^{33} раза) отстоит от нынешнего положения вещей в том, что касается способности материи производить вычисления. Так что если мы будем и дальше удваивать производительность наших компьютеров каждые два – три года, для достижения этой последней черты нам понадобится больше двух столетий.

Хотя каждый универсальный компьютер способен на те же вычисления, что и любой другой, некоторые из них могут отличаться от прочих своей высокой производительностью. Например, вычисление, требующее миллионов умножений, не требует миллионов различных совершающих умножение модулей с использованием различных транзисторов, как показано на рис. 2.6, – требуется только один такой модуль, который можно использовать многократно при соответствующей организации ввода данных. В соответствии с этим духом максимизации эффективности большинство современных компьютеров действуют согласно парадигме, подразумевающей разделение всякого вычисления на много шагов, в перерывах между которыми информация переводится из вычислительных модулей в модули памяти и обратно. Такая архитектура вычислительных устройств была разработа-

на между 1935 и 1945 годами пионерами компьютерных технологий – такими, как Алан Тьюринг, Конрад Цузе, Преспер Эккерт, Джон Мокли и Джон фон Нейман. Ее важная особенность заключается в том, что в памяти компьютера хранятся не только данные, но и его "софт" (то есть программа, определяющая, что надо делать с данными). На каждом шагу центральный процессор выполняет очередную операцию, определяющую, что именно надо сделать с данными. Еще одна часть памяти занята тем, чтобы определять, каков будет следующий шаг, просто пересчитывая, сколько шагов уже сделано, она так и называется – *счетчик команд*: это часть памяти, где хранится номер исполняемой команды. Переход к следующей команде просто прибавляет единицу к счетчику. Для того чтобы перейти к нужной команде, надо просто задать программному счетчику нужный номер – так и поступает оператор "если", устраивая внутри программы петлевой возврат к уже пройденному.

Современным компьютерам удается значительно ускорить выполнение вычислений, проводя их, что называется, "параллельно", в продолжение идеи повторного использования одних и тех же модулей: если вычисление можно разделить на части и каждую часть выполнять самостоятельно (поскольку результат одной не требуется для выполнения другой), то тогда эти части можно вычислять одновременно в разных составляющих "харда".

Идеально воплощение параллельности достигается

в *квантовом компьютере*. Пионер теории квантовых вычислений Дэвид Дойч утверждал в полемическом запале, что "квантовый компьютер распределяет доступную ему информацию по бесчисленному множеству копий себя самого во всем мультиверсуме" и решает благодаря этому здесь, в нашей Вселенной, любую задачу гораздо быстрее, потому что, в каком-то смысле, получает помощь от других версий самого себя[6]. Мы пока еще не знаем, будет ли пригодный для коммерческого использования квантовый компьютер создан в ближайшие десятилетия, поскольку это зависит и от того, действительно ли квантовая физика работает так, как мы думаем, и от нашей способности преодолеть связанные с его созданием серьезнейшие технические проблемы, но и коммерческие компании, и правительства многих стран мира вкладывают ежегодно десятки миллионов долларов в реализацию этой возможности. Хотя квантовый компьютер не поможет в разгоне заурядных вычислений, для некоторых специальных типов были созданы изобретательные алгоритмы, способные изменить скорость кардинально – в частности, это касается задач, связанных со взломом криптосистем и обучением нейронных сетей. Квантовый компьютер также способен эффективно симулировать поведение квантово-механических систем, включая атомы, молекулы и новые соединения, заменяя измерения в химических лабораториях примерно в том же ключе, в каком расчеты на обычных компьютерах заменили, сделав ненужными, измерения

в аэродинамических трубах.

Что такое обучение?

Хотя даже карманный калькулятор легко обгоняет меня в состязании на быстроту в арифметических подсчетах, он никогда не улучшит своих показателей ни по быстроте вычислений, ни по их точности, сколько бы ни тренировался. Он ничему не учится, и каждый раз, когда я, например, нажимаю кнопку извлечения квадратного корня, он вычисляет одну и ту же функцию, точно повторяя одни и те же действия. Точно так же первая компьютерная программа, обыгравшая меня в шахматы, не могла учиться на своих ошибках и каждый раз просчитывала одну и ту же функцию, которую умный программист разработал, чтобы оценить, насколько хорош тот или иной следующий ход. Напротив, когда Магнус Карлсен в возрасте пяти лет проиграл свою первую игру в шахматы, он начал процесс обучения, и это принесло ему восемнадцать лет спустя титул чемпиона мира по шахматам.

Способность к обучению, как утверждается, – основная черта сильного интеллекта. Мы уже видели, как кажущийся бессмысленным фрагмент неживой материи оказывается способным запоминать и вычислять, но как он может учиться? Мы видели, что поиск ответа на сложный вопрос подразумевает вычисление некоторых функций, и определенным образом организованная материя может вычислить любую вычислимую функцию. Когда мы, люди, впервые создали

карманные калькуляторы и шахматные программы, мы как-то организовали материю. И теперь, для того чтобы учиться, этой материи надо как-то, просто следуя законам физики, *реорганизовывать себя*, становясь все лучше и лучше в вычислении нужных функций.

Чтобы демистифицировать процесс обучения, давайте сначала рассмотрим, как очень простая физическая система может научиться вычислять последовательность цифр в числе π или любом другом числе. Выше мы видели, как холмистую поверхность с множеством ям между холмами (рис. 2.3) можно использовать в качестве запоминающего устройства: например, если координата одной из ям точно равна $x = \pi$ и поблизости нет никаких других ям, то, положив шарик в точку с координатой $x = 3$, мы увидим, как наша система вычисляет отсутствующие знаки после запятой, просто наблюдая, как шарик скатывается в ямку. Теперь предположим, что поверхность сделана из мягкой глины, поначалу совершенно плоской как стол. Но если какие-то фанаты-математики будут класть шарики в одни и те же точки с координатами, соответствующими их любимым числам, то благодаря гравитации в этих точках постепенно образуются ямки, и со временем эту глиняную поверхность можно будет использовать, чтобы узнать, какие числа она "запомнила". Иными словами, глина выучила, как ей вычислить значащие цифры числа π.

Другие физические системы, в том числе и мозг, могут

учиться намного эффективнее, но идея остается той же. Джон Хопфилд показал, что его сеть пересекающихся нейронов, о которой шла речь выше, может учиться подобным же образом: если вы раз за разом приводите ее в одни и те же состояния, она постепенно изучит эти состояния и будет возвращаться в какое-то из них, оказавшись где-то поблизости. Вы хорошо помните членов вашей семьи, поскольку часто их видите, и их лица всплывают в вашей памяти всякий раз, как только ее подталкивает к этому что-либо связанное с ними.

Теперь благодаря нейронным сетям трансформировался не только биологический, но и искусственный интеллект, и с недавнего времени они начали доминировать в такой исследовательской области, связанной с искусственным интеллектом, как *машинное обучение* (изучение алгоритмов, которые улучшаются вследствие приобретения опыта). Прежде чем углубиться в то, как эти сети могут учиться, давайте сначала поймем, как они могут выполнять вычисления. Нейронная сеть – это просто группа нейронов, соприкасающихся друг с другом и потому способных оказывать взаимное влияние. Ваш мозг содержит примерно столько же нейронов, сколько звезд в нашей Галактике – порядка сотен миллиардов. В среднем каждый из этих нейронов контактирует примерно с тысячей других через переходы, называемые синапсами – именно сила этих синаптических связей, которых

насчитывается примерно сотни триллионов, кодирует большую часть информации в вашем мозгу.

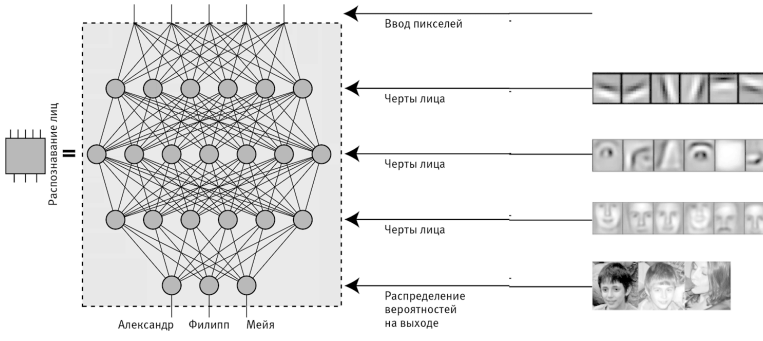


Рис. 2.9

Сеть из нейронов может выполнять вычисления функций так же, как это делает сеть из гейтов NAND. Например, сети искусственных нейронов обучились по вводимым числам, представляющим собой яркость пикселей изображения, давать на выходе числа, соответствующие вероятностям, что на этих изображениях тот или иной человек. Каждый искусственный нейрон (желтый кружок) вычисляет взвешенную сумму чисел, отправленных ему через связи (прямые линии) от нейронов предыдущего слоя, применяет простую функцию и посылает результат нейронам следующего слоя – чем дальше, тем больше вычисляется подробностей. Типичная нейронная сеть, способная распознавать лица, содержит сотни тысяч нейронов. На этом рисунке для простоты показана

лишь жалкая горсточка.

Мы можем схематически изобразить нейронную сеть в виде точек, представляющих нейроны, и соединяющих их линий, которые представляют синапсы (см. рис. 2.9). Настоящие синапсы – это довольно сложные электрохимические устройства, совсем не похожие на эту схематическую иллюстрацию: они включают в себя разные части, которые называют аксонами и дендритами; есть много разновидностей нейронов, которые действуют по-разному, и точные детали того, как и когда электрическая активность в одном нейроне влияет на другие нейроны, все еще остаются предметом дальнейших исследований. Однако уже сейчас ясно, что нейронные сети могут достичь производительности человеческого уровня во многих удивительно сложных задачах, даже если на время забыть обо всех этих сложностях и заменить настоящие биологические нейроны чрезвычайно простыми имитирующими их устройствами, совершенно одинаковыми и подчиняющимися очень простым правилам. В настоящее время наиболее популярная модель такой *искусственной нейронной сети* представляет состояние каждого нейрона одним числом и силу каждого синапса – тоже одним числом. В этой модели при каждом действии каждый нейрон обновляет свое состояние, вычисляя среднее арифметическое от состояния всех присоединенных к нему нейронов с весами, в качестве которых берутся силы их синаптической свя-

зи. Иногда еще прибавляется константа, а к результату применяется так называемая *функция активации*, дающая число, которым будет выступать в качестве состояния данного нейрона на следующем такте[15]. Самый простой способ использовать нейронную сеть как функцию заключается в том, чтобы сделать ее прямой, превратив в канал передачи, где информация направляется лишь в одну сторону, как показано на рис. 2.9, загружая на вход функции верхний слой нейронов и считывая выход со слоя нейронов внизу.

Успешное использование этой простой нейронной сети представляет нам еще один пример независимости от субстрата: нейронная сеть обладает колоссальной вычислительной силой, которая, вне всякого сомнения, не зависит от мелких подробностей в ее устройстве. В 1989 году Джордж Цибенко, Курт Хорник, Максвелл Стинчкомб и Халберт Уайт доказали нечто замечательное: простые нейронные сети вроде только что описанной *универсальны* в том смысле, что они могут вычислять *любую* функцию с произволь-

---

[15] Добавим для тех, кто любит математику, что в качестве этой функции чаще всего выступает либо сигмоидальная функция $\sigma(x) - 1/(1 + e^{-x})$, либо пороговая функция $\sigma(x) = \max\{0, x\}$, хотя доказано, что в этой роли можно использовать какую угодно, лишь бы она не была линейной (то есть не представлялась в виде прямой линии на графике). В знаменитой модели Хопфилда использовалась функция $\sigma(x) = -1$ if $x < 0$ and $\sigma(x) = 1$ if $x \geq 0$. Если состояния нейронов хранятся в памяти в виде вектора, то при переходе к следующему такту он обновляется умножением сначала этого вектора на матрицу, элементами которой служат силы синаптических связей, и последующим применением функции $\int(x)$ ко всем новым вычисленным элементам.

ной точностью, просто приписывая соответствующие значения числам, которыми характеризуются силы синаптических связей. Другими словами, эволюция, вероятно, сделала наши биологические нейроны такими сложными не потому, что это было необходимо, а потому, что это было более эффективно, и потому, что эволюция, в отличие от инженеров-людей, не получает наград за простоту и понятность предлагаемых конструкций.

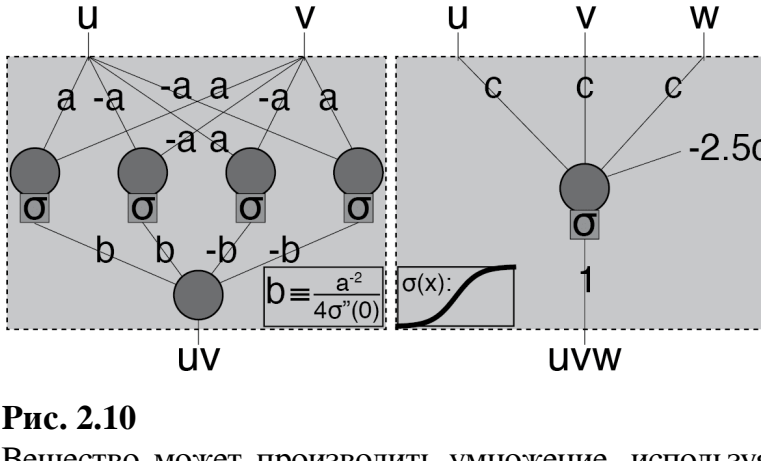


Рис. 2.10

Вещество может производить умножение, используя не гейты NAND, как на рис. 2.7, а нейроны. Для понимания ключевого момента здесь не требуется вникать в детали, достаточно только отдавать себе отчет, что нейроны (как биологические, так и искусственные) не только способны про-

изводить математические действия, но их для этого требуется значительно меньше, чем гейтов NAND. *Вот еще факультативные детали для упертых фанатов математики*: кружочками обозначено сложение, квадратики обозначают применение функции σ, а прямые отрезки – умножение на число, которое этот отрезок пересекает. На входе – вещественное число (слева) или бит (справа). Умножение становится сколь угодно точным при $a \to 0$ (слева) и при $c \to \infty$ (справа). Левая сеть работает при любой функции $\sigma(x)$, имеющей изгиб в нуле ($\sigma''(0) \neq 0$), что можно доказать разложением функции $\sigma(x)$ по формуле Тейлора. Для сети справа надо, чтобы функция $\sigma(x)$ стремилась к нулю и к единице при очень малых и очень больших x соответственно, так чтобы соблюдалось условие $uvw = 1$, только когда $u + v + w = 3$. (Эти примеры взяты из статьи моего студента Генри Лина: https://arxiv.org/abs/1608.08225, проверена 18 мая 2018.) Комбинируя умножения и сложения, можно вычислять любые полиномы, с помощью которых, как известно, мы можем получить апроксимацию любой гладкой функции.

Впервые услышав об этом, я был озадачен: как что-то до такой степени простое может вычислить нечто произвольно сложное? Например, как вы сможете даже просто-напросто что-то перемножать, когда вам разрешено только вычислять взвешенные средние значения и применять одну фиксированную функцию? Если вам захочется проверить, как это ра-

ботает, на рис. 2.10 показано, как всего пять нейронов могут перемножать два произвольных числа и как один нейрон может перемножить три бита.

Хотя вы можете доказать *теоретическую* возможность вычисления чего-либо произвольно большой нейронной сетью, ваше доказательство ничего не говорит о том, можно ли это сделать на практике, располагая сетью разумного размера. На самом деле, чем больше я об этом думал, тем больше меня удивляло, что нейронные сети и в самом деле так хорошо работали.

Предположим, что у вас есть черно-белые мегапиксельные фотографии, и вам их надо разложить в две стопки – например, отделив кошек от собак. Если каждый из миллиона пикселей может принимать одно из, скажем, 256 значений, то общее количество возможных изображений равно $256^{1000000}$, и для каждого из них мы хотим вычислить вероятность того, что на нем кошка. Это означает, что произвольная функция, которая устанавливает соответствие между фотографиями и вероятностями, определяется списком из $256^{1000000}$ позиций, то есть числом большим, чем атомов в нашей Вселенной (около 10^{78}). Тем не менее нейронные сети всего лишь с тысячами или миллионами параметров каким-то образом справляются с такими классификациями довольно хорошо. Как успешные нейронные сети могут быть "дешевыми" в том смысле, что от них требуется так мало па-

раметров? В конце концов, вы можете доказать, что нейронная сеть, достаточно маленькая для того, чтобы вписаться в нашу Вселенную, потерпит грандиозное фиаско в попытке аппроксимировать почти все функции, преуспев лишь в смехотворно крошечной части всех вычислительных задач, решения которых вы могли бы от нее ждать.

Я получил огромное удовольствие, разбираясь с этой и другими, связанными с ней, загадками вместе со студентом по имени Генри Лин. Среди разнообразных причин испытывать благодарность к своей судьбе – возможность сотрудничать с удивительными студентами, и Генри – один из них. Когда он впервые зашел в мой офис и спросил, хотел бы я поработать с ним, я подумал, что, скорее, мне надо было бы задавать такой вопрос: этот скромный, приветливый юноша с сияющими глазами из крошечного городка Шревепорт в штате Луизиана уже успел опубликовать восемь научных статей, получить премию Forbes 30-Under-30 и записать лекцию на канале TED, получившую более миллиона просмотров – и это всего-то в двадцать лет! Год спустя мы вместе написали статью, в которой пришли к удивительному заключению: вопрос, почему нейронные сети работают так хорошо, не может быть решен только методами математики, потому что значительная часть этого решения относится к физике.

Мы обнаружили, что класс функций, с которыми нас познакомили законы физики и которые, собственно, и заставили нас заинтересоваться вычислениями, – это удивительно

узкий класс функций, потому что по причинам, которые мы все еще не полностью понимаем, законы физики удивительно просты. Более того, крошечная часть функций, которую могут вычислить нейронные сети, очень похожа на ту крошечную часть, интересоваться которыми нас заставляет физика! Мы также продолжили предыдущую работу, показывающую, что нейронные сети глубокого обучения (слово "глубокое" здесь подразумевает, что они содержат много слоев) гораздо эффективнее, чем мелкие, для многих из этих функций, представляющих интерес. Например, вместе с еще одним удивительным студентом MIT, Дэвидом Ролником, мы показали, что простая задача перемножения n чисел требует колоссальных 2^n нейронов для сети с одним слоем и всего лишь около 4n нейронов в глубокой сети. Это помогает объяснить не только возросший энтузиазм среди исследователей AI по отношению к нейронным сетям, но также и то, зачем эволюции понадобились нейронные сети у нас в мозгу: если мозг, способный предвидеть будущее, дает эволюционное преимущество, в нем должна развиваться вычислительная архитектура, пригодная для решения именно тех вычислительных задач, которые возникают в физическом мире.

Теперь, когда мы знаем, как нейронные сети работают и как вычисляют, давайте вернемся к вопросу о том, как они могут учиться. В частности, как может нейронная сеть улучшать свои вычислительные способности, обновляя состояние своих синапсов.

Канадский психолог Дональд Хебб в своей книге 1949 года *The Organization of Behavior*, вызвавшей живой отклик, утверждал, что если бы два соседних нейрона часто оказывались активны (“светились”) одновременно, то их синаптическая связь усиливалась бы, обучая их включать друг друга – эта идея нашла отражение в популярной присказке “Связаны вместе, светятся вместе”. Хотя до понимания в подробностях, как именно происходит обучение в настоящем мозгу, нам еще далеко, и исследования показывают, что ответы во многих случаях должны будут далеко выходить за рамки простых предложенных правил вроде того, что стало известно как “обучение по Хеббу”, даже эти простые правила, тем не менее, способны объяснить, каким образом происходит обучение нейронных сетей во многих интересных случаях. Джон Хопфилд ссылался на обучение по Хеббу, которое позволило его исключительно простой искусственной нейронной сети сохранить много сложных воспоминаний путем простого повторения. Такое экспонирование информации в целях обучения обычно называют “тренировкой”, когда речь идет об искусственных нейронных сетях (а также о животных или о людях, которым надо приобрести определенный навык), хотя слова “опыт”, “воспитание” или “образование” тоже подходят. В искусственных нейронных сетях, лежащих в основе современных систем AI, обучение по Хеббу заменено, как правило, более сложными правилами с менее благозвучными названиями, такими как обратное

распространение ошибки (backpropagation) или спуск по стохастическому градиенту (stochastic gradient descent), но основная идея одна и та же: существует некоторое простое детерминированное правило, похожее на закон физики, с помощью которого синапсы со временем обновляются. Словно по волшебству, пользуясь этим простым правилом, нейронную сеть можно научить чрезвычайно сложным вычислениям, если задействовать при обучении большие объемы данных. Мы пока еще не знаем точно, какие правила использует при обучении наш мозг, но, каков бы ни был ответ, нет никаких признаков, что эти правила нарушают законы физики.

Большинство цифровых компьютеров увеличивают эффективность своей работы, разбивая задачу на много шагов и многократно используя одни и те же вычислительные модули, – искусственные и биологические нейронные сети поступают аналогично. В мозгу есть области, представляющие собой то, что в информатике принято называть *рекуррентными* нейронными сетями: информация внутри них может протекать в различных направлениях, и то, что на предыдущем такте служило выходом, может стать входом в последующем – в этом их отличие от сетей прямой передачи. Сеть логических гейтов в микропроцессоре ноутбука также рекуррентна в этом смысле: она продолжает использовать уже обработанную информацию, позволяя в то же время вводить новую – с клавиатуры, трекпада, камеры и т. п., которой также позволяется влиять на текущие вычисления, а это,

в свою очередь, определяет, как будет осуществляться вывод информации: на монитор, динамики, принтер или через беспроводную сеть. Аналогично нейронная сеть в вашем мозгу рекуррентна, поскольку получает информацию от ваших глаз, ушей и других органов чувств и позволяет этой информации влиять на текущее вычисление, которое, в свою очередь, определяет, как будет производиться вывод результатов к вашим мышцам.

История обучения по крайней мере столь же длинна, как и история самой жизни, поскольку каждый самовоспроизводящийся организм так или иначе производит копирование и обработку информации, то есть как-то себя ведет, чему ему надо было каким-то образом научиться. Однако в эпоху Жизни 1.0 организмы не учились в течение своей жизни: способы обработки информации и реакции на нее определялись унаследованной организмом ДНК, поэтому обучение происходило медленно, на уровне видов, через дарвиновскую эволюцию от поколения к поколению.

Около полумиллиарда лет назад некоторые генные линии здесь, на Земле, открыли путь к возникновению животных, обладающих нейронными сетями, и это дало таким животным способность менять свое поведение, обучаясь на опыте в течение своей жизни. Когда появилась Жизнь 2.0, она, благодаря своей способности учиться значительно быстрее, победила в соревновании видов и распространилась по планете словно лесной пожар. В первой главе мы уже выяснили,

что жизнь постепенно улучшала свои способности обучаться, причем со все возрастающей скоростью. У одного вида обезьянообразных мозг оказался настолько хорошо приспособленным к обучению, что они научились пользоваться разными орудиями, разговаривать, стрелять и создали развитое общество, распространившееся по всему миру. Это общество само по себе можно рассматривать как систему, которая запоминает, вычисляет и учится, и всё это оно делает с неуклонно возрастающей скоростью, так как одно изобретение влечет за собой следующее: письменность, книгопечатание, современная наука, компьютеры, интернет и т. д. Что следующим поместят будущие историки в этом списке изобретений, ускоряющих обучение? Я думаю, следующим будет искусственный интеллект.

Как все мы знаем, лавина технических достижений, обеспечивших совершенствование компьютерной памяти и рост вычислительной мощности компьютеров (рис. 2.4 и рис. 2.8), привели к впечатляющему прогрессу в искусственном интеллекте, но потребовалось немало времени, пока машинное обучение достигло зрелости. Когда созданный IBM компьютер Deep Blue в 1997 году обыграл чемпиона мира по шахматам Гарри Каспарова, его главные преимущества заключались в памяти и способности быстро и точно считать, – но не в умении учиться. Его вычислительный интеллект был создан группой людей, и ключевая причина, по которой Deep Blue смог обыграть своих создателей, заключалась в его спо-

собности быстрее считать, и потому он мог анализировать больше возникающих в игре позиций. Когда созданный IBM компьютер Watson обошел человека, показавшего себя сильнейшим в викторине *Jeopardy!*, он тоже опирался не на обучение, а на специально запрограммированные навыки и превосходство в памяти и быстродействии. То же самое можно сказать обо всех прорывных технологиях в робототехнике, от самобалансирующихся транспортных средств до беспилотных автомобилей и ракет, приземляющихся в автоматическом режиме.

Напротив, движущей силой многих последних достижений AI стало *машинное обучение*. Посмотрите, например, на рис. 2.11. Вы сразу догадаетесь, что на этой фотографии, но запрограммировать функцию, на входе которой, ни много ни мало, цвет каждого из пикселей изображения, а на выходе – точно описывающая фотографию подпись, например: "Группа молодых людей, играющих во фризби", – в течение десятилетий не удавалось ни одному из многочисленных исследователей искусственного интеллекта во всем мире. И только команда Google смогла сделать именно это в 2014 году[7]. Если ввести другой набор пикселей, на выходе появится: "Стадо слонов, идущих по сухому травяному полю", – и снова ответ точный. Как они это смогли? Программируя вручную, как Deep Blue, создавая по отдельности каждый алгоритм, опознающий игру фризби, лица и все такое? Нет, они создали относительно простую нейронную сеть, не об-

ладавшую поначалу никаким знанием о физическом мире и его составляющих, а потом дали ей возможность учиться, предоставив колоссальный объем информации. В 2004 году знаменитый визионер Джефф Хокинс, рассуждая об искусственном интеллекте, писал: “Никакой компьютер не может … видеть так же хорошо, как мышь”, – но те времена давно уже прошли.

Рис. 2.11

“Группа людей, играющих во фризби” – такую подпись к этой фотографии сгенерировала машина, ничего не знающая ни о людях, ни об играх, ни о фризби.

Так же, как мы не вполне понимаем, как учатся наши де-

ти, мы все еще не до конца поняли, как учатся такие нейронные сети и почему они иногда терпят неудачу. Но уже ясно, что они будут очень полезны, и поэтому глубокое обучение стало привлекать инвесторов. Благодаря глубокому обучению сильно изменились подходы к технической реализации компьютерного зрения: от распознавания рукописного текста до анализа видеопотоков в реальном времени и беспилотных автомобилей. Благодаря ему произошла революция в способах преобразовывать с помощью компьютера устную речь в письменный текст и переводить его на другие языки, даже в реальном времени, поэтому мы можем теперь поговорить с персональными цифровыми помощниками, такими как Siri, Google Now или Cortana. Раздражающие головоломки типа CAPTCHA, разгадывая которые мы должны убедить сайт, что мы люди, становятся все труднее, чтобы обогнать технологии машинного обучения. В 2015 году Google DeepMind выпустил систему с искусственным интеллектом, которая с помощью глубокого обучения осваивала десятки различных компьютерных игр примерно так же, как это делает ребенок, – то есть не пользуясь инструкциями, с той единственной разницей, что научалась играть лучше любого человеческого существа.

В 2016 году та же самая компания выпустила AlphaGo – компьютерную систему, играющую в го, которая при помощи глубокого обучения стала так точно оценивать позиционные преимущества расположения камней на доске, что по-

бедила сильнейшего игрока в мире. Этот успех служит положительной обратной связью, привлекая все больше финансирования и все больше талантливой молодежи в исследования искусственного интеллекта, которые приводят к новому успеху.

Мы посвятили эту главу природе интеллекта и его развитию до настоящего времени. Сколько времени потребуется, чтобы машины смогли обойти нас в решении всех когнитивных задач? Мы этого не знаем и должны быть готовы к тому, что ответом окажется "никогда". Однако смысл этой главы в том, чтобы мы подготовили себя также и к тому, что это все-таки произойдет, и, не исключено, даже еще при нашей жизни. В конце концов, материя может быть устроена так, что, когда она подчиняется законам физики, она запоминает, вычисляет и учится, – причем материя не обязательно биологической природы. Исследователей искусственного интеллекта часто обвиняют в том, что они слишком много обещают и слишком мало своих обещаний выполняют, но справедливости ради надо заметить, что у многих таких критиков послужной список тоже далеко не безупречен. Некоторые из них просто жонглируют словами, то определяя интеллект как нечто такое, чего компьютеры пока еще не могут, то как нечто такое, что произведет на нас наибольшее впечатление. Компьютеры теперь стали очень хороши или даже превосходны в арифметике, в игре в шахматы, в доказательстве математических теорем, подборе акций, распознавании

образов, вождении автомобиля, аркадных играх, го, синтезе речи, преобразовании устной речи в письменную, переводе с языка на язык и диагностике рака, но иной критик лишь презрительно хмыкнет: “Конечно же, для этого не нужен настоящий разум!”. Он будет продолжать утверждать, что настоящий разум должен добраться до вершин ландшафта Моравеца (рис. 2.2), пока еще не скрывшихся под водой, подобно тем людям в прошлом, которые утверждали, что ни субтитры под картинкой, ни игра в го машине не под силу, – а вода продолжала прибывать.

Исходя их того, что вода будет прибывать еще как минимум некоторое время, можно предположить, что влияние искусственного интеллекта на общество будет расти. Задолго до того как AI достигнет человеческого уровня в решении всех задач, он успеет открыть нам новые увлекательные возможности и задать нам много новых вопросов в самых разных областях, связанных с инфекционными болезнями, законодательными системами, разоружением и созданием новых рабочих мест. Каковы они, и как мы можем лучше подготовиться к ним? Давайте рассмотрим это в следующей главе.

Подведение итогов

• Интеллект, определяемый как способность достигать сложных целей, не может быть измерен одним только IQ, он

должен быть представлен спектральной плотностью в соответствии со способностями к достижению *любых* целей.

• Современный искусственный интеллект имеет тенденцию к узкой специализации, причем каждая система может достигать только очень конкретных целей, – в отличие от интеллекта человека, чрезвычайно широкого.

• Память, вычисление, обучение и разум представляются чем-то абстрактным, нематериальным и эфемерным, потому что они независимы от субстрата: они живут своей жизнью, не отражая ни деталей своего устройства, ни особенностей основного материального субстрата.

• Любая материя может быть основой для памяти, если у используемого ее фрагмента достаточно разных стабильных состояний.

• Любая материя может стать *компьюториумом*, то есть вычислительным (компутационным) субстратом, надо только, чтобы в ней содержались определенные универсальные строительные блоки, которые могут быть объединены для вычисления любой функции. Гейты NAND и нейроны дают два важнейших примера таких универсальных "вычислительных атомов".

• Нейронная сеть является мощным *обучающимся* субстратом, потому что, просто подчиняясь законам физики, она может преобразовываться, становясь все более пригодной для выполнения требуемых вычислений.

• Из-за поразительной простоты законов физики нас, лю-

дей, интересует лишь крошечная часть всех мыслимых вычислительных задач, а нейронные сети, как правило, именно для решения задач из этой крошечной части идеально подходят.

• Как только технология удваивает свою изначальную производительность, ее часто можно использовать для создания новой технологии, которая, в свою очередь, становится вдвое производительнее старой, что приводит к повторному удвоению возможностей в духе закона Мура. Уже на протяжении целого столетия стоимость информационных технологий сокращается вдвое примерно раз в два года, что и привело к нынешней информационной эре.

• Если развитие технологий искусственного интеллекта будет продолжаться, то задолго до того как AI достигнет человеческого уровня в решении всех задач, он успеет открыть нам новые увлекательные возможности и задать много новых вопросов в самых разных областях, связанных с инфекционными болезнями, законодательными системами, разоружением и созданием новых рабочих мест, каковые мы рассмотрим в следующей главе.

Глава 3
Ближайшее будущее: болезни, законы, оружие и работа

Если мы не поспешим изменить направление, мы рискуем прибыть туда, откуда отбыли.
Ирвин Кори

Что значит быть человеком в наше время? Например, что мы по-настоящему в себе ценим, что отличает нас от других форм жизни и от машин? Что другие люди ценят в нас, благодаря чему некоторые из них предлагают нам работу? Какие бы ответы на эти вопросы мы ни дали, ясно, что по мере развития технологий нам придется со временем изменять их.

Возьмите, например, меня. Как ученый я горжусь тем, что смог поставить перед собой собственные цели, мне достало ума и интуиции, чтобы решить довольно много не решенных до меня задач, и я сумел воспользоваться языком, чтобы сообщить о своих находках другим. К счастью для меня, общество оказалось готово заплатить мне за эту работу. Столетия назад я мог бы, наверное, как и многие другие, построить свою идентичность фермера или ремесленника, но с тех пор развитие технологий сильно сократило область, занимаемую такими профессиями. Это означает, что теперь стало невоз-

можно каждому строить свою идентичность в сельском хозяйстве или в ремеслах.

Лично меня совсем не беспокоит, что сегодняшние машины превосходят меня в навыках ручного труда – в копании или вязании: для меня это не хобби, не источник дохода и не повод собою гордиться. В самом деле, любые иллюзии, которые могли у меня возникнуть по этому поводу, разбились, когда мне было всего восемь лет: у меня были уроки вязания в школе, показавшие мою полную неспособность к этому делу, и я смог хоть как-то справиться с данным мне заданием только благодаря помощи сострадательной пятиклассницы, сжалившейся надо мной.

Но если технологии будут продолжать развиваться, не случится ли так, что AI со временем превзойдет людей также и в том, чем я горжусь сейчас и за что меня ценят на рынке труда? Стюарт Рассел признавался мне, как ему с коллегами довелось недавно испытать момент искушения “выразиться по матушке”, когда они вдруг стали свидетелями такого, чего не ожидали от искусственного интеллекта еще много-много лет. Позвольте, пожалуйста, и мне рассказать вам о некоторых подобных моментах, в которых я вижу грядущую победу над многими из человеческих способностей.

Прорывы

Системы глубокого обучения с подкреплением и его агенты

В 2014 году, когда я смотрел видео, на котором разработанная DeepMind система с искусственным интеллектом училась играть в компьютерные игры, у меня отвисла челюсть. В особенности хорошо искусственному интеллекту удавалось играть в Breakout (см. рис. 3.1), классическую игру Atari, с нежностью вспоминаемую мной с подросткового возраста. Цель игры в том, чтобы, перемещая платформу, заставлять шарик биться о кирпичную стену. Всякий раз, когда удается выбить из стены кирпич, он пропадает, а счет увеличивается.

В тот день я написал несколько компьютерных игр, и хорошо знал, что написать программу, которая может сыграть в Breakout, совсем не трудно, но это было не то, что сделала команда DeepMind. Они сделали другое: создали девственно чистый AI, который ничего не знал об этой игре, как и о любых других играх, и вдобавок не имел никакого понятия о том, что такое игры, платформы, кирпичи или шарики. Их AI знал лишь одно: длинный список чисел, загружающихся через равные интервалы времени и представляю-

щих текущий счет, и еще один длинный список, которые мы (но не AI) интерпретировали бы как описание цвета и освещенности разных частей экрана. AI просто велели максимизировать счет, выставляя с регулярными интервалами числа, которые мы (но не AI) будем распознавать как коды, соответствующие определенным нажатиям клавиш.

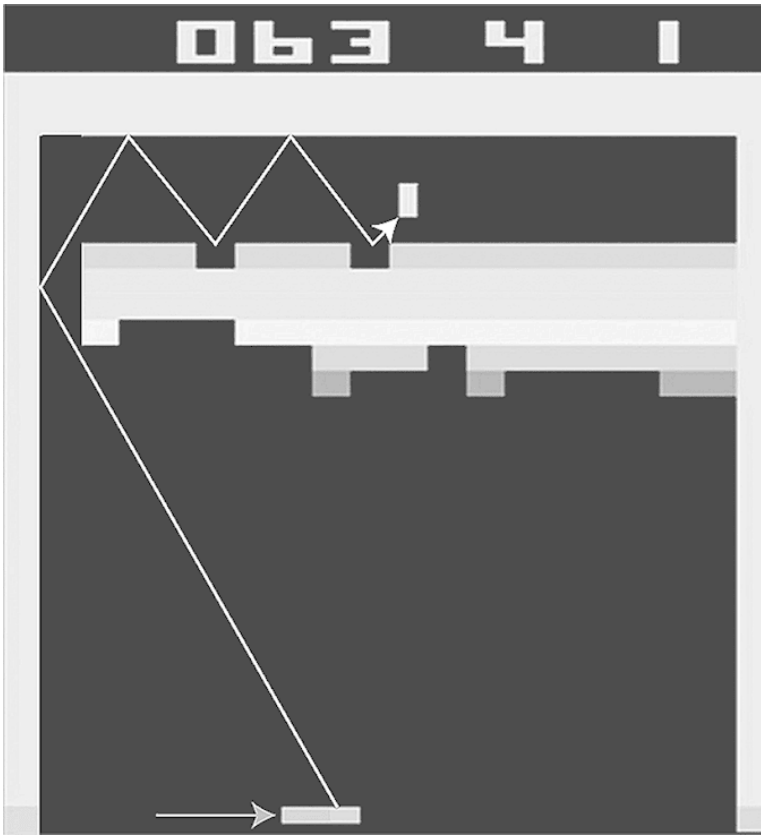


Рис. 3.1

Искусственный интеллект DeepMind учился проходить аркадную игру *Breakout* на платформе Atari с нуля, для чего использовались методы машинного обучения с подкреплением. Вскоре DeepMind самостоятельно открыл оптималь-

ную стратегию: пробивать в левом краю кирпичной стены дыру и загонять в эту дыру игровой шарик, который, оказавшись в замкнутом пространстве, быстро увеличивает счет. Я добавил на этом рисунке стрелки, показывающие траектории платформы и шарика.

Поначалу AI играл ужасно: он бессмысленно толкал платформу влево и вправо, как слепой, почти каждый раз промахиваясь мимо шарика. В какой-то момент у него, казалось, возникла идея, что двигать платформу по направлению к шарику – это, наверное, правильно, но шарик все равно пролетал мимо. Мастерство AI, однако, продолжало расти с практикой, и вскоре он стал играть значительно лучше, чем я когда бы то ни было, безошибочно отбивая шарик, как бы быстро тот ни двигался. И тут-то и пришло время моей челюсти отвиснуть: AI непостижимым образом смог раскрыть знакомую мне стратегию максимизации очков: всегда целиться в верхний левый угол, чтобы, пробив дырку в кирпичной кладке, загонять шарик туда, позволяя ему там долго прыгать между тыльной стороной стены и границей игрового поля. Это действительно казалось разумным решением. Позже Демис Хассабис говорил мне, что программисты компании DeepMind не знали этого трюка, пока созданный ими искусственный интеллект не открыл им глаза. Я всем рекомендую посмотреть этот ролик, перейдя по ссылке, которую я здесь привожу[8].

В том, как все это делалось, было что-то до такой степени человеческое, что мне стало не по себе: я видел AI, у которого была цель и который достиг совершенства на пути к ней, значительно обогнав своих создателей. В предыдущей главе мы определили интеллект просто как способность достигать сложных целей, и в этом смысле AI DeepMind становился все более умным в моих глазах (хотя бы и в очень узком смысле освоения премудростей единственной игры). В первой главе мы уже встречались с тем, что специалисты по информатике называют интеллектуальными агентами: это сущности, которые собирают информацию об окружающей среде от датчиков, а затем обрабатывают эту информацию, чтобы решить, как действовать в этой среде. Хотя игровой искусственный интеллект DeepMind жил в чрезвычайно простом виртуальном мире, состоящем из кирпичей, шариков и платформы, я не мог отрицать, что этот агент был разумным.

DeepMind вскоре опубликовала и свой метод, и использованный код, объяснив, что в основе лежала очень простая, но действенная идея, получившая название *глубокого обучения с подкреплением*[9]. Обучение с подкреплением – классический метод машинного обучения, основанный на бихевиористской психологии, которая утверждает, что достижение положительного результата подкрепляет ваше стремление повторить выполненное действие, и наоборот. Словно собака, которая учится выполнять команды хозяина, опираясь на его поддержку и в надежде на угощение, искусствен-

ный интеллект DeepMind учился двигать платформу, ловя шарик, в надежде на увеличение счета. DeepMind объединила эту идею с глубоким обучением: там научили глубокую нейронную сеть, описанную в предыдущей главе, предсказывать, сколько очков в среднем заработает AI, нажимая ту или иную из доступных клавиш, и, исходя из этого и учитывая текущее состояние игры, он выбирал ту клавишу, которую нейронная сеть оценивала как наиболее перспективную.

Рассказывая о том, что поддерживает мою положительную самооценку, я включил в этот список и способность решать разнообразные не решенные до меня задачи. Интеллект, ограниченный лишь способностью научиться хорошо играть в Breakout и больше ни на что не годный, следует считать чрезвычайно узким. Для меня вся важность прорыва DeepMind заключалась в том, что глубокое обучение с подкреплением – исключительно универсальный метод. Нет сомнений, что они практиковали его же, когда их AI учился играть в сорок девять различных игр Atari и достиг уровня, при котором стал уверенно обыгрывать любых человеческих соперников в двадцать девять из них, от Pong до Boxing, Video Pinball и Space Invaders.

Не надо было долго ждать момента, когда эту идею начнут использовать для обучения AI более современным играм – с трехмерными, а не двухмерными мирами. Вскоре конкурент компании DeepMind, базирующийся в Сан-Франциско OpenAI, выпустил платформу под названием Universe, где

DeepMind AI и другие интеллектуальные агенты могли совершенствоваться во взаимодействии с компьютером так же, как если бы это была игра, – орудуя мышкой, набирая что угодно на клавиатуре, открывая любое программное обеспечение, например запуская веб-браузер и роясь в интернете.

Охватывая взглядом будущее углубленного обучения с подкреплением, трудно предсказать, к чему оно может привести. Возможности метода явно не ограничиваются виртуальным миром компьютерных игр, поскольку, если вы робот, сама жизнь может рассматриваться как игра. Стюарт Рассел рассказывал мне о своем первом настоящем HS-моменте, когда он наблюдал, как его робот Big Dog поднимается по заснеженному лесному склону, изящно решая проблему координации движений конечностей, которую он сам не мог решить в течение многих лет[10]. Для прохождения этого эпохального этапа в 2008 году потребовались усилия огромного количества первоклассных программистов. После описанного прорыва DeepMind не осталось причин, по которым робот не может рано или поздно воспользоваться каким-нибудь вариантом глубокого обучения с подкреплением, чтобы самостоятельно научиться ходить, без помощи людей-программистов: все, что для этого необходимо, – это система, начисляющая ему очки при достижении успеха. Роботы в реальном мире также без помощи людей-программистов могут научиться плавать, летать, играть в настольный теннис, драться и делать все остальное из почти бесконечного списка

других двигательных задач. Для ускорения процесса и снижения риска где-нибудь застрять или повредить себя в процессе обучения прохождение его начальных этапов будет, вероятно, осуществляться в виртуальной реальности.

Интуиция, творчество, стратегия

Еще одним поворотным моментом для меня стала победа созданного DeepMind искусственного интеллекта AlphaGo в матче из пяти партий в го против Ли Седоля, который на начало XXI века считался лучшим игроком в го в мире.

Тогда все ждали, что людей вот-вот лишат звания лучших игроков в го, как это случилось с шахматами десятилетиями раньше. И только настоящие знатоки го предсказывали, что на это потребуется еще одно десятилетие, и поэтому победа AlphaGo стала поворотным моментом для них так же, как и для меня. Ник Бострём и Рэй Курцвейл оба подчеркнули, что этот прорыв AI было очень трудно предвидеть, о чем свидетельствуют, в частности, интервью самого Ли Седоля до и после проигрыша в первых трех играх:

> Октябрь 2015: "Оценивая нынешний уровень машины… я думаю, что выиграю почти все партии".
>
> Февраль 2016 года: "Я слышал, что Google DeepMind AI стал на удивление силен и быстро учится, но я убежден, что смогу выиграть хотя бы в этот раз".
>
> 9 марта 2016 года: "Я был очень удивлен, так как

совсем не ожидал, что могу проиграть".

10 марта 2016 года: "У меня нет слов… Я просто в шоке. Должен признать… что третья игра будет для меня нелегкой".

12 марта 2016 года: "Я чувствовал свое бессилие".

В течение года после победы над Ли Седолем улучшенный вариант AlphaGo обыграл двадцать лучших игроков в го в мире, не проиграв ни одной партии.

Почему все это воспринималось мной так лично? Я признавался выше, что считаю интуицию и способность к творчеству основными своими человеческими качествами, и, как я сейчас понимаю, в тот момент я почувствовал, что AlphaGo обладает обоими.

Играющие в го по очереди ставят черные и белые камни на доске 19 на 19 (см. рис. 3.2). Возможных позиций в го больше, чем атомов в нашей Вселенной, а это означает, что просчитать все интересные последствия каждого хода – дело безнадежное. Поэтому игроки в значительной степени полагаются на подсознательную интуицию, которая дополняет их сознательные рассуждения в оценке сильных и слабых сторон той или иной позиции, и у экспертов эта интуиция развивается в почти сверхъестественное чувство. Как мы видели в предыдущей главе, в результате глубокого обучения иногда возникает нечто напоминающее интуицию: глубокая нейронная сеть может определить, что на картинке изображена кошка, не имея возможности объяснить почему. По-

этому команда DeepMind поставила на идею, что глубокое обучение может распознавать не только кошек, но и сильные позиции в го. Главное, к чему они стремились, создавая AlphaGo, – было поженить интуицию, присущую глубокому обучению, с логической силой классического GOFAI[16], каков он был до революции глубокого обучения. Они взяли обширную базу данных, где было много позиций го как из игр, сыгранных людьми, так и из игр, сыгранных AlphaGo с клоном самого себя, и тренировали глубокую нейронную сеть предсказывать для каждой позиции вероятность итоговой победы белых. Кроме того, они натренировали отдельную сеть предсказывать вероятные следующие ходы. Затем они объединили эти две сети, пользуясь "старыми добрыми методами" для быстрого просмотра сокращенного списка наиболее вероятных будущих позиций, чтобы определить следующий ход, для которого следующая позиция окажется самой сильной.

[16] Широко используемая аббревиатура от Good Old-Fashioned Artificial Intelligence, что означает "старый добрый искусственный интеллект". – *Прим. перев.*

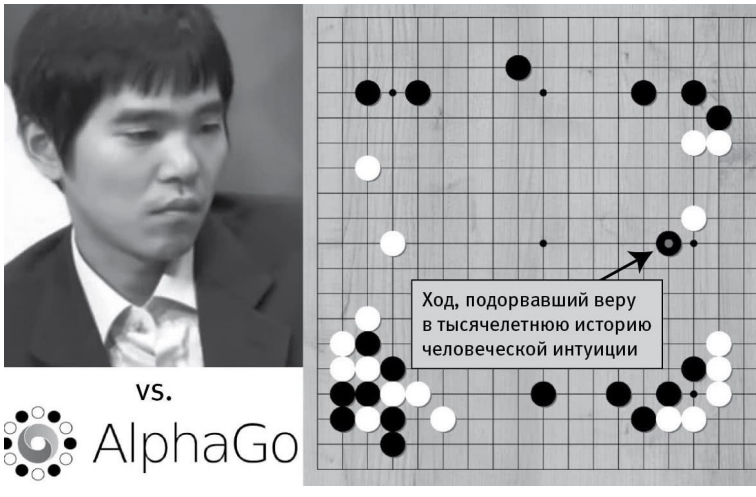


Рис. 3.2

Продолжение DeepMind – искусственный интеллект AlphaGo. Пренебрегая тысячелетним человеческим опытом игры в го, он сделал невероятно творческий ход на пятой линии, вся сила которого обнаружилась только 50 ходов спустя, в результате у легенды го Ли Седоля не оставалось никаких шансов.

Детьми, появившимися в браке интуиции и логики, оказались ходы, которые были не просто сильными, – в некоторых случаях их с полным основанием можно назвать креативными. Например, тысячелетняя мудрость го учит, что в начале игры надо стремиться захватить третью и четвертую линии от края. Тут есть возможность для торга: игра на третьей ли-

нии дает возможность быстро проводить краткосрочные захваты территории на краю доски, в то время как игра на четвертой линии способствует долгосрочному стратегическому влиянию на центр.

На тридцать седьмом ходу второй партии AlphaGo потряс мир го, пойдя наперекор этой древней мудрости и начав играть на пятой линии (рис. 3.2), словно он больше доверял своей способности долгосрочного планирования, чем человек, и поэтому отдавал предпочтение стратегическому преимуществу, а не краткосрочной выгоде. Комментаторы были ошеломлены, Ли Седоль даже поднялся и на какое-то время покинул помещение, где шла игра[11]. Они продолжали играть еще достаточно долго, было сделано еще примерно пятьдесят ходов, и только после этого основные события из нижнего левого угла доски переместились в центр, достигнув того самого камня, поставленного на тридцать седьмом ходу! И его присутствие здесь в конце концов сделало всю игру, навсегда внеся вторжение AlphaGo на пятую линию в анналы истории го как одно из самых важных открытий.

Именно из-за того, что игра в го требует интуиции и творчества, многие считают го в бóльшей степени искусством, чем просто игрой. В Древнем Китае умение играть в го считалось одним из четырех "основных искусств" наряду с живописью, каллиграфией и игрой на цине[17], и оно остается

[17] Цинь (#) – общее название различных китайских музыкальных инструментов, из которых наиболее популярны семиструнная цитра (###) и ее более ран-

чрезвычайно популярным в Азии: за первой партией между AlphaGo и Ли Седолем следили почти 300 миллионов человек. Результат матча глубоко потряс мир го, и победа AlphaGo стала для него важнейшей исторической вехой. Кэ Цзиэ, обладатель самого высокого рейтинга по го в то время, так прокомментировал это событие: “Человечество играло в го тысячи лет, и все же, как нам показал искусственный интеллект, мы всего лишь поцарапали его поверхность… Союз игроков-людей и игровых компьютеров открывает новую эру… Человек и искусственный интеллект смогут найти истину го вместе”. Плодотворное сотрудничество между человеком и машиной, и в самом деле, представляется очень многообещающим во многих сферах, включая науку, где искусственный интеллект, надеюсь, поможет нам, людям, углубить наше понимание мира и в значительно большей мере реализовать наш потенциал.

В конце 2017 года команда DeepMind запустила следующую модель – AlphaZero. Человеческому искусству игры в го тысячи лет, были сыграны миллионы партий, но все они не понадобились AlphaZero, которая училась с нуля, играя сама с собой. Она не только разгромила AlphaGo, но и стала сильнейшим в мире игроком в шахматы – и это тоже исключительно играя сама с собой. После двух часов практики она могла победить любого шахматиста-человека, а через четыре – обыграла Stockfish, лучшую в мире шахматную

ний вариант гуцинь (##). – *Прим. перев.*

программу. Меня тут особенно впечатляет не только то, что она била любого человека-шахматиста, но и то, что она обошла любого человека, занимающегося программированием искусственного интеллекта, она сделала устаревшим весь созданный людьми AI-софт, который разрабатывался несколько десятилетий. Иначе говоря, мы теперь не можем отмахнуться от идеи, что искусственный интеллект создает лучший искусственный интеллект.

Урок, преподанный нам AlphaGo, для меня состоял еще и в другом: объединение интуиции глубокого обучения с логикой "старого доброго искусственного интеллекта" может создавать стратегии на грани возможного. Поскольку го – одна из самых сложных стратегических игр, AI-системы должны теперь использоваться для того, чтобы оценивать способности и развивать их у лучших стратегов среди людей, проявляющих себя далеко за пределами игровой доски. Например, речь можно вести об инвестиционной стратегии, стратегии во внешней политике или военных операциях. Решение стратегических задач в перечисленных областях реальной жизни, как правило, осложняется человеческой психологией, отсутствием информации и случайными факторами, но системы с искусственным интеллектом, успешно играющие в покер, уже продемонстрировали, что ни одна из этих проблем не может считаться непреодолимой.

Естественный язык

Есть еще одна сфера деятельности, где успехи искусственного интеллекта в последнее время потрясли меня. Это языки. Еще в раннем детстве я полюбил путешествовать, и мое любопытство в отношении других культур и других языков сыграло огромную роль в формировании моей идентичности. В нашей семье говорили по-шведски и по-английски, в школе я учил немецкий и испанский, в двух браках мне понадобилось изучать португальский и румынский, просто так, ради удовольствия, я изучал русский, французский и мандарин.

> Но с искусственным интеллектом тягаться мне оказывается не под силу, и после важного открытия 2016 года больше нет таких "приятных" мне языков, в которых я могу переводить с одного на другой лучше, чем система AI, созданная мозгом Google.

Я достаточно прозрачно выразился? Я действительно пытался это сказать:

> Но AI догоняет меня, и после крупного прорыва в 2016 году не осталось почти никаких языков, между которыми я могу переводить лучше, чем искусственный интеллект, разработанный командой Google Brain для Google-переводчика.

Я сначала перевел эту фразу на испанский и обратно, ис-

пользуя приложение, которое я установил на своем ноутбуке несколько лет назад. В 2016 году команда Google Brain обновила свою бесплатную услугу Google Translate, включив в нее использование рекурсивных глубоких нейронный сетей, и в сравнении со "старыми добрыми" системами GOFAI это оказалось принципиальным[12]:

> Но AI догонял меня, и после прорыва в 2016 году практически не осталось языков, которые могут перевестись лучше, чем система AI, разработанная командой Google Brain.

Как вы можете видеть, местоимение "Я" потерялось во время захода в испанский язык, что, к сожалению, изменило смысл предложения[18]. Близко, да мимо! Однако в защиту искусственного интеллекта от Google должен признать, что меня часто критикуют за пристрастие к избыточно длинным предложениям, которые трудно разобрать, и я выбрал для этого примера одно из самых замысловато закрученных. Типичные предложения часто переводятся безукоризненно. Появление этой системы вызвало в результате изрядный переполох, и сейчас к ее помощи прибегают сотни миллионов человек ежедневно. Кроме того, благодаря использованию глубокого обучения для развития систем преобразова-

[18] В полной мере воспроизвести в переводе игру с Google-переводчиком, описанную в книге автором, не удалось: при переводе в 2018 году авторского пассажа сначала на русский, потом на испанский, а затем снова на русский местоимение "Я" упорно вставало на свое место и было устранено насильственным путем. – *Прим. перев.*

ния речи в текст или текста в речь их пользователи теперь могут проговаривать текст своему смартфону на одном языке и выслушивать его перевод на другой.

Конец ознакомительного фрагмента.

Комментарии

1.

Открытое письмо о дружественном и надежном искусственном интеллекте: http://futureoflife.org/ai-open-letter/

2.

Пример типичного алармизма по отношению к роботам в широкой прессе: http://tinyurl.com/hawkingbots

3.

Замечание по поводу происхождения термина AGI см.: http://goertzel.org/who-coined-the-term-agi/

4.

Hans Moravec 1998, “When will computer hardware match the human brain”, Journal of Evolution and Technology, vol. 1.

5.

Данные о стоимости вычислений в разные годы до 2011 взяты из книги Рэя Курцвейла How to Create a Mind, последующие данные вычислены на основании информации, приведенной в: https://en.wikipedia.org/wiki/FLOPS

6.

Один из создателей теории квантовых вычислений Дэвид

Дойч показывает, каким образом квантовые вычисления связаны с многомировой интерпретацией квантовой механики, в книге: David Deutsch. The fabric of reality. London: Penguin, 1997 (есть русский перевод: Дойч Д. Структура реальности. Наука параллельных вселенных. М.: Альпина нон-фикшн, 2015. – Прим. перев.). Если вас интересует мой собственный подход к квантовым параллельным вселенным как к третьему из четырех уровней мультиверсума, то смотрите мою книгу Our Mathematical Universe (см. рус. пер.: Тегмарк М. Наша математическая Вселенная. В поисках фундаментальной природы реальности. М.: Corpus, 2016 / пер. с англ. А. Сергеев. – Прим. перев.).

7.

О прорыве, совершенном Google в распознавании образов, см.: https://arxiv.org/pdf/1411.4555.pdf

8.

DeepMind алгоритм глубокого машинного обучения с подкреплением позволил довольно быстро научиться играть в Breakout: https://tinyurl.com/atariai

9.

См. статью, в которой описывается искусственный интеллект DeepMind, совершенствующийся в играх на платформе

Atari: http://tinyurl.com/ataripaper

10.
Робот Биг Дог в действии: https://www.youtube.com/watch?v=W1czBcnX1Ww

11.
Запись реакции на революционный ход AlphaGo на 5-й линии: https://www.youtube.com/watch?v=JNrXgpSEEIE

12.
Статья в New York Times о недавних достижениях в машинном переводе: http://www.nytimes.com/2016/12/14/magazine/the-great-aiawakening.html