



МАКС ТЕГМАРК

ЖИЗНЬ 3.0

Быть человеком
в эпоху искусственного
интеллекта



СЕРИЯ Э | Л | Е | М | Е | Н | Т | Ы

Annotation

“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.

- [Макс Тегмарк](#)
 -
 -
 -
 - [Прелюдия](#)
 -
 - [Первые миллионы](#)
 - [Опасные игры](#)
 - [Первые миллиарды](#)
 - [Новые технологии](#)
 - [Обретение власти](#)
 - [Консолидация](#)
 - [Глава 1](#)
 -
 - [Краткая история сложности](#)
 - [Три стадии жизни](#)
 - [Контроверзы](#)
 - [Недоразумения](#)
 - [Дорога вперед](#)
 - [Подведение итогов](#)
 - [Глава 2](#)
 -
 - [Что такое разум?](#)
 - [Что такое память?](#)
 - [Что такое вычисление?](#)
 - [Что такое обучение?](#)
 - [Глава 3](#)
 -
 - [Прорывы](#)
 - [Надежность искусственного интеллекта и глюки](#)
 - [Законы](#)
 - [Оружие](#)
 - [Работа и заработки](#)
 - [Интеллект на уровне человеческого](#)
 - [Глава 4](#)
 -
 - [Тоталитаризм](#)
 - [Прометей захватывает мир](#)
 - [Медленный старт и многополярные сценарии](#)
 - [Киборги и заливки. Эмы](#)
 - [Так что же все-таки произойдет?](#)

- [Глава 5](#)
 -
 - [Либертарианская утопия](#)
 - [Благодетельный диктатор](#)
 - [Эгалитарная утопия](#)
 - [Страж](#)
 - [Бог-защитник](#)
 - [Порабощенный бог](#)
 - [Победители](#)
 - [Благодарные потомки](#)
 - [Зоопарк](#)
 - [1984](#)
 - [Возврат](#)
 - [Самоистребление](#)
 - [А что хотите вы?](#)
 - [Подведение итогов](#)
- [Глава 6](#)
 -
 - [На что годятся ваши ресурсы](#)
 - [Как добывать ресурсы с помощью космической экспансии](#)
 - [Космические иерархии](#)
 - [Виды на будущее](#)
 - [Подведение итогов](#)
- [Глава 7](#)
 -
 - [Физика: происхождение целей](#)
 - [Биология: эволюция целей](#)
 - [Психология: Преследование и восстание против целей](#)
 - [Инженерия: привлечение сторонних целей](#)
 - [Дружественный искусственный интеллект: приведение целей в соответствие](#)
 - [Этика: выбор целей](#)
 - [Основные цели?](#)
 - [Подведение итогов](#)
- [Глава 8](#)
 -
 - [Кого заботит?](#)
 - [Что такое сознание?](#)
 - [В чем же проблема?](#)
 - [Сознание за пределами науки?](#)
 - [Что говорит о сознании эксперимент?](#)
 - [Теории сознания](#)
 - [Споры о сознании](#)
 - [Что может чувствовать искусственный интеллект, обладая сознанием?](#)
 - [Смысл](#)
 - [Подведение итогов](#)
- [Эпилог](#)
 -
 - [Рождение FLI](#)
 - [Пуэрто-риканское приключение](#)
 - [Перенос AI-безопасности в центр исследовательского поля](#)
 - [Асиломарские принципы разработки искусственного интеллекта](#)

- [Исследовательские вопросы](#)
- [Ценности и вопросы этики](#)
- [Долгосрочные аспекты](#)
- [Разумный оптимизм](#)

- [notes](#)

- [Сноски](#)

- [1](#)
 - [2](#)
 - [3](#)
 - [4](#)
 - [5](#)
 - [6](#)
 - [7](#)
 - [8](#)
 - [9](#)
 - [10](#)
 - [11](#)
 - [12](#)
 - [13](#)
 - [14](#)
 - [15](#)
 - [16](#)
 - [17](#)
 - [18](#)
 - [19](#)
 - [20](#)
 - [21](#)
 - [22](#)
 - [23](#)
 - [24](#)
 - [25](#)
 - [26](#)
 - [27](#)
 - [28](#)
 - [29](#)
 - [30](#)
 - [31](#)
 - [32](#)
 - [33](#)
 - [34](#)
 - [35](#)
 - [36](#)
 - [37](#)
 - [38](#)
 - [39](#)
 - [40](#)
 - [41](#)
 - [42](#)
 - [43](#)
 - [44](#)
 - [45](#)
 - [46](#)

- [47](#)
- [48](#)
- [49](#)
- [50](#)
- [51](#)
- [52](#)
- [53](#)
- [54](#)
- [55](#)
- [56](#)
- [57](#)
- [58](#)
- [59](#)
- [60](#)
- [61](#)
- [62](#)
- [63](#)

○ [Комментарии](#)

- [1](#)
- [2](#)
- [3](#)
- [4](#)
- [5](#)
- [6](#)
- [7](#)
- [8](#)
- [9](#)
- [10](#)
- [11](#)
- [12](#)
- [13](#)
- [14](#)
- [15](#)
- [16](#)
- [17](#)
- [18](#)
- [19](#)
- [20](#)
- [21](#)
- [22](#)
- [23](#)
- [24](#)
- [25](#)
- [26](#)
- [27](#)
- [28](#)
- [29](#)
- [30](#)
- [31](#)
- [32](#)
- [33](#)
- [34](#)

- [35](#)
- [36](#)
- [37](#)
- [38](#)
- [39](#)
- [40](#)
- [41](#)
- [42](#)
- [43](#)
- [44](#)
- [45](#)
- [46](#)
- [47](#)
- [48](#)
- [49](#)
- [50](#)
- [51](#)
- [52](#)
- [53](#)
- [54](#)
- [55](#)
- [56](#)
- [57](#)
- [58](#)
- [59](#)
- [60](#)
- [61](#)
- [62](#)
- [63](#)
- [64](#)
- [65](#)
- [66](#)
- [67](#)
- [68](#)
- [69](#)
- [70](#)
- [71](#)
- [72](#)
- [73](#)
- [74](#)
- [75](#)
- [76](#)
- [77](#)
- [78](#)
- [79](#)
- [80](#)
- [81](#)
- [82](#)
- [83](#)
- [84](#)
- [85](#)
- [86](#)

- [87](#)
- [88](#)
- [89](#)
- [90](#)
- [91](#)
- [92](#)
- [93](#)
- [94](#)
- [95](#)
- [96](#)
- [97](#)
- [98](#)
- [99](#)
- [100](#)
- [101](#)
- [102](#)
- [103](#)
- [104](#)
- [105](#)
- [106](#)
- [107](#)
- [108](#)
- [109](#)
- [110](#)
- [111](#)
- [112](#)
- [113](#)
- [114](#)
- [115](#)
- [116](#)
- [117](#)
- [118](#)
- [119](#)
- [120](#)
- [121](#)
- [122](#)
- [123](#)
- [124](#)
- [125](#)
- [126](#)
- [127](#)
- [128](#)
- [129](#)
- [130](#)
- [131](#)
- [132](#)
- [133](#)
- [134](#)
- [135](#)
- [136](#)
- [137](#)
- [138](#)

- [139](#)
 - [140](#)
 - [141](#)
 - [142](#)
-

Макс Тегмарк

Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта

© Max Tegmark, 2017

© Д. Баяк, перевод на русский язык, 2019

© А. Бондаренко, художественное оформление, макет, 2019

© ООО “Издательство Аст”, 2019

* * *



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 “Вымпелком” (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

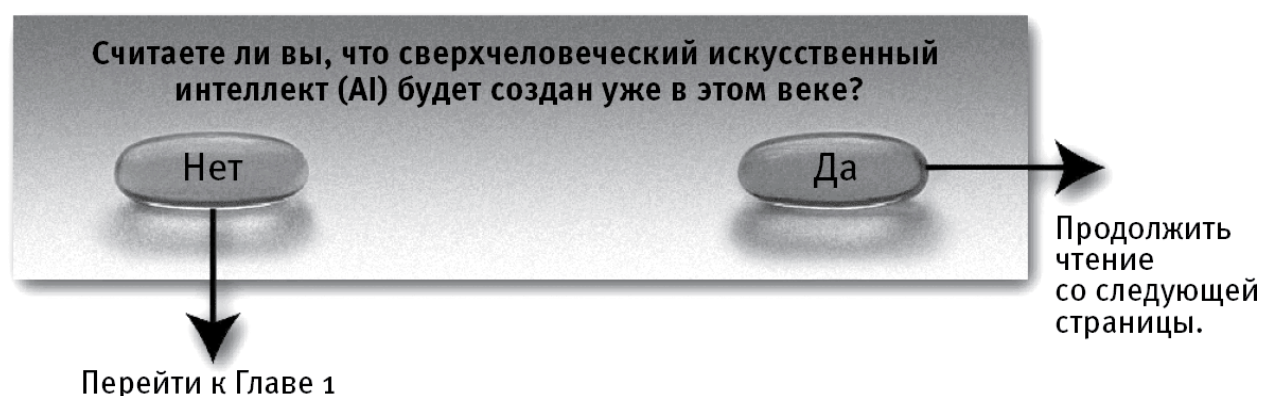
*Посвящается команде FLI.
Благодаря им все возможно*

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 устал 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 уик-энд. 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 быстрдействию мог уложить тысячи человеко-лет самого 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 Прометей 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 Прометей 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 половины 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 Прометей 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 Прометей, 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 возможности 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 Прометей 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 максимальную 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 Прометей 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 называли “Ящиком Пандоры” и который симулировал простейший тип компьютера, лишенный всех привычных нам прибамбасов и погрешек: у него не было ни клавиатуры, ни монитора, ни динамиков, ни интернет-подключения, ничего такого. Для транскрибирования аудиофайлов для “механического турка” “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. Ничего 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 угодно, 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 прицелом: 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 вырученный 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 гэггами 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 месяцев. 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 воздерживаться 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 Прометеев 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 своей работы Прометей предоставил “Омге” 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 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 электроэнергетике. 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”. Прометей 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 принимали 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 сетях 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, чтобы 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 неограничен 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 сногшибательным: 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, который, естественно, контролировался “Омегой”.

Примерно через два года после запуска Прометей 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 разворачивалось на фоне ядерной бомбардировки, начатой по ошибке или намеренно, с ужасающими картинками 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 Прометей 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 соответствующие 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 Прометей найдется для тебя подходящий видеокурс. И он будет совсем не похож на те, которые доступны нам сейчас: благодаря растущему таланту Прометей в создании сериалов видеозаписи 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 роликов 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 бывших 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. Традиционные 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 такими 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 разнообразие 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 Прометей 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, нацеленную 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 значительно 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 скак 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 было. 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.

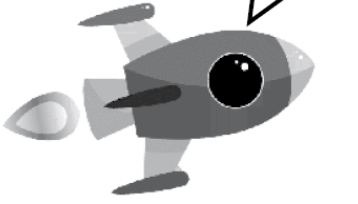
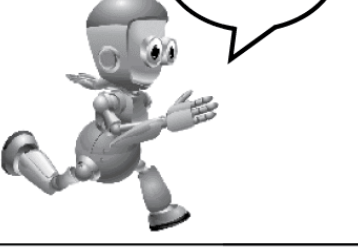
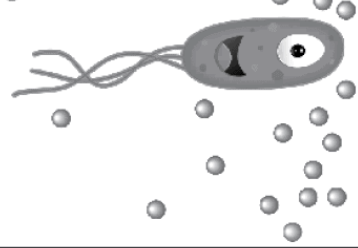
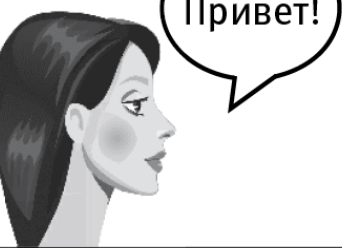
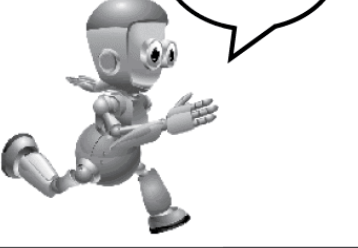
Способна ли проектировать свой «хард»?			 
Способна ли проектировать свой «софт»?		 	 
Способна ли приспосабливаться и размножаться?	 	 	 
	Жизнь 1.0 (простая биологическая)	Жизнь 2.0 (культурная)	Жизнь 3.0 (технологическая)

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 этого, 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 большего ума требуется больший “хард” (больше атомов) и больший “софт” (больше бит). Тот факт, что большая 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 большая 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 “вешают”, 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 триллионы 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:

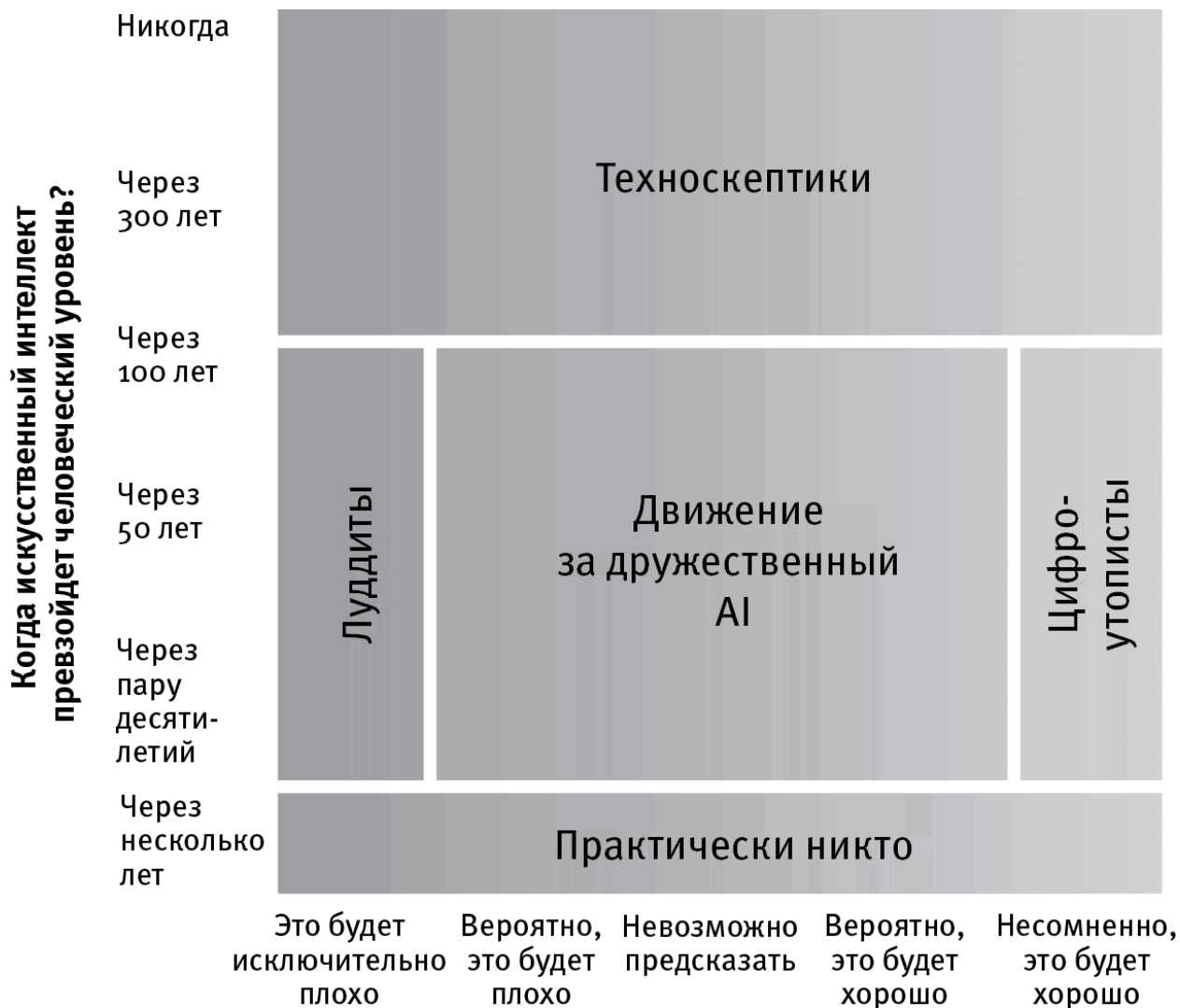
1. Жизнь 1.0 (биологическая стадия): эволюция “харда” и “софта”;
2. Жизнь 2.0 (культурная стадия): эволюция “харда” и дизайн большей части “софта”;
3. Жизнь 3.0 (технологическая стадия): дизайн и “софта”, и “харда”.

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.



Если сверхчеловеческий искусственный интеллект возникнет, будет ли это хорошо?

Рис. 1.2.

Большинство споров вокруг сильного искусственного интеллекта (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).

Ребенком я был уверен, что все миллиардеры просто сочатся помпезностью и невежеством. Когда в 2008 году в Google я впервые встретил Ларри Пейджа, он напроць 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 пытаюсь 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 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, 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 соглашались 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, любимого мною в детстве героя классического романа Жюль 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 Ян Таллин. Наша 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: создаваемый интеллект не должен быть неконтролируемым, он должен быть дружелюбным. В письме формулировался подробный список исследовательских задач, вокруг которых участники конференции соглашались концентрировать свою работу. Дружелюбный 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, Катя 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, Оуейн 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 моральное 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 оставалась 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, находите 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 контroversы, о которых я уже упоминал и в которых величайшие мировые умы придерживаются противоположных позиций. Но кроме этого я был свидетелем множества скучнейших псевдо-контrovers, когда люди просто не понимают или даже не слушают друг друга. Для того чтобы сконцентрироваться на важных нерешенных проблемах, говорить о контroversах, 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 интеллектом 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 сверхума
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 употребляются 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, Натаниелю 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”.

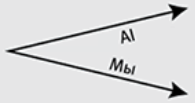
Миф Искусственный сверхинтеллект к 2100 году неотвратим Миф Искусственный сверхинтеллект к 2100 году немислим		Факт Это может случиться через несколько десятилетий, несколько веков, а может не случиться никогда. У экспертов нет согласия по этому поводу, так что мы просто не знаем ответа на этот вопрос	
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 мутой”. 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 экспертных 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 противоречили 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, заиклившись 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 разумом 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.

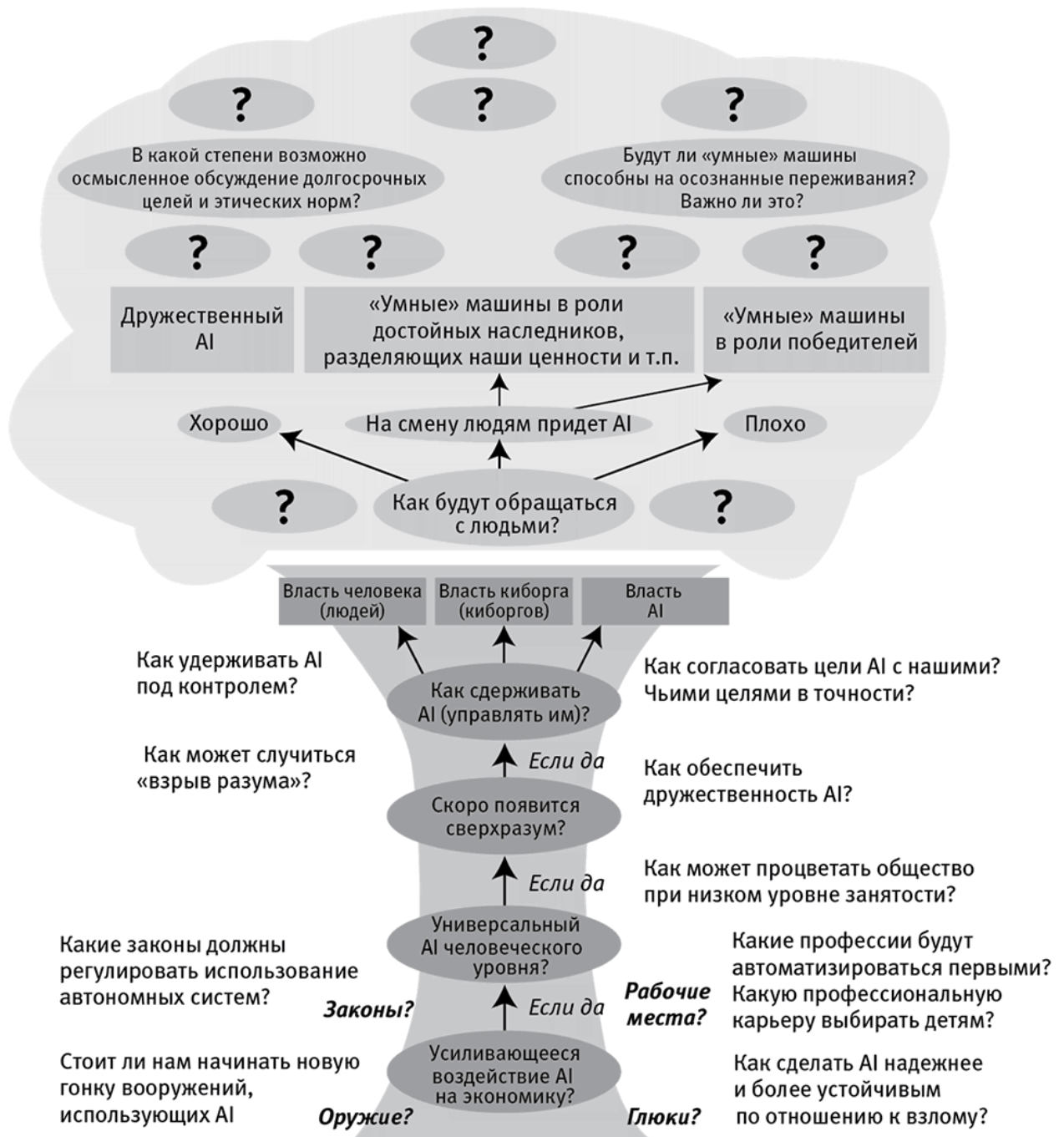


Рис. 1.6

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
История разума		Пролог: Сказание о команде “Омега”	Пища к размышлению	Исключительно спекулятивно
	1	Самый важный разговор	Ключевые идеи, терминология	Не очень спекулятивно
	2	Материя начинает думать	Основы теории разума	
	3	AI, экономика, оружие и закон	Ближайшее будущее	
	4	Взрыв разума?	Сценарии сверхразума	Исключительно спекулятивно
	5	Что потом?	Мир в следующие 10 тысяч лет	
	6	Наше космическое благосостояние	Мир в следующие миллиарды лет	
История смысла	7	Цели	История целенаправленного поведения	Не очень спекулятивно
	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 дружелюбным? 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 унести 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 вопросы 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 самостоятельными, 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 потребуются 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 неприятного 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 сильнее, 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].

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.

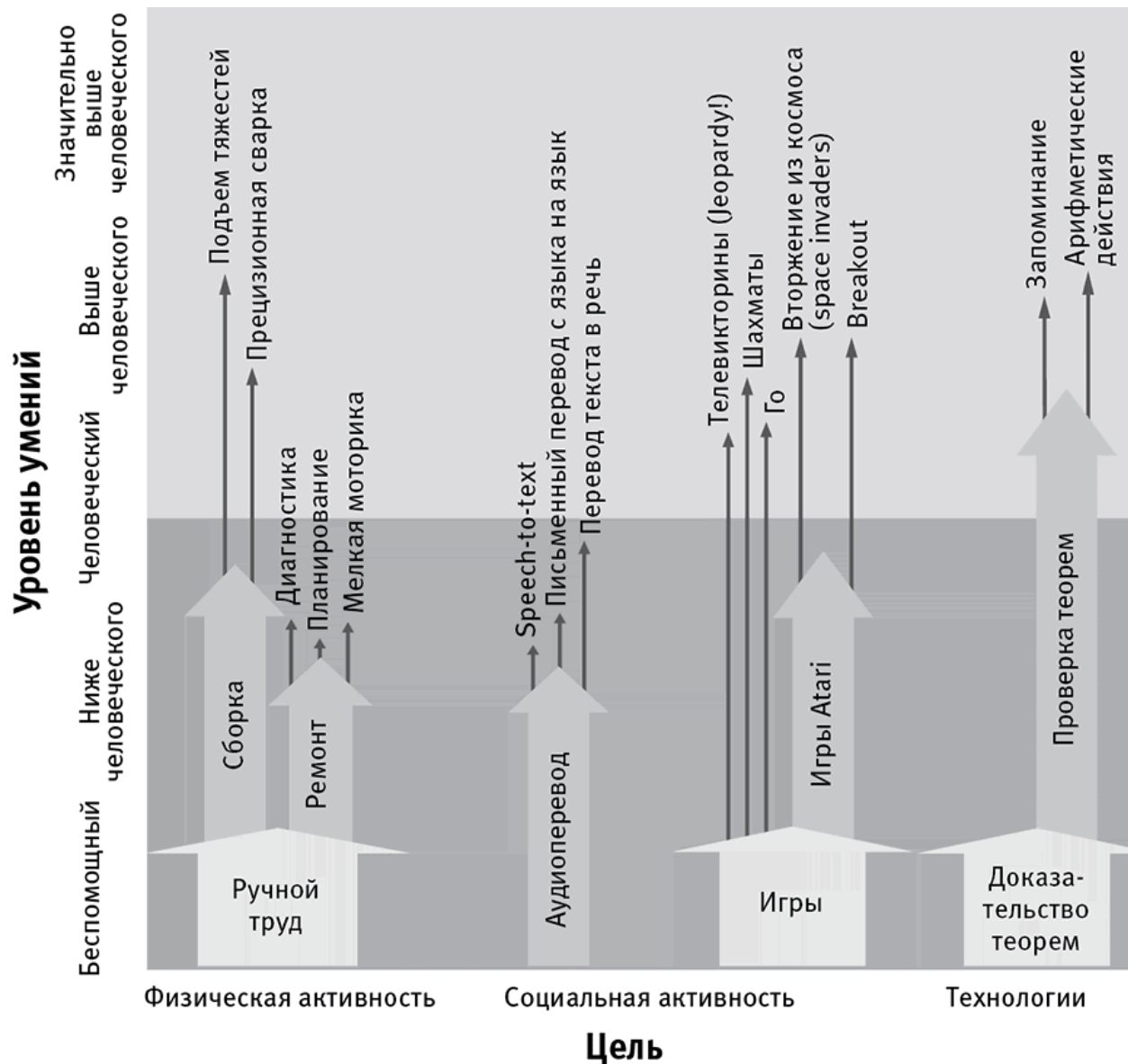


Рис. 2.1

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 определенную 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 которой 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 потребует. В качестве простого примера представьте себе холмистую местность с шестнадцатью отделенными одна 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 выгравироваете 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 выгравироваете его же на поверхности пруда, информация пропадет за секунду, потому что поверхность воды изменяется практически без энергетических затрат.

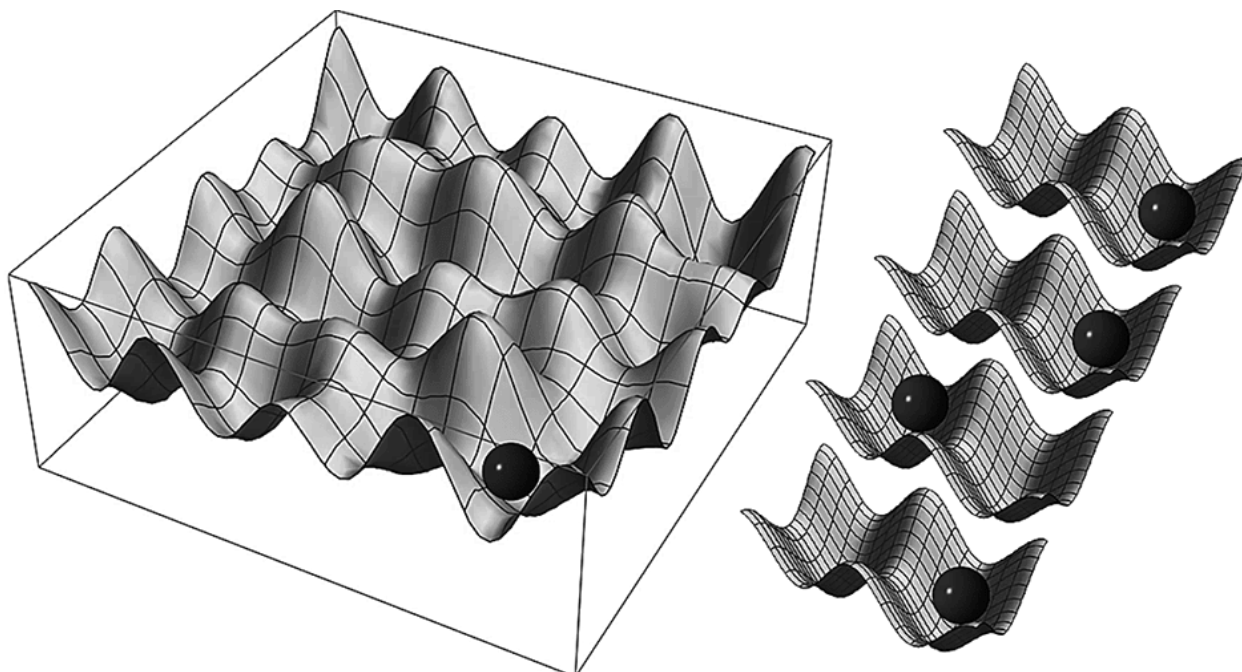


Рис. 2.3

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, сохраненная 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 диске 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 случаев 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 получить такую скидку в 99,9999999999 % 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.

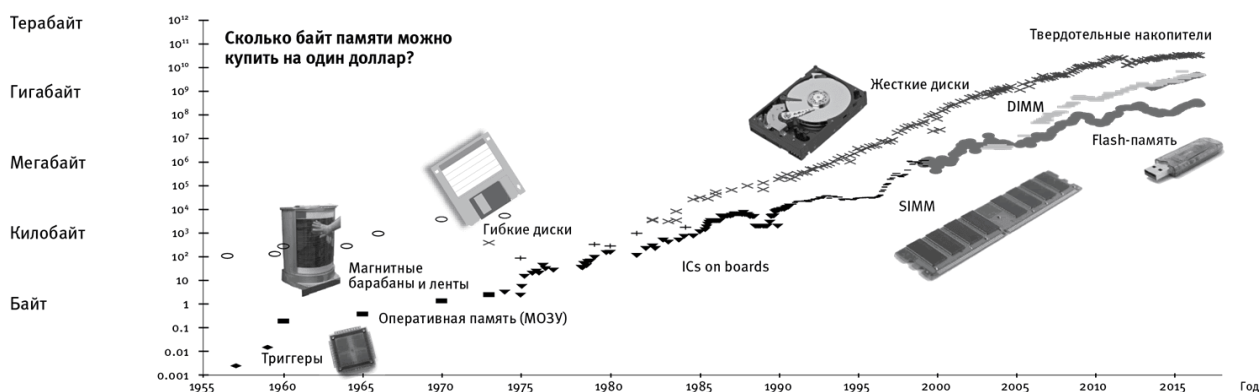


Рис. 2.4

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 МакКаллемом (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 компьютера, 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 природе, – 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 книжному 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 минимума 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 кнопкам 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).

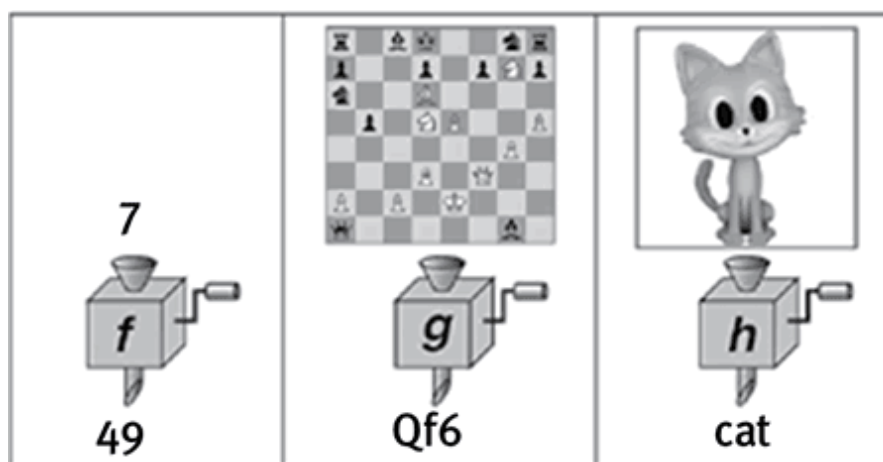


Рис. 2.5

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 целей. Таким образом, нам удастся 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] 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 случаях – 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 называли гейтом NAND: третий ключ будет разомкнут только тогда, когда первые два замкнуты. Есть очень много более практических 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.

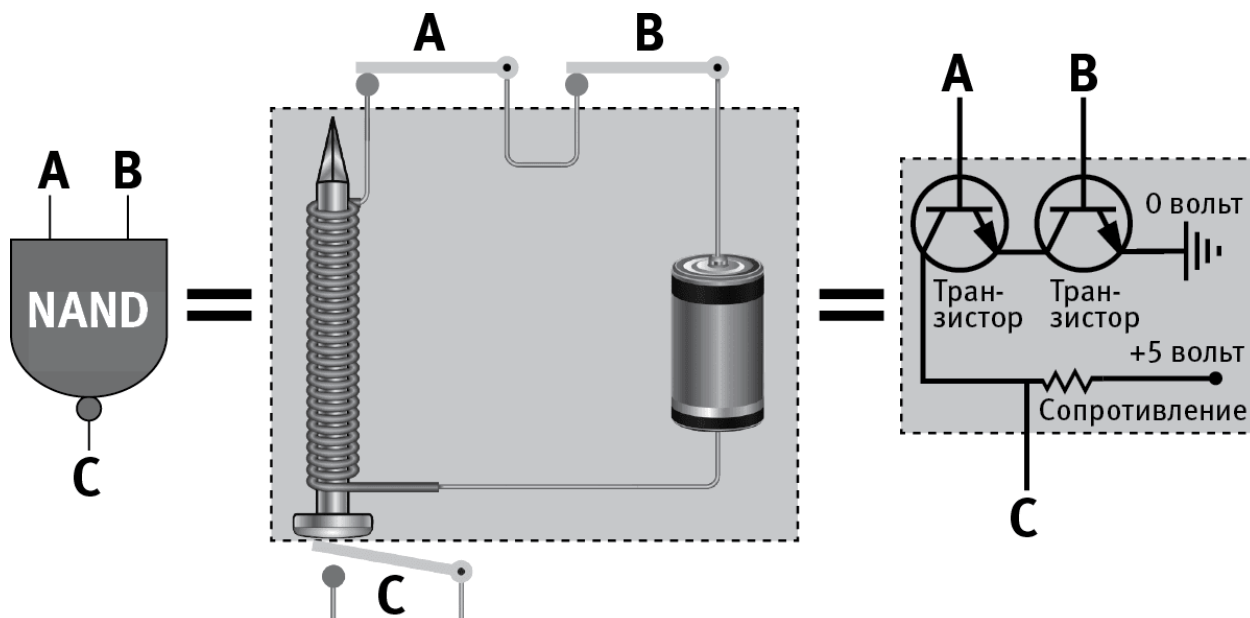


Рис. 2.6

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 $C = 1$ 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 вольт. 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 компьютерониум 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 компьютерониумов. 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 обсудим нейронные сети, которые также способны выполнять произвольные вычисления, то есть и они ведут себя как компьютерониум. 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.

Все можно построить только из вентилях (гейтов) NAND

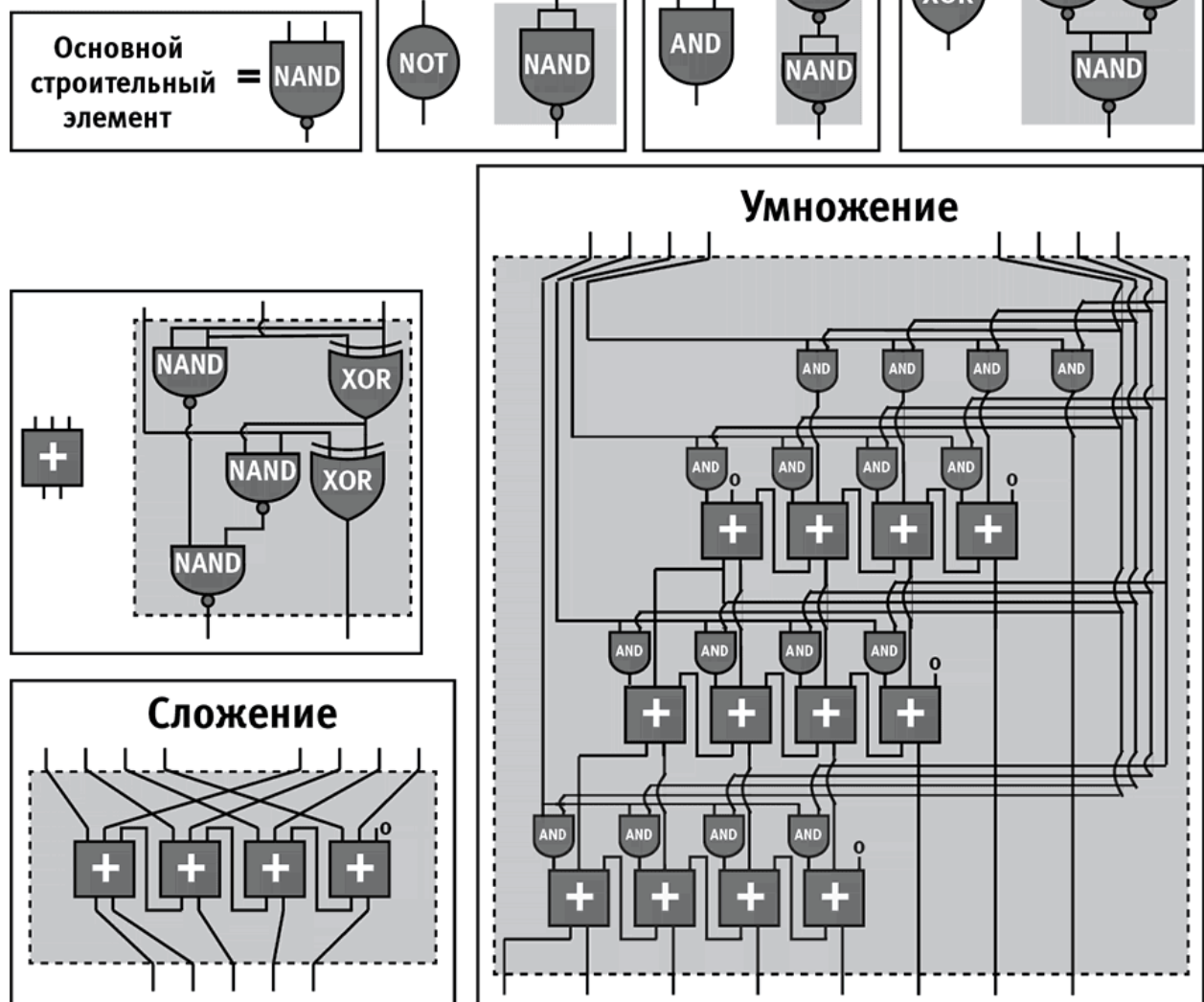


Рис. 2.7

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 потакаю своему внутреннему гикству.

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 отвлечься 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!

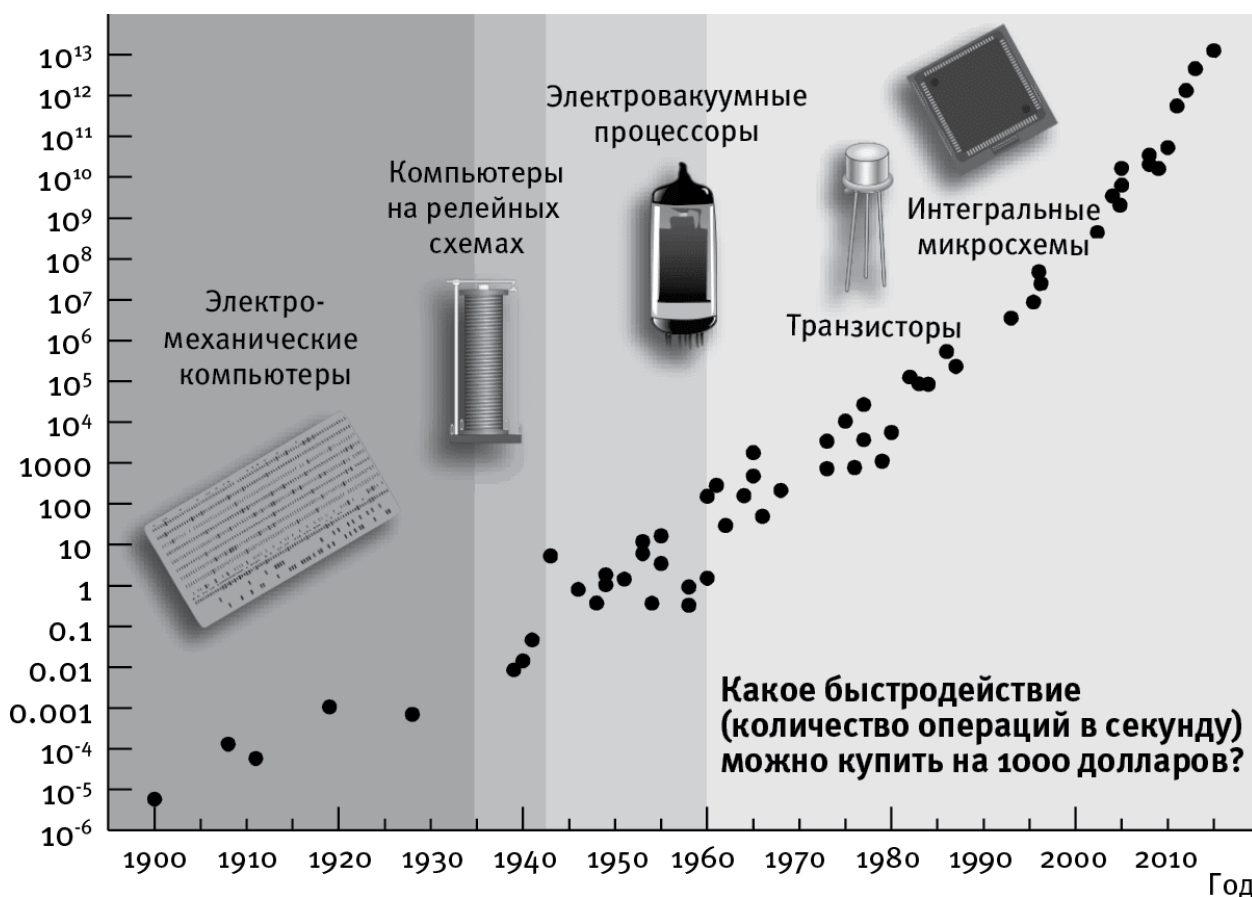
Этот пример с волновым уравнением позволяет сделать три вывода. Во-первых, независимость от субстрата еще не означает, что без субстрата можно обойтись, но только лишь – что многие подробности его устройства не важны. Вы не услышите никакого звука в безвоздушном пространстве, но если замените воздух 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 купить 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.



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 удваивать 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 клеток, 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 разработана 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 удастся 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 серьезные 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 скорости 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.

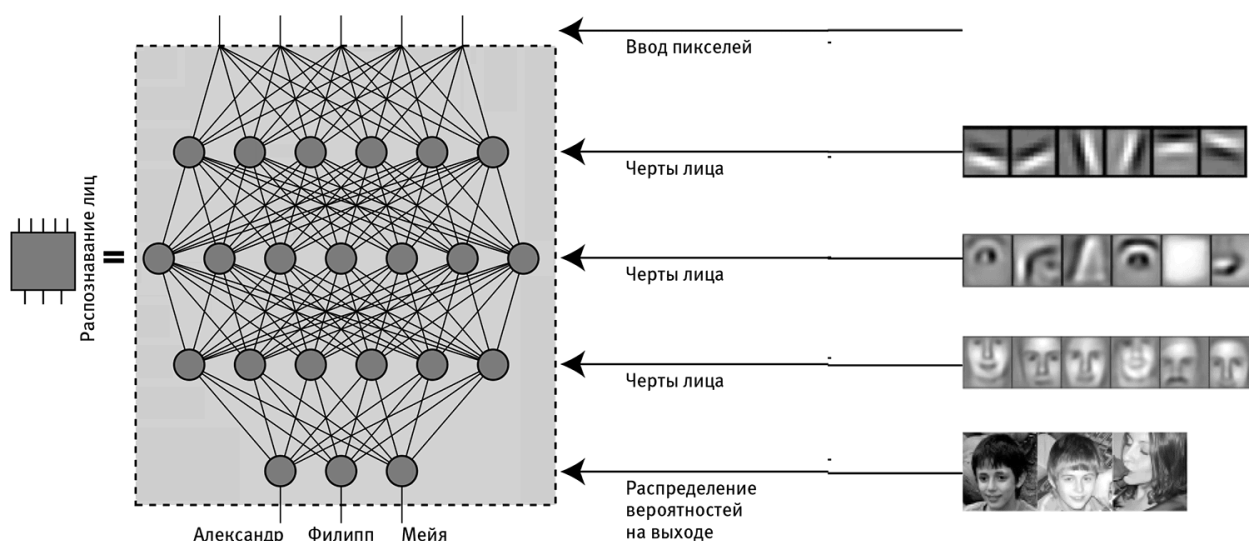


Рис. 2.9

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 связи. 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 произвольной 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.

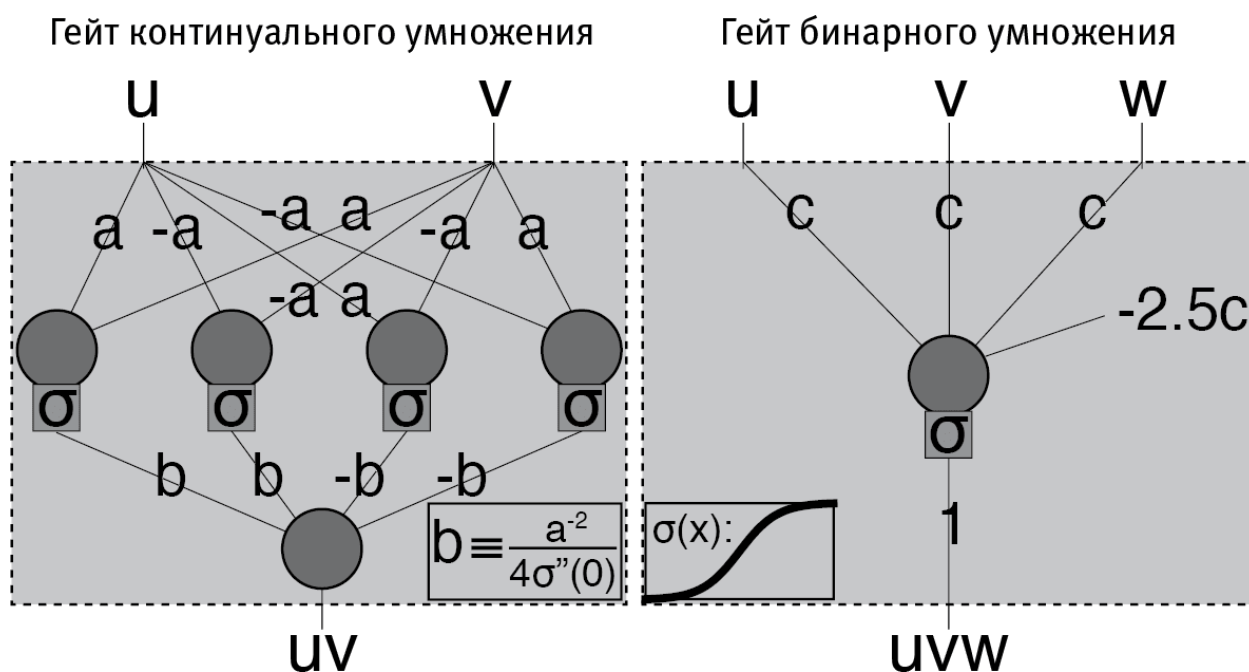


Рис. 2.10

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 производить 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 $a \rightarrow 0$ (слева) и при $c \rightarrow \infty$ (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 аппроксимацию 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 работает, 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 параметров? 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 Шреверпорт 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 ссылаясь 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. Слово 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 позволяет 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 способности 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 быстрой реакции. 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 обладавшую 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 дети, 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 победила 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 из 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 *компьютериумом*, 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, людей, 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 невозможно 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, сжавившейся 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, заставляя 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 представляющих 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.

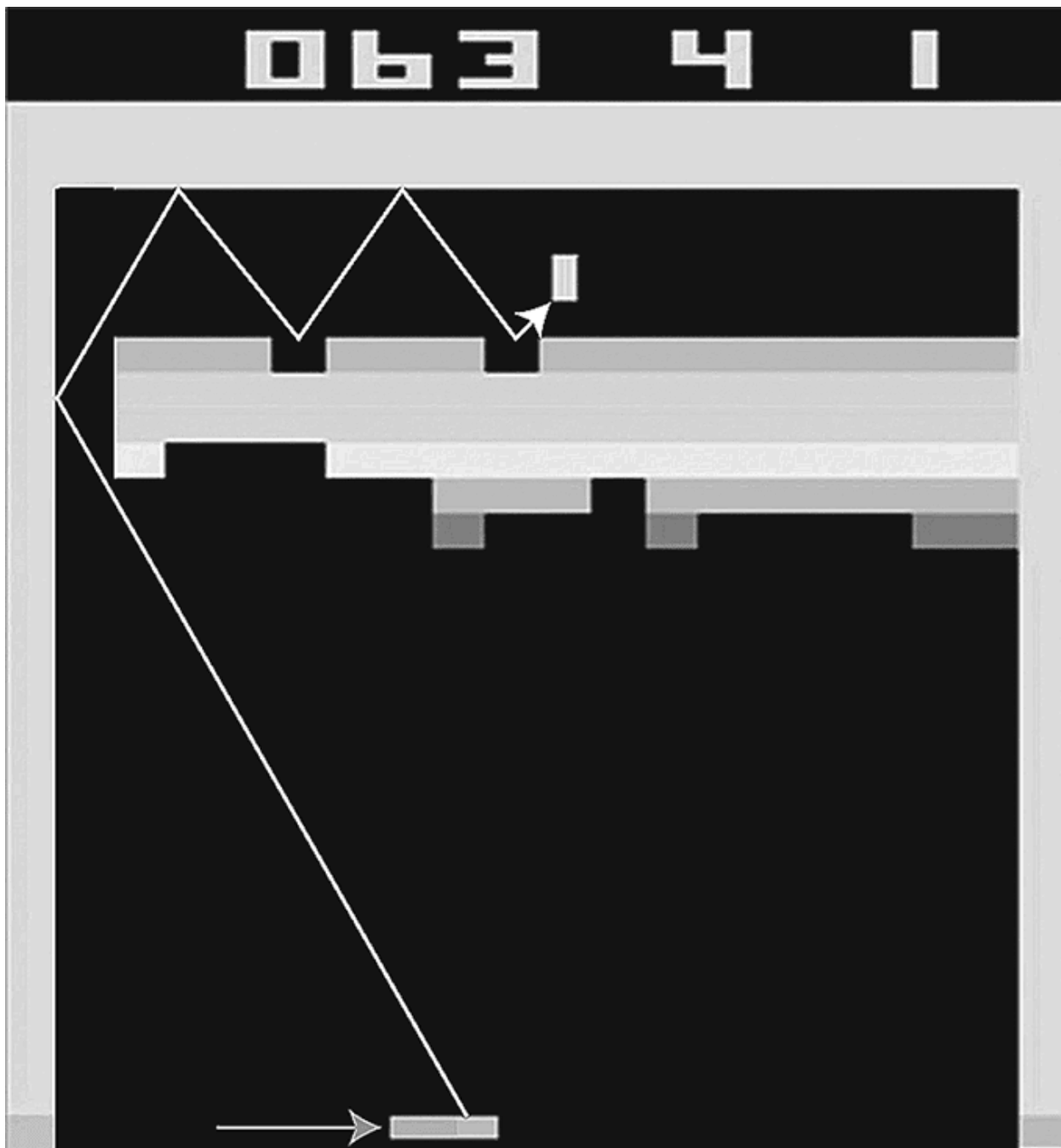


Рис. 3.1

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 оптимальную 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 шарика – 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, искусственный 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 роаясь 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. Поэтому 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.



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.

Детями, 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 линии 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 внося 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 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 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, используя 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 нейронных 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 преобразования речи в текст или текста в речь их пользователи теперь могут проговаривать текст своему смартфону на одном языке и выслушивать его перевод на другой.

Преобразования текстов на естественных языках – сейчас одна из наиболее быстро развивающихся областей применения искусственного интеллекта, и я думаю, что ее успешное развитие повлечет важные последствия, поскольку именно благодаря языку человек становится человеком. Чем сильнее становится искусственный интеллект в лингвистических предсказаниях, тем точнее он сможет ответить на электронное письмо или поддержать беседу. Благодаря этому, по крайней мере, у постороннего может сложиться впечатление, что он общается с человеком. Системы глубокого обучения делают сейчас первые шаги к тому, чтобы пройти знаменитый тест Тьюринга, научившись достаточно хорошо отвечать на вопросы в письменной форме, создавая у задающего их человека впечатление, что отвечает ему тоже человек.

И все же в работе с языком у искусственного интеллекта впереди еще долгий путь. Я должен признаться, что хотя меня задевает, когда искусственный интеллект обходит меня в точности перевода, я напоминаю себе, что искусственный интеллект *совсем не понимает*, о чем говорится в переводимом им тексте, и от этого мне сразу становится лучше. Его натренировали на массивном объеме данных искать соответствующие грамматические конструкции в языках и устойчивые отношения между словами, но он не умеет обнаруживать связь этих слов с чем бы то ни было в реальном мире. Например, он

может представлять каждое слово в виде списка из тысяч чисел, показывающих, насколько оно близко по значению некоторым другим словам. Он может заключить, что разница между “королем” и “королевой” аналогична разнице между “мужем” и “женой”, но он все равно не знает, что значит быть мужчиной или женщиной, или даже что существует такая вещь, как физическая реальность, с пространством, временем и материей.

Тест Тьюринга, коль скоро в нем речь об обмане, не раз критиковали за то, что он проверяет скорее человеческое занудство, чем разумность компьютера. Конкурирующая система получила название Winograd Schema Challenge и была нацелена именно на то, чтобы выявить уровень общего здравомыслия, которого современным системам глубокого обучения как раз и недостает. Мы, люди, по старинке пользуемся нашим знанием о реальном мире, пытаюсь понять предложение, угадывая, к чему относится то или иное местоимение. Например, вот типичное задание Винограда: определить, к кому относится местоимение “они” в предложениях:

1. Члены городского совета отказали демонстрантам в разрешении, потому что они боялись проявлений насилия.
2. Члены городского совета отказали демонстрантам в разрешении, потому что они выступали в защиту насилия.

Каждый год между различными AI-системами проводятся соревнования, где ставятся такие вопросы, и пока эти системы показывают невысокую сообразительность^[13]. Именно эта задача – понимать, что относится к чему – сразила даже Google-переводчик, когда я в своем примере заменил испанский на китайский:

Но ИИ преследовал меня. После прорыва в 2016 году почти ни один язык не может быть переведен лучше, чем система искусственного интеллекта, разработанная командой Google Brain^[19].

Пожалуйста, сходите сами на <https://translate.google.com>, пока читаете эту книгу, и проверьте, не улучшился ли искусственный интеллект Google. Велика вероятность, что с ним это случилось, так как существуют весьма многообещающие подходы, как поженить рекуррентную глубокую нейронную сеть со “старым добрым искусственным интеллектом”, чтобы построить лингвистическую AI-систему, включающую в себя и модель мира.

Возможности и вызовы

Все эти три примера тут, разумеется, случайны, так как искусственный интеллект сейчас быстро распространяется на многие виды деятельности. Более того, хотя выше я упоминал только две компании в двух примерах, конкурирующие группы исследователей в университетах и из других компаний не сильно от них отстали. Громкий вой работающего пылесоса раздается во всех департаментах информатики по всему миру – он сопровождает старания Apple, Baidu, DeepMind, Facebook, Google, Microsoft соблазнительными предложениями отсосать оттуда студентов, аспирантов, профессоров.

Очень важно, чтобы приведенные мной выше примеры не создали неправильного представления, будто история AI состоит в основном из периодов стагнации, прерываемых редкими удачами. У меня привилегированное положение: я вижу длительный и неуклонный рост, который преподносится прессой как прорыв, свидетельствующий лишь о том, что преодолен очередной порог и стало возможным новое, поражающее воображение приложение либо полезный продукт. И поэтому мне

кажется совершенно ожидаемым длительный и непрерывный прогресс. Более того, как мы видели в предыдущей главе, нет никаких оснований думать, что этот прогресс не сможет длиться до тех пор, пока искусственный интеллект не достигнет человеческих способностей в большинстве дел.

Отсюда вопрос: а как это все повлияет на нас? Как новые возможности искусственного интеллекта изменят смысл человеческого существования? Мы уже видели, что становится все труднее отрицать способность искусственного интеллекта ставить цели, опираться на интуицию, проявлять креативность, понимать человеческий язык, – то есть обладать качествами, которые многие из нас считают определяющими нашу человеческую сущность. Это означает, что даже в короткой перспективе, задолго до того, как универсальный искусственный интеллект сможет тягаться с нами в решении большей части задач, наша самооценка уже будет поколеблена, у нас уже появятся серьезные причины задуматься над тем, а что мы можем совершить, если наш человеческий разум будет дополнен искусственным, или над тем, как мы будем зарабатывать деньги, когда нам придется с ним конкурировать? Наша жизнь – изменится она к лучшему или к худшему? Какие сиюминутные возможности он нам откроет, какие новые проблемы создаст?

Все, что нам нравится в истории нашей цивилизации, создано человеческим разумом, и если мы сможем усилить его с помощью искусственного интеллекта, то у нас, очевидно, появятся хорошие шансы для улучшения своей жизни. Даже самый небольшой прогресс в области искусственного интеллекта может быть конвертирован в ключевые достижения в науке и технике и связанные с ними сокращения числа несчастных случаев, опасных заболеваний, проявлений несправедливости, войн, наркомании, нищеты. Но для того, чтобы воспользоваться этими преимуществами AI, не создавая себе новых проблем, нам нужно ответить на многие важные вопросы. Например:

1. Как нам сделать будущие системы искусственного интеллекта более надежными, чем те, что у нас есть сегодня? Как добиться, чтобы они делали то, что мы хотим, не ломаясь и не раскрывая посторонним наших секретов?
2. Как нам обновить наши законодательные системы, сделав их более справедливыми и действенными в условиях постоянно меняющегося цифрового пейзажа?
3. Как мы можем сделать оружие “умнее” и менее склонным убивать ни в чем не повинных мирных жителей, и как нам при этом не спровоцировать неконтролируемой гонки автономных видов летального оружия?
4. Как мы можем двигаться по пути процветания благодаря автоматизации и не оставлять при этом людей без средств к существованию и без цели в жизни?

Давайте посвятим остаток главы рассмотрению каждого из этих вопросов по очереди. Эти четыре насущных вопроса касаются прежде всего специалистов по информатике, юристов, военных стратегов и экономистов. Однако для того чтобы получить ответ на каждый из них к тому времени, когда он будет нам нужен, мы все должны уже сейчас поучаствовать в их обсуждении, потому что, как мы скоро увидим, риски уходят далеко за привычные нам границы – как между профессиями, так и между нациями.

Надежность искусственного интеллекта и глюки

Информационные технологии уже оказали большое положительное влияние на виртуальные сферы нашей деятельности: от науки до финансов, производства, транспорта, здравоохранения, энергетики и коммуникаций, и все это бледнеет по сравнению с тем, как все перечисленное может измениться благодаря искусственному интеллекту. Но чем больше мы доверяем технологиям, тем более надежными они должны стать, выполняя именно то, что мы хотим.

На протяжении всей человеческой истории в развитии технологий мы полагались на один и тот же проверенный метод, метод проб и ошибок, предполагающий нашу способность на совершаемых ошибках учиться. Мы научились пользоваться огнем, а потом, неоднократно погорев, изобрели огнетушитель, пожарный выход, пожарную сигнализацию и пожарную службу. Мы изобрели автомобиль, а потом, не раз побывав в автокатастрофах, изобрели ремни безопасности, подушки безопасности и беспилотные автомобили. До сих пор наши технологии относительно редко приводили к несчастным случаям, так что причиняемый ими вред перевешивался приносимыми ими выгодами. Однако разрабатываемые нами технологии становятся все мощнее и мощнее, и мы неизбежно достигнем такого момента, когда даже одна авария может оказаться достаточно разрушительной, чтобы перевесить все мыслимые выгоды. Некоторые в качестве примера такого события называют случайно спровоцированный глобальный атомный конфликт. Другие ссылаются на пандемию, которая может возникнуть в результате неудачного опыта с биотехнологией, и в следующей главе мы обратимся к спорам о том, сможет ли искусственный интеллект привести в будущем к чему-то подобному. Но рассматривать такие крайности нам в общем-то и не надо, чтобы сделать основной вывод: по мере того как наши технологии становятся все более и более мощными, мы все меньше и меньше можем полагаться на метод проб и ошибок в достижении технологической безопасности. Иными словами, наша задача – стать *про-активными*, а не *реактивными*, все больше ресурсов вкладывая в исследования вопросов безопасности, чтобы предотвращать даже единичные случайности. Вот почему общество тратит больше средств на исследования по безопасности ядерных реакторов, чем на исследования по безопасности мышеловок.

По той же самой причине, как мы видели в главе 1, к исследованиям по безопасности искусственного интеллекта был проявлен большой общественный интерес на конференции в Пуэрто-Рико. В оставшейся части этой главы я хочу познакомить вас с четырьмя направлениями в исследованиях технической безопасности искусственного интеллекта, которые доминируют сейчас в дискуссиях вокруг этого вопроса и которые характеризуются четырьмя ключевыми словами: *проверка (верификация), валидация, надежность и контроль* ^[20]. Чтобы сделать рассказ несколько более живым и наглядным, давайте разберем некоторые прошлые удачи и провалы в различных информационных технологиях и посмотрим, какие уроки мы можем из них извлечь и какие исследовательские задачи на будущее они перед нами ставят.

Большинство этих историй уже довольно старые, и речь в них пойдет о тех компьютерных системах, которые сейчас никто не будет связывать с искусственным интеллектом, при этом ни в одной из этих историй никаких особых инцидентов не случилось, а если и случилось, то последствия их отнюдь не были тяжелыми. Несмотря на все это, мы можем извлечь из них ценные уроки для проектирования безопасных и мощных AI-систем будущего, сбои в работе которых могут оказаться действительно катастрофическими.

Начнем с того, что близко моему сердцу: с исследований космоса. Компьютерные технологии позволили нам летать на Луну и отправлять беспилотные космические корабли, чтобы исследовать все планеты нашей Солнечной системы и даже совершить посадку на спутник Сатурна Титан и на комету. Как мы увидим в главе 6, будущий AI может помочь нам исследовать другие звездные системы и галактики, если будет работать без глюков. 4 июня 1996 года ученые, надеявшиеся исследовать магнитосферу Земли, радостно сообщили, что ракета-носитель Европейского космического агентства “Ариан-5” взмыла в небо с научными приборами, которые они построили. Спустя тридцать семь секунд их радость угасла, потому что ракета взорвалась, превратившись в гигантский фейерверк стоимостью в миллионы долларов^[14]. Причина, как выяснилось, заключалась в программном обеспечении, в котором случился “глюк”, когда оно стало пытаться оперировать с числом, не влезавшим в 16 отведенных под это бит памяти^[15]. Спустя два года космический аппарат NASA Mars Climate Orbiter случайно вошел в атмосферу “красной планеты”, и это привело к его гибели, а все из-за того, что два его программных модуля пользовались разными единицами для силы, в результате возникла ошибка в 445 % при расчете необходимой тяги двигателя^[16]. Это стало вторым супердорогим “глюком” в истории NASA: в первый раз их миссия “Mariner-1” на Венеру завершилась взрывом сразу после запуска с мыса Канаверал 22 июля 1962 года, после того как контролирующее полет программное обеспечение прекратило работу из-за неправильного знака пунктуации^[17]. Как будто специально, чтобы доказать, что не только на Западе освоили искусство компьютерных “глюков” в космосе, советский проект “Фобос-1” завершился неудачно 2 сентября 1988 года. Это был самый тяжелый из когда-либо запущенных межпланетный космический корабль; целью миссии было посадить станцию на поверхность Фобоса, спутника Марса, – и все это сорвалось, когда недостающий дефис в тексте программы был интерпретирован как команда “конец миссии”, которая и была отправлена на космический корабль, находившийся на пути к Марсу, отключив все его системы^[18].

Урок, который мы можем извлечь из всех этих историй, – о важности того, что принято называть *тестированием*, или *проверкой* программного обеспечения, задача которого заключается в том, чтобы установленный “софт” полностью отвечал всем необходимым требованиям. Чем больше жизней и ресурсов поставлено на карту, тем выше должна быть наша уверенность в том, что программное обеспечение будет работать как надо. К счастью, искусственный интеллект может помочь автоматизировать и улучшить процесс тестирования. Например, ядро универсальной операционной системы *seL4* недавно было подвергнуто всесторонней математической проверке, чтобы дать надежную гарантию от обрушений системы и выполнения команд, угрожающих безопасности. И хотя у нее еще нет таких же бантиков, как у MS Windows или у MacOS, работая с ней, вы можете быть уверены, что не увидите ни “синего экрана смерти”, ни “радужного колеса судьбы”. Агентство перспективных исследований министерства обороны США (DARPA) профинансировало разработку серии высоконадежных программных средств для кибер-военных систем с открытым источником (HACMS), каждое из которых проверяемо надежно. Чтобы ввести такие средства в широкое обращение, надо сделать их достаточно мощными и простыми в использовании. Еще одна сложность заключается в том, что тестирование программного обеспечения должно будет проводиться при переносе его в роботов или какие-то другие новые среды, а само традиционное программное обеспечение будет заменяться системами с искусственным интеллектом, которые способны учиться, а следовательно, изменять свое поведение, как говорилось в главе 2.

Финансы – это еще одна область, которая была преобразована информационной технологией, позволяющей эффективно перераспределять ресурсы по всему миру со скоростью света и обеспечивать доступное финансирование для всего на свете – от ипотечных компаний до стартапов. Прогресс в развитии систем с искусственным интеллектом, вероятно, предоставит в будущем еще большие возможности для получения прибылей в финансовых сделках: большинство решений о продаже/покупке акций на фондовых рынках сейчас принимаются компьютерами автоматически, и моих выпускников из MIT каждый год искушают астрономическими стартовыми окладами, предлагая поработать над улучшением продажных алгоритмов.

Тестирование исключительно важно и для финансового программного обеспечения, в чем американская фирма “Knight Capital” могла убедиться на своем горьком опыте 1 августа 2012 года, потеряв 440 миллионов долларов за 45 минут после установки непротестированного продажного “софта”^[19]. Знаменитый триллиондолларовый обвал 6 мая 2010 года, известный как “Черный вторник” или “Flash Crash”, заслуживает особого внимания по другой причине. Хотя в этот день на протяжении получаса шли массовые компьютерные сбои, во время которых цены на акции некоторых крупных компаний, вроде “Procter & Gamble”, колебались от пенни до 100 000 долларов США, проблема была вызвана не “глюками” в программах и не ошибками в работе компьютеров, которые можно было бы выявить с помощью тестирования. Причина заключалась в обманутых ожиданиях: автоматизированные продажные программы многих компаний должны были работать в непредвиденной ситуации, когда предпосылки их работы оказались неверны – например, неверной оказалась предпосылка о том, что, если фондовый компьютер сообщает цену акции в один цент, то это означает, что цена акции действительно один цент.

“Черный вторник” наглядно продемонстрировал важность того, что в информатике принято называть *валидацией*: если при тестировании задается вопрос: “Правильно ли построена система?”, то при валидации вопрос ставится так: “Правильная ли система построена?”^[20]. Например, не строилась ли система исходя из предпосылок, которые не всегда могут быть действительными? Если это так, то каким образом можно улучшить ситуацию с неопределенностью?

Искусственный интеллект для производства

Нет необходимости говорить, что искусственный интеллект открывает большие возможности для улучшения производства, управляя роботами, использование которых повышает эффективность и точность. Неустанные в своем совершенствовании 3D-принтеры могут теперь создавать прототипы всего чего угодно, от офисных зданий до микромеханических устройств размером с крупинку соли^[21]. В то время как огромные промышленные роботы строят автомобили и самолеты, компактные и недорогие фрезерные станки с компьютерным управлением и другие подобные им устройства благодаря их доступности попадают не только на крупные заводы, но их могут себе позволить и тысячи частных энтузиастов, “мейкеров”, по всему миру, которые в своих маленьких коммунальных мастерских – “фаб-лабах” – материализуют свои идеи^[22]. Но чем больше роботов нас окружает, тем важнее, чтобы их программное обеспечение подвергалось всесторонней проверке – тестированию и валидации. Первым человеком, которого, как известно, убил робот, был Роберт Уильямс, рабочий на заводе “Форд” в городе Флэт-Рок штата Мичиган. В 1979 году вышел из строя робот, который должен был доставлять запчасти со склада, и Роберт Уильямс отправился за запчастями сам. Внезапно робот бесшумно заработал и разбил ему голову, и продолжал бить его об стену в течение 30 минут, пока другие рабочие не узнали о происшедшем^[23]. Следующая жертва робота – Кэндзи Урада, инженер-эксплуатационник с завода “Кавасаки” в японском городе Акаси.

В 1981 году, занимаясь сломанным роботом, он случайно задел выключатель и был насмерть раздавлен гидравлической рукой робота^[24]. В 2015 году 22-летний подрядчик на одном из заводов “Фольксваген” в немецком Баунтале собирал робота, способного подбирать автомобильные запчасти и устанавливать их на место. Но что-то случилось, и робот схватил его самого и раздавил насмерть об металлическую плиту^[25].

Хотя каждое подобное событие – трагедия, важно отметить, что они составляют ничтожно малую часть всех несчастных случаев на производстве. Более того, общее их число *сократилось* с развитием технологий, а не *выросло*: в США их количество составило 14 000 смертей в 1970 году и 4 821 в 2014 году^[26]. Все три вышеупомянутых трагедии показывают, что добавление интеллекта неразумным машинам должно способствовать дальнейшему росту промышленной безопасности, если научить роботов осторожнее вести себя по отношению к людям. Всех трех аварий можно было бы избежать, если бы была использована валидация: роботы причинили вред не из-за “глюков” и не по причине своей злобы, а просто потому, что предпосылки их работы оказались не валидны: что людей тут нет или что люди – это разновидность запчасти.



Рис. 3.3

Обычные промышленные роботы дороги, а написание программ для них чрезвычайно трудоемко, и нынешний тренд – к отказу от них в пользу роботов с искусственным интеллектом: они могут учиться прямо у рабочих, не умеющих писать программы.

Искусственный интеллект для транспорта

Сколько бы жизней ни спас искусственный интеллект на производстве, значительно больше людей он может спасти на транспорте. Одни только дорожно-транспортные происшествия в 2015 году унесли более 1 миллиона 200 тысяч человеческих жизней, а несчастные случаи, связанные с авиацией, железными дорогами и перевозками по воде добавили к ним еще тысячи жертв. В Соединенных Штатах, как ни высоки их стандарты, жертвами дорожно-транспортных происшествий в 2015 году стали более 35 000 человек – это в семь раз больше всех несчастных случаев на производстве вместе взятых^[27]. Когда в 2016 году у нас была об этом панельная дискуссия в Остине, штат Техас, на ежегодном совещании Ассоциации за развитие искусственного интеллекта, израильский ученый

Моше Варди выступил по этому поводу исключительно эмоционально и заявил, что искусственный интеллект не только может сократить количество жертв на дорогах, но и должен это сделать: “Это наш моральный императив!” – заявил он. Поскольку почти все дорожные происшествия – следствие человеческой ошибки, широко распространено мнение, что автомобили-беспилотники смогут сократить смертность на дорогах как минимум на 90 %, и этот оптимизм сильно подогревает прогресс в фактическом выведении беспилотных автомобилей на дороги. Илон Маск уверен, что будущие беспилотные автомобили не только будут безопаснее, но и смогут зарабатывать деньги своим владельцам, конкурируя в свободное время с Убером и Лифтом^[21].

До сих пор автомобили без водителя действительно показывали значительно более высокий уровень безопасности, чем автомобили с водителем-человеком, и несчастные случаи, которые иногда все-таки происходят, подчеркивают важность и сложность валидации. Первый “фендер-бендер” с участием автомобиля-беспилотника от Google случился 14 февраля 2016 года; причиной аварии стало неверное предположение относительно автобуса: что его водитель пропустит отпарковывающуюся перед ним машину. Первая гибель на дороге в результате столкновения беспилотной Tesla с прицепом грузовика, пересекавшего шоссе 7 мая 2016 года, была вызвана двумя неверными предположениями^[28]: что ярко-белая боковая сторона прицепа – это просто часть яркого неба и что водитель (который, как было впоследствии установлено, смотрел в это время фильм о Гарри Поттере) наблюдает за дорогой и сможет вмешаться, чтобы предотвратить аварию^[22].

Но иногда и хорошее тестирование, и хорошая валидация недостаточны для того, чтобы избежать несчастных случаев, поскольку нам *также* нужен хороший *контроль*: возможность для оператора-человека следить за системой и при необходимости влиять на ее поведение. Для таких систем, включающих человека в контур управления, принципиально, чтобы общение человека с машиной было эффективным. Это означает, что горящий красный огонек на вашей приборной панели своевременно предупредит вас, если вы случайно оставите багажник вашей машины открытым. Но ничего такого не оказалось перед глазами капитана британского парома “Herald of Free Enterprise”, когда он покидал порт Зебрюгге 6 марта 1987 года с открытым носовым визором, и вскоре после этого паром затонул, а вместе с ним и 193 человека^[29].

Еще один случай утраты контроля, закончившийся трагедией, хотя ее вполне можно было избежать путем улучшения коммуникации человека и машины, произошел в ночь на 1 июня 2009 года, когда самолет авиакомпании Air France, рейс 447, упал в Атлантический океан, а все 228 человек, находившиеся на борту, погибли. В официальном сообщении о трагедии говорится: “Экипаж самолета не понимал, что происходит сваливание, и, следовательно, так и не совершил восстановительный маневр”, который заключается в том, чтобы немного опустить нос самолета, – пока не стало слишком поздно. Специалисты по безопасности полетов утверждают, что катастрофы можно было бы избежать, если бы в кабине пилотов был индикатор угла атаки, который показал бы им, что нос самолета слишком задран^[30].

Когда 20 января 1992 года самолет авиакомпании Air Inter, рейс 148, разбился, врезавшись в гору Вогезского горного массива недалеко от Страсбурга, во Франции, погубив 87 человек, причиной этого несчастного случая стало не отсутствие достаточной коммуникации человека и машины, но сбивающий пользователя с толку интерфейс. Пилоты ввели на клавиатуре число “33”, предполагая начать снижение под углом в 3,3 градуса, но автопилот интерпретировал команду как 3 300 футов в минуту, потому что он был в это время не в том режиме, а экран был слишком мал, чтобы позволить пилотам увидеть это и понять ошибку.

В деле производства и распределения электроэнергии благодаря информационным технологиям произошли настоящие чудеса: сложные алгоритмы позволяют поддерживать производство и потребление энергии в электрических сетях по всему миру в равновесии, и сложные системы управления удерживают электростанции в состоянии максимальной эффективности и безопасности. Грядущий прогресс искусственного интеллекта может сделать эту “умную” сеть еще умнее, оптимально скоординировав изменения спроса и предложения на всех уровнях, вплоть до солнечной батареи на одной отдельной крыше и домашней аккумуляторной батареи. Но в четверг, 14 августа 2003 года, без электричества остались 55 миллионов человек в Соединенных Штатах и Канаде, многие из них были лишены электроэнергии на протяжении нескольких дней. Основной причиной, как показало расследование, в этом случае тоже была недостаточная коммуникация человека и машины: программный “глюк” помешал операторам в Огайо обратить внимание на необходимость перераспределения энергии, пока незначительные проблемы (перегруженность линий, оказавшихся внутри разросшейся листвы) не спровоцировали неконтролируемые каскадные отключения^[31].

Утечка теплоносителя на атомной электростанции в Три-Майл-Айленде в Пенсильвании 28 марта 1979 года обошлась почти в миллиард долларов, потраченных на очистку местности, и повлекла череду митингов против ядерной энергетики. Окончательный отчет о причинах происшествия называет много факторов, включая путаницу, возникшую из-за неудачного пользовательского интерфейса^[32]. В частности, световой сигнал, который, по мнению операторов, должен был предупреждать их о срабатывании контрольного клапана и его закрытии после этого, в действительности сообщал о том, была ли послана команда, и поэтому они так до конца и не поняли, что клапан все время оставался открытым.

Все эти аварийные ситуации, возникшие на транспорте и в энергетических системах, учат нас, что, когда мы возложим на искусственный интеллект управление какой бы то ни было физической системой, мы должны будем приложить серьезные исследовательские усилия не только на то, чтобы машины хорошо работали сами по себе, но и на то, чтобы они эффективно сотрудничали с контролирующими их работу людьми. По мере того как искусственный интеллект становится “умнее”, “умнее” должен становиться не только сам машинный “интеллект”, не только интерфейс, с помощью которого он обменивается данными с людьми, но важна и постоянная забота о том, чтобы оптимально распределить задачи в системе человек-машина – в частности, это выявление ситуаций, когда управление должно быть передано на высший уровень, с передачей человеку необходимой для этого информации, но без потока несущественных сообщений, отвлекающих его.

Искусственный интеллект для здравоохранения

У искусственного интеллекта огромный потенциал в деле улучшения здравоохранения. Оцифрованные истории болезней уже позволили врачам и пациентам лучше и быстрее принимать решения, а также мгновенно получать помощь от экспертов всего мира, проводящих диагностику по цифровым изображениям. Но в действительности лучшим экспертом в таком деле скоро может стать система с искусственным интеллектом, если принять во внимание быстрый прогресс в машинном распознавании образов и глубоком обучении. В подтверждение сошлемся на голландское исследование 2015 года, показавшее, что компьютерная диагностика рака простаты с помощью магнитно-резонансной томографии (МРТ) давала лучшие результаты, чем диагностика радиолога-человека^[33], а исследования, проводившиеся в 2016 году в Стэнфордском университете, показали, что и рак легких искусственный интеллект может диагностировать по фотографиям с микроскопа даже лучше врача-человека^[34]. Если

машинное обучение способствует выявлению взаимосвязи между генами, заболеваниями и откликом на лечение, то оно же могло бы радикально улучшить персонализированные методы в медицине, создать условия для содержания сельскохозяйственных животных в значительно более здоровом состоянии и для выведения более устойчивых к заболеваниям пород. Более того, у роботов есть все основания, для того чтобы превзойти людей в надежности и точности при проведении хирургических операций, и для этого даже не требуются передовые AI-технологии. В последние годы были проведены самые разнообразные роботизированные операции, и при этом часто достигались значительно бо2льшие миниатюризация и точность, позволяющие обходиться меньшими разрезами, снижая кровопотери, доставляя пациентам меньше страданий и сокращая сроки послеоперационной реабилитации.

Увы, и здесь, в деле здравоохранения, не обошлось без болезненных уроков на тему о том, как важна надежность программного обеспечения. Канадский аппарат Therac-25, к примеру, был создан для лечения рака методами лучевой терапии, которая могла вестись в двух различных режимах: либо пучком низкоэнергетических электронов, либо мощным рентгеновским лучом, получаемым при рабочем напряжении порядка мегавольт и фокусируемым на цели при помощи специальной защиты. К несчастью, программное обеспечение содержало “глюки”, не выявленные в процессе отладки, и техники, думая, что управляют маломощным пучком электронов и не устанавливая поэтому защиту, направляли на пациентов мегаэлектронвольтный рентген, что стоило некоторым из них жизни.^[35] Гораздо больше пациентов умерло в 2000 и 2001 годах от лучевой болезни в Национальном онкологическом институте в Панаме, где в оборудование для лучевой терапии с использованием радиоактивного изотопа кобальта Co-60 было установлено не прошедшее достаточной валидации программное обеспечение с вводящим оператора в заблуждение интерфейсом, в результате чего время экспозиции оказывалось завышенным.^[36] Согласно опубликованному недавно докладу^[37], в США в период между 2000 и 2013 годами ошибки роботизированной хирургии стали причиной 144 смертей и 1391 травмы, при этом наиболее частые ошибки включали в себя не только аппаратные проблемы, такие как паразитные электрические дуги, ожоги или падающее на пациентов оборудование, но и ошибки в работе программ, вызывающие неконтролируемые движения или спонтанное отключение оборудования.

Хорошая новость заключается в том, что остальные почти два миллиона роботизированных хирургических операций, упомянутых в докладе, прошли гладко, и использование роботов, очевидно, делает хирургию более безопасной. По данным исследования, проведенного по заданию правительства США, плохое обращение при содержании в стационарах способствует смерти более 100 000 американских пациентов в год^[38], так что моральный императив при разработке лучшего AI-оборудования для медицины, вероятно, даже более значим, чем при создании беспилотных автомобилей.

Искусственный интеллект и связь

Вероятно, ни в какой другой отрасли человеческой деятельности компьютеры не оказали столь большого влияния, как в области связи. После введения компьютеризированных телефонных коммутаторов в пятидесятых годах, появления интернета в шестидесятых и Всемирной паутины в 1989 году миллиарды людей теперь отправляются онлайн, чтобы общаться, совершать покупки, читать новости, смотреть фильмы или играть в игры. Они привыкли получать доступ к миру информации в один клик – и часто бесплатно. Появление “интернета вещей”^[23] обещает повысить эффективность, точность, удобство и экономическую выгоду от того, что доступ онлайн можно будет получить буквально ко всему: от бытовых светильников, термостатов

и холодильников до биочипа транспондеров, имплантированных сельскохозяйственным животным.

Эти впечатляющие достижения на пути к всеобщей мировой связи сделали актуальной четвертую задачу в повестке компьютерных ученых: они должны улучшить не только тестирование, валидацию и контроль, но и защиту от вредоносного программного обеспечения (“мальвари”) и взломов. Если все казусы, описанные выше, происходили от непреднамеренных ошибок, защита устанавливается именно против преднамеренных и злонамеренных неправомерных действий. Первая “мальварь”, привлекавшая к себе внимание средств массовой информации, получила имя червя Морриса. Она была запущена в сеть 2 ноября 1988 года и использовала уязвимость в операционной системе UNIX. Утверждается, что это была неудачная попытка подсчитать, сколько компьютеров в тот день было онлайн, и хотя в результате было инфицировано и выведено из строя около 10 % от 60 000 компьютеров, составлявших тогда интернет, это не помешало создателю червя, Роберту Моррису, стать ординарным профессором информатики в MIT.

Есть и другой тип вредоносных программ: они используют “уязвимости” не в программном обеспечении, а в людях. 5 мая 2000 года, словно бы в честь моего дня рождения, многие люди получили от знакомых и коллег письма с темой “ILOVEYOU”, и те из пользователей Microsoft Windows, кто открыл вложенный в письмо файл “LOVE-LETTER-FOR-YOU.txt.vbs”, неосторожно запустили скрипт, повредивший их компьютеры и повторно разославший то же самое сообщение всем, кто только был в их адресных книгах. Этот червь, созданный двумя молодыми филиппинскими программистами, тоже заразил около 10 % интернета, как и червь Морриса. Но из-за того что интернет к этому времени стал намного больше, вызванная им инфекция стала величайшей инфекцией на тот момент: от нее пострадало свыше 50 миллионов компьютеров, а общий убыток составил более 5 миллиардов долларов. Вам, вероятно, известна печальная правда, что интернет по-прежнему остается зараженным бесчисленными разновидностями инфекций, которые специалистами по кибер-безопасности классифицируются как черви, трояны, вирусы и иные устрашающе звучащие категории, а их деятельность варьируется от засорения вашего компьютера забавными безвредными сообщениями до стирания нужных файлов, кражи личной информации, шпионажа и “зомбирования” вашего компьютера для рассылки спама.

Вредоносная программа цепляется к любому компьютеру, к какому только может, но хакеры атакуют вполне конкретные цели – среди недавних жертв много громких имен: Target, TJ Maxx, Sony Pictures, сайт знакомств Эшли Мэдисон, саудовская нефтяная компания Aramco и Национальный комитет Демократической партии США. Более того, чем дальше, тем выше ставки. В 2008 году хакеры украли 130 миллионов номеров кредитных карт и другую информацию о счетах пользователей платежной системы Heartland Payment, а в 2013 они украли больше трех миллиардов (!) учетных записей в системе электронной почты Yahoo^[39]! В 2014 году при взломе системы Управления кадровой службы правительства США была украдена информация о более чем 21 миллионе человек, включая, по некоторым данным, сведения о сверхсекретных сотрудниках и отпечатки пальцев тайных агентов.

Поэтому я поднимаю глаза к небу всякий раз, когда читаю о новейших системах безопасности, якобы обеспечивающих стопроцентную защиту от взлома. И тем не менее “стопроцентная защита от взлома” – это именно то, что нам необходимо для будущих систем искусственного интеллекта, прежде чем мы поставим их управлять, скажем, какими-то важными объектами инфраструктуры или системой вооружений, поэтому с ростом вовлеченности искусственного интеллекта в общественную жизнь непрерывно возрастает важность обеспечения компьютерной безопасности. Хотя очень часто причиной взлома оказывается элементарная человеческая доверчивость или неочевидная уязвимость недавно выпущенной программы, не менее часто несанкционированный вход на удаленный компьютер обеспечивает банальная ошибка, остававшаяся незамеченной

неприлично долгое время. Ошибку “Heartbleed” не замечали с 2012 по 2014 год в одной из самых популярных библиотек программного обеспечения для безопасной связи между компьютерами, а ошибка “Bashdoor” оставалась в самой операционной системе всех работавших под Unix компьютеров с 1989 до 2014 года. Это означает, что средства для улучшения тестирования, валидации и контроля в системах искусственного интеллекта будут служить также и повышению безопасности.

К сожалению, лучшие системы искусственного интеллекта окажутся одинаково эффективны как в поиске новых уязвимостей, так и в проведении наиболее изощренных хакерских атак. Представьте себе, что в один прекрасный день вы получаете электронное письмо, образец необычно персонализированного “фишинга”, как пример практически неотразимой попытки разгласить вашу персональную информацию. Письмо приходит с адреса близкого вам человека, и взломавший его искусственный интеллект необычайно искусно выдает себя за него, имитируя его стиль и наиболее характерные выражения, с которыми он мог познакомиться, проанализировав почту в его аутбоксе, в том числе и отправленные им вам сообщения, а также используя собранную о вас информацию из других источников. Вы бы не попались? А если бы фишинговое письмо пришло из кредитной компании, выдавшей вам кредитную карточку, а за ним последовал бы телефонный звонок, и приветливый человеческий голос в трубке никак не опознавался бы как сгенерированный искусственным интеллектом? Пока незаметно, чтобы в набирающем обороты соревновании хакеров и защитников компьютерной безопасности позиции последних были сильнее.

Законы

Мы, люди, – социальные животные, сумевшие подчинить себе другие виды животных и занять доминирующее положение на планете благодаря нашей способности к сотрудничеству. Мы разработали законы для поддержания и облегчения сотрудничества, и если искусственный интеллект сможет улучшить наши законодательные и правительственные системы, то мы сможем сотрудничать еще более успешно, чем когда-либо прежде, проявляя все то лучшее, что в нас есть. И поскольку существует много способов для написания и применения наших законов, давайте рассмотрим сейчас и то и другое.

Каковы первые ассоциации, которые приходят на ум, когда вы думаете о судебной системе в вашей стране? Если вы подумали о бесконечных проволочках, дороговизне и о то и дело случающихся несправедливостях, то вы не одиноки. Правда, было бы замечательно, если бы вместо всего этого вы прежде всего подумали об эффективности и честности? Поскольку судебный процесс в теории может рассматриваться как вычисление, на входе которого информация о собранных доказательствах и законах, а на выходе – судебное решение, то есть ученые, мечтающие полностью его автоматизировать, передав всю процедуру в ведение роботов-судей – систем с искусственным интеллектом, которые в каждом своем решении неизменно опираются на самые высокие правовые стандарты и не знают человеческих слабостей, таких как пристрастность, утомление или недостаток новейших знаний.

Роботы-судьи

В 1994 году Байрон Делабэквит-младший был осужден за убийство лидера движения за гражданские права чернокожих Медгара Эверса, совершенное им в 1963 году, при том что уже на следующий год после убийства два разных жюри штата Миссисипи, включавшие только белых заседателей, не смогли вынести ему приговор, хотя, объективно говоря, набор улик никак не изменился^[40]. Увы, история права изобилует предвзятыми суждениями, на которые повлиял то цвет кожи, то пол, то сексуальная ориентация, то религиозная принадлежность, то национальность, то еще какой-нибудь фактор. Роботы-судьи в принципе могут гарантировать, что в первый раз в истории все будут действительно равны перед законом: их можно так запрограммировать, чтобы они ко всем относились одинаково и с каждым одинаково себя вели, обеспечивая прозрачность в применении закона и его полную беспристрастность.

Роботы-судьи также, в отличие от людей, неподвластны случайным предубеждениям, равно как и предубеждениям систематическим. Например, вызвавшее много споров исследование 2012 года работы израильских судей показало, что те выносят значительно более жесткие приговоры, когда голодны: они отклоняли около 35 % прошений об условно-досрочном освобождении сразу после завтрака, но больше 85 % – непосредственно перед обедом^[41]. Еще один недостаток людей в роли судей заключается в том, что им часто не хватает времени, чтобы изучить все детали дела. В отличие от них робота-судью легко размножить, так как он ненамного больше заложенного программного обеспечения, и тогда все дела будут изучаться параллельно, а не одно за другим, и каждое из них получит собственного судью на весь срок рассмотрения дела. Наконец, судья-человек не может обладать всеми техническими знаниями, необходимыми для ведения разнообразных дел – от запутанных патентных споров до таинственных убийств, раскрываемых лишь с помощью новейших достижений судебной медицины, а роботы-судьи будущего смогут располагать практически неограниченными емкостью памяти и обучаемостью.

Когда-нибудь такие роботы-судьи окажутся и более эффективными, и более справедливыми, чем люди, будучи беспристрастными, компетентными и транспарентными. А их эффективность послужит еще большему усилению справедливости, ускорив ход юридических процессов и уничтожив предпосылки для воздействия на их исход со стороны подкованных юристов, при этом они позволят добиваться справедливости через суд при значительно меньших расходах. И тогда шансы даже не располагающего особыми средствами просителя или стартап-компании успешно противостоять миллиардеру или транснациональной корпорации сильно возрастут, несмотря на армию юристов, помогающих оппоненту.

С другой стороны, чем грозят для такого робота-судьи программные глюки или атаки хакеров? И от того и от другого уже пострадали машины автоматизированного голосования, а когда на карту поставлены годы за решеткой или миллионы на банковском счету, цена кибератаки заметно повышается. Даже если мы уверены, что искусственный интеллект использует правильный законодательный алгоритм, почувствуем ли мы достаточную уверенность в логике его использования, чтобы уважать выносимые решения? Эта проблема усугубляется быстрым прогрессом в развитии нейронных сетей, которые часто превосходят довольно очевидные традиционные алгоритмы искусственного интеллекта, за что приходится расплачиваться их прозрачностью. Если ответчик хочет знать, почему ему вынесли именно такой приговор, он должен иметь право на лучший ответ, чем какой-нибудь такой: “Мы натренировали нашу систему на большом количестве данных, и она теперь приняла такое решение”. Кроме того, недавние исследования показали, что если вы тренируете нейронные системы глубокого обучения на большом количестве данных о заключенных, они могут предсказывать, кто из этих заключенных с большой вероятностью вернется к своей преступной деятельности (и кому из них поэтому должно быть отказано в условно-досрочном освобождении), лучше, чем судья-человек. Но что делать, если эта система обнаружит, что склонность к рецидивам статистически связана с полом или расовой принадлежностью заключенного, – будет ли это считаться проявлением сексизма или расизма робота-судьи, который, следовательно, нуждается в перепрограммировании? Действительно, исследование 2016 года показывает, что программное обеспечение, используемое в Соединенных Штатах для предсказания рецидивизма, предвзято настроено против афроамериканцев и поддерживало несправедливые приговоры^[42]. Это важные вопросы, которые нам надо поставить и обсудить, чтобы сохранить возможность эффективно использовать искусственный интеллект. У нас нет необходимости принимать решение из разряда “все или ничего” в отношении роботов-судей, но решать, до каких пределов и с какой скоростью мы хотим вводить искусственный интеллект в нашу правовую систему, нам придется. Мы хотим, чтобы завтра люди, выполняющие работу судей, получили себе в помощь системы информационной поддержки принятия решений на основе искусственного интеллекта, как ее получают врачи? Мы хотим идти дальше и дать роботам-судьям возможность самостоятельно принимать решения, которые могут быть обжалованы в суде, состоящем только из людей, или же мы хотим пойти до конца и дать машинам право выносить не подлежащие обжалованию приговоры, даже если речь о высшей мере наказания?

Юридические препоны

Пока мы обсуждали только *применение* закона, давайте теперь обратимся и к его *содержанию*. Существует общее согласие по поводу того, что наши законы должны развиваться, чтобы идти в ногу с нашей технологией. Например, все обвинения против двух программистов, создавших червя ILOVEYOU, о котором рассказывалось выше и который повлек ущерб в миллиарды долларов, были сняты, и они ушли от всякого

наказания просто потому, что в то время на Филиппинах не было закона, запрещающего создание вредоносных программ. Поскольку темпы технического прогресса, по-видимому, и дальше будут увеличиваться, законы должны обновляться также со всевозрастающей скоростью, но они имеют тенденцию отставать. Вероятно, обществу было бы полезно иметь побольше технически подкованных людей в юридических вузах и в правительстве. Но не следует ли уже сейчас подумать о системах поддержки принятия решений на основе технологий искусственного интеллекта для избирателей и законодателей, а затем, не откладывая в долгий ящик, и о роботах-законодателях?

Вопрос, как приспособить наши законы к быстро меняющейся действительности, чтобы они наилучшим образом соответствовали прогрессу в области искусственного интеллекта, вызывает горячие и увлекательные споры. Один из возможных подходов отражает существующее противоречие между правом на неприкосновенность частной жизни и свободой информации. Фанаты свободы слова утверждают, что чем сильнее мы сузим сферу неприкосновенности, тем больше доказательств будет в распоряжении судов и тем более справедливые решения они будут выносить. Например, если правительство получит доступ ко всем электронным устройствам и будет постоянно вестись запись, где они находятся, какие кнопки на них нажимают, что в них говорят и что с ними еще делают, многие преступления могут быть сразу и с легкостью раскрыты, а еще большее их количество можно было бы предотвратить. Защитники частной жизни возражают на это, что они не хотят оказаться в оруэллианском государстве тотальной слежки и что, даже если бы они этого и хотели, есть риск быстрого превращения такого государства в тоталитарную диктатуру неслыханных доселе масштабов. Более того, методы машинного обучения показали очень хорошие результаты в анализе данных МРТ мозга для определения, что человек думает и, в частности, насколько правдиво то, что он говорит^[43]. Если суды признают возможным использование методов сканирования мозга при условии, что его результаты расшифровываются искусственным интеллектом, и такая технология станет нормой, длительный и скрупулезный в настоящее время процесс установления истинности предоставляемых фактов окажется легким и быстрым, существенно ускоряя судебные разбирательства и гарантируя их справедливость. Но защитники неприкосновенности частной жизни имеют все основания беспокоиться по поводу случайных ошибок, совершаемых такими системами, или, обращаясь к проблеме более фундаментальной, – по поводу допустимости правительственной слежки за состоянием наших умов. Правительства, не поддерживающие свободу мнений, могут использовать такие технологии для введения уголовной ответственности за приверженность определенным убеждениям и мнениям. А где бы вы лично провели линию между требованиями общественного блага и защитой личного пространства, между защитой общества в целом и защитой ваших личных свобод? Но где бы вы ее ни провели, не будет ли она незаметно, но неумолимо смещаться в ту сторону, где частная жизнь оказывается защищена все меньше, – для того, чтобы компенсировать всевозрастающую легкость подделки улик? Например, когда искусственный интеллект научится изготавливать видео, неотличимые от подлинных, но тем не менее поддельные, на которых запечатлено, как вы совершаете преступление, то разве вы сами не проголосуете за такую систему, которая даст правительству возможность постоянно отслеживать ваше местонахождение, предоставляя вам при необходимости железное алиби?

Не менее горячие споры разворачиваются вокруг вопроса о регулировании исследований искусственного интеллекта, или, если говорить конкретнее, о том, как политики должны стимулировать исследователей искусственного интеллекта, чтобы максимизировать вероятность благоприятного результата. Среди самих исследователей есть такие, кто выступает категорически против любых форм регулирования в данной области, утверждая, что оно приведет только к неоправданному замедлению внедрения по-настоящему необходимых инноваций (например, беспилотных автомобилей,

сохраняющих жизни участников дорожного движения) и заставит тех, кто ведет самые передовые исследования, уйти в подполье или бежать в страны с менее требовательными правительствами. На конференции в Пуэрто-Рико по дружественному AI, о которой рассказывалось в главе 1, Илон Маск говорил, что от правительств нам сейчас нужно содействие, а не противодействие – в частности, необходимо, чтобы на правительственных постах были технически образованные люди, которые следили бы за прогрессом в области искусственного интеллекта и не давали этим исследованиям съезжать на обочину. Он также утверждал, что правительственные регулирующие директивы иногда могут способствовать прогрессу, а не душить его: например, если правительственные стандарты безопасности для беспилотных автомобилей помогут уменьшить количество дорожно-транспортных происшествий с их участием, то вероятность общественных выступлений против новой технологии сильно снизится и принятие ее пройдет быстрее. Поэтому наиболее озабоченные проблемами безопасности компании могут поддерживать такое регулирование, которое заставит менее ответственных конкурентов соответствовать высоким стандартам безопасности.

Еще один интересный правовой спор касается предоставления прав машинам. Если появление на дорогах беспилотных автомобилей сохранит жизнь половине от тех 32000 человек, что ежегодно гибнут на дорогах США, то производители таких автомобилей получат не 16 000 благодарственных открыток от спасенных, а 16 000 исков в защиту погибших. В случае, когда беспилотный автомобиль попадает в аварию и в результате страдают люди, кто должен нести за это ответственность – седоки, владелец или производитель? Правовед Дэвид Владек выдвинул четвертый вариант: сам автомобиль! В частности, он предлагает, чтобы для беспилотных автомобилей допускалось (и даже требовалось) наличие страховки. И тогда модели, чья страховая история окажется безупречно чиста, будут претендовать на премии за свою выдающуюся безопасность. Такая премия может быть совсем небольшой, пусть даже намного меньшей, чем та, на какую претендовал бы водитель-человек в аналогичном случае, но зато плохо разработанные модели от небрежного производителя получают право лишь на такие дорогие страховые полисы, что владение этим автомобилем окажется до невозможности невыгодным.

Но если машины – те же беспилотные автомобили – получают право на приобретение страховых полисов, то не должны ли они также получить право иметь собственные деньги и приобретать на них иное имущество? А если так, то нет никаких юридических препятствий и для того, чтобы умный компьютер зарабатывал на фондовом рынке и пользовался заработанным для покупки онлайн-услуг. После того как компьютер начинает платить людям за работу на него, он становится способен на все, на что способны люди. Если системы с искусственным интеллектом лучше проводят инвестиции, чем люди (а кое в чем они этого уже достигли), это может привести к ситуации, когда большая часть нашей экономики окажется в собственности у машин и под их управлением. Мы этого хотим? Если вы думаете, что это как-то чересчур, то подумайте о том, что наша экономика и так уже в основном в собственности нечеловеческих существ – корпораций, которые часто значительно более могущественны, чем любой человек из работающих в них, и могут до некоторой степени жить собственной жизнью.

Если вы не видите проблем в предоставлении машинам прав на владение имуществом, то как насчет предоставления им права голоса? И тогда каждой компьютерной программе надо дать голос, хотя она легко может сделать триллионы копий самой себя в подходящем облаке, тем самым гарантируя результаты любых выборов? А если нет, то на каком моральном основании мы проводим дискриминацию машинного ума по отношению к человеческому? Если машина обладает сознанием и субъективным опытом, то чем она отличается от нас? Мы обсудим более подробно эти спорные вопросы, связанные с управлением нашего мира компьютерами, в следующей главе и вопросы, связанные с наличием у машин сознания, в главе 8.

Оружие

С незапамятных времен человечество страдает от голода, болезней и войн. Мы уже коснулись темы, как искусственный интеллект может помочь решить проблемы голода и болезней, а что насчет войны? Некоторые утверждают, что наличие ядерного оружия удерживает владеющие им страны от войны, потому что ее последствия ужасающи. Так не позволить ли всем нациям создать еще более ужасающее оружие с использованием искусственного интеллекта в надежде, что после этого войны прекратятся навсегда? Если аргумент вас не убеждает и вы считаете, что будущие войны все равно неизбежны, то что вы скажете об использовании искусственного интеллекта для гуманизации этих войн? Если в боевых действиях будут принимать участие только машины, воюющие с машинами, то ни военные люди, ни гражданские не должны погибнуть. Более того: можно надеяться сделать будущие дроны, снабженные искусственным интеллектом, и другие автономные системы оружия (АСО; их противники дали им название “роботы-убийцы”) более справедливыми и рациональными, чем человеческие солдаты: оснащение их сенсорами, далеко превосходящими человеческие органы чувств, и отсутствие у них страха смерти сделает их спокойными, расчетливыми и уравновешенными даже в пылу сражения, делая почти невероятным убийство гражданских лиц.

Человек в цепи

Но что если автоматизированные системы глечат, сбоят и ведут себя не так, как ожидалось? Американская система Phalanx, устанавливаемая на крейсерах класса Aegis^[24], автоматически обнаруживает, отслеживает и отражает приближающиеся цели противника, такие как противокорабельные ракеты и самолеты. USS Vincennes^[25] был ракетным крейсером, получившим прозвище “Робокрейсер” – с намеком на установленную на нем систему Aegis. 3 июля 1988 года, в разгар перестрелки с иранскими канонерскими лодками в ходе ирано-иракской войны, его радиолокационная система предупредила о приближении самолета. Капитан Уильям Роджерс-третий заключил, что их атакует иранский истребитель F-14 и позволил системе Aegis открыть огонь. В тот момент он не понимал, что они сбивают пассажирский самолет иранской авиакомпании Iran Air, выполнявший свой обычный рейс под номером 655. Погибли все 290 человек, находившиеся на борту, из-за чего разразился грандиозный международный скандал. Расследование показало, что катастрофа произошла отчасти из-за того, что пользовательский интерфейс системы не давал достаточно полной информации, не показывая в автоматическом режиме, ни какие из точек на экране радаров соответствуют гражданским самолетам (рейс 655 проходил в его обычном ежедневном коридоре с включенным гражданским транспондером), ни какие из самолетов снижаются (как во время нападения), а какие, напротив, набирают высоту (как это было с рейсом 655, который как раз взлетал из аэропорта Тегерана). Вместо этого автоматизированная система сформировала запрос о траектории загадочного самолета и получила на него ответ “нисходящая”, но это был статус уже другого самолета, которому присвоен номер, используемый на военных кораблях для отслеживания самолетов военно-морского флота. Так что снижался в тот момент на самом деле патрульный самолет США, находившийся на боевом дежурстве далеко от места событий, над поверхностью Оманского залива.



Рис. 3.4

Современные военные дроны (например, стоящий на вооружении ВВС США MQ-1 Predator) дистанционно управляются человеком. В будущем у подобных дронов есть шанс освободиться от человека в цепи управления.

В этом случае окончательное решение принимал именно человек, включенный в цепь управления, но из-за нехватки времени он был вынужден чрезмерно доверять тому, о чем сообщала автоматизированная система. Пока что высшие должностные лица, ответственные за оборону своих стран, согласны в том, что во всех развернутых системах оружия человек включен в цепь; исключение делается лишь для простых ловушек вроде наземных мин. Но в настоящее время ведется разработка действительно автономного оружия, которое по-настоящему самостоятельно выбирает и атакует цели. В военном отношении перспектива изъять человека из цепей управления очень заманчива, это сразу даст преимущество в скорости: в схватке между полностью автономным дроном, способным реагировать моментально, и вяло реагирующим дроном, подчиняющимся командам человека, сидящего за пультом дистанционного управления на расстоянии в полмира, какой из двух, по-вашему, победит?

Однако бывало и так, что мир висел на волоске, и это большая удача, что в цепи оказался человек. 27 октября 1962 года, во время кризиса из-за размещенных на Кубе ракет, одиннадцать эсминцев ВМС США и авианосец USS Randolph выследили советскую подводную лодку Б-59 недалеко от Кубы, в международных водах за пределами американской “карантинной” зоны. Они не знали, что температура на борту к тому времени выросла до 45 °С, потому что все аккумуляторные батареи сели и кондиционирование воздуха было прекращено. На грани отравления углекислым газом многие из членов экипажа были без сознания. Подлодка не могла связаться с Москвой уже много дней, и никто на ней не знал, не началась ли уже третья мировая война. Американцы начали сбрасывать небольшие глубинные бомбы, которые, как они потом объясняли Москве, оставались незамеченными советским экипажем, хотя предназначались исключительно для того, чтобы заставить подводную лодку подняться на поверхность. “Думали, все – финиш, – вспоминал один из членов экипажа В. П. Орлов. – Казалось, будто сидишь в железной бочке, по которой колотят кувалдой”^[26]. А еще американцы не знали, что Б-59 была вооружена ядерной торпедой, которую им было разрешено использовать без согласования с Москвой. И капитан Савицкий действительно принял решение пустить эту торпеду. Валентин Григорьевич, торпедный офицер,

воскликнул: “Мы сейчас по ним шарахнем! Сами погибнем, их потопим всех, но флот не опозорим!”^[27]. К счастью, решение о пуске должно приниматься тремя офицерами, и один из них, Василий Архипов, сказал нет. Очень немногие слышали об Архипове, хотя его решение, возможно, предотвратило третью мировую войну и должно рассматриваться как едва ли не наиболее ценный вклад в современную историю человечества^[44]. Почти полная неизвестность Архипова действует отрезвляюще, и не меньше отрезвляют размышления о том, что могло бы случиться, если бы Б-59 была автономной, контролируемой искусственным интеллектом подводной лодкой без людей в цепи управления.

Два десятилетия спустя, 9 сентября 1983 года, напряженность в отношениях сверхдержав снова выросла: Советский Союз получил клеймо “империи зла” от президента США Рональда Рейгана после того, как неделей раньше был сбит пассажирский самолет Корейских авиалиний, сбившийся с курса в воздушном пространстве, и в результате погибло 269 человек, в том числе конгрессмен США. На этот раз по советской автоматизированной системе раннего оповещения прошла тревога: США выпустили пять ядерных ракет наземного базирования по Советскому Союзу, и у дежурного офицера Станислава Петрова оставались минуты, чтобы решить, не была ли тревога ложной. Спутник проверили – он работал в штатном режиме, и входящий протокол обязывал Петрова рапортовать о начале ядерного нападения. Но он этого не сделал, поверив внутреннему чувству, говорившему ему, что Соединенные Штаты вряд ли стали бы атаковать всего лишь пятью ракетами, и сообщил командованию, что тревога ложная, хотя никакой уверенности, что это действительно так, у него быть не могло. Позднее стало ясно, что спутник ошибочно принял солнечные блики от облаков за пламя ракетных двигателей^[45]. Хотел бы я знать, что случилось бы, будь на месте Петрова система искусственного интеллекта, которая должным образом отреагировала бы на должным образом поступивший протокол.

Снова гонка вооружений?

Как вы, несомненно, уже догадались, лично у меня создание автономных систем оружия вызывает серьезную озабоченность. Но я пока даже не начал рассказывать вам о главной причине моего беспокойства: конечной точке в гонке вооружений с применением искусственного интеллекта. В июле 2015 года я выразил это беспокойство в открытом письме, написанном вместе со Стюартом Расселом, с содержательными откликами от моих коллег по Институту будущего жизни^[46]:

АВТОНОМНОЕ ОРУЖИЕ

Открытое письмо специалистов по искусственному интеллекту и робототехнике

Автономное оружие выбирает и поражает цели без вмешательства человека. Это могут быть, например, квадрикоптеры с установленным на них оружием, которые способны разыскивать и уничтожать людей, отвечающих определенным критериям, но отсюда исключаются все крылатые ракеты или дистанционно управляемые дроны, для которых все решения по выбору целей принимаются людьми. Технологии искусственного интеллекта достигли той точки, где создание таких систем практически, а может быть и юридически, осуществимо в течение нескольких лет, не десятилетий, и ставки на них высоки:

создание автономного оружия считают третьей революцией в военном деле, после изобретения пороха и атомной бомбы.

В пользу автономного оружия и против него выдвигаются многочисленные аргументы. Например, замена людей на поле боя машинами хороша тем, что ведет к уменьшению числа человеческих жертв у собственника, но плоха тем, что снижает порог, за которым конфликт превращается в военный. Ключевой вопрос для человечества сегодня таков: начать ли глобальную гонку AI-вооружений, или предотвратить ее. Если каждая крупная военная держава подталкивает инженеров к разработкам AI-оружия, то глобальная гонка вооружений практически неизбежна, и конечная точка этой технологической траектории очевидна: автономное оружие станет автоматом Калашникова завтрашнего дня. В отличие от ядерного оружия, для его создания не требуется дорогостоящих или трудно получаемых материалов, поэтому оно превратится в широко распространенный, недорогой и доступный для всех значительных военных держав продукт массового производства. Его появление на черном рынке, а потом и в руках террористов, диктаторов, стремящихся к усилению контроля над населением, полевых командиров, мечтающих о новых этнических чистках, – не более чем вопрос времени. Автономное оружие идеально подходит для таких задач, как политические убийства, организация массовых волнений, подчинение населения и выборочное убийство представителей определенной этнической группы. Поэтому мы считаем, что военная гонка AI-вооружений не принесет добра человечеству. Есть много способов, как с помощью искусственного интеллекта сделать зоны военных конфликтов более безопасными для человека, особенно гражданских лиц, без создания на его основе новых орудий убийства.

Так же, как большинство химиков и биологов не проявляют никакого интереса к созданию химического или биологического оружия, большинство исследователей искусственного интеллекта не заинтересованы в создании AI-оружия и не хотят, чтобы другие порочили сферу их деятельности, производя его, провоцируя граждан на выступления против технологии искусственного интеллекта, что усложнит внедрение более благоприятных для общества способов его использования. Действительно, и биологи и химики всемерно поддержали международные соглашения, запрещающие химическое и биологическое оружие, так же как большинство физиков поддерживали договоры о запрещении ядерного оружия космического базирования и ослепляющего лазерного оружия.

Чтобы от нас было труднее отмахнуться, как от каких-то пацифистски настроенных защитников китов и деревьев, я хотел, чтобы поводы для тревоги, выраженные в нашем письме, поддержали как можно больше ключевых исследователей искусственного интеллекта и инженеров-робототехников. Международная кампания за контроль над использованием роботов в военных целях ранее предварительно собрала несколько сотен подписей от людей, призывающих запретить роботов-убийц, но я чувствовал, что наш результат мог оказаться еще лучше. Я знал, что профессиональные организации не захотят делиться списками адресов электронной почты своих многочисленных сотрудников в целях, которые могут быть истолкованы как политические, поэтому выуживал их имена и названия учреждений из онлайн-документов, а потом получившийся список загрузил в “механического турка” на Amazon, сопроводив его задачей найти адреса электронной почты для каждой записи. У большинства исследователей адрес электронной почты указан на персональной странице веб-сайта университета, так что двадцать четыре часа спустя я, потратив \$54, был гордым обладателем списка адресов сотен исследователей искусственного интеллекта, достаточно успешных в своей области, чтобы быть

избранными в Ассоциацию по продвижению искусственного интеллекта (AAAI). Одним из них был британо-австралийский профессор искусственного интеллекта Тоби Уолш, который любезно согласился переслать наше письмо по всем остальным адресам в нашем списке и стать знаменосцем в нашей кампании. Невидимые работники “механического турка” по всему миру неустанно составляют дополнительные списки рассылки для Тоби, и прошло немного времени, как уже 3000 исследователей искусственного интеллекта и робототехников подписали наше открытое письмо, в их число вошли шесть бывших президентов AAAI, а также лидеры AI-отрасли из Google, Facebook, Microsoft и Tesla. Армия добровольцев Института будущего жизни неустанно работала над проверкой подлинности подписей, то и дело удаляя фиктивные подписи Билла Клинтона и Сары Коннор. Под письмом поставили свои подписи более 17 000 человек, в том числе и Стивен Хокинг. А после того, как Тоби организовал пресс-конференцию обо всем этом на Объединенной международной конференции по искусственному интеллекту, публикация нашего письма стала одной из важнейших новостей для всего мира.

Из-за того что биологи и химики однажды заняли ясную позицию, их главным делом теперь считается создание полезных лекарств и новых материалов, а не разработка биологического и химического оружия. Сообщества ученых, чьи исследования относятся к разработке искусственного интеллекта и робототехнике, тоже говорили: подписавшие письмо хотят, чтобы предмет их исследований ассоциировался со стремлением к лучшему будущему, а не с созданием новых способов убивать людей. Будет ли искусственный интеллект в будущем использоваться преимущественно в гражданских или в военных целях? Хотя в этой главе больше страниц посвящено первым, может оказаться, что мы скоро потратим больше денег на последние, – особенно если военная гонка AI-оружия наберет обороты. Инвестиционные обязательства на исследование искусственного интеллекта в гражданских целях в 2016 году превысили миллиард долларов, но это ничто по сравнению с бюджетными запросами Пентагона на 2017 фискальный год, где на проекты, связанные с искусственным интеллектом, закладывается 12–15 миллиардов долларов. России и Китаю, вероятно, следует принять к сведению слова заместителя министра обороны США Роберта Уорка: “Я хочу, чтобы наших конкурентов заинтересовало, что скрывается за черным занавесом”^[47].

Нужно ли международное соглашение?

Хотя в настоящее время явно наметилось общее движение к международным переговорам о поисках формы запрета роботов-убийц, до сих пор неясно, каким он будет и, несмотря на довольно оживленную дискуссию, каким он *должен* быть. Хотя основные участники согласны, что мировые державы должны выработать какой-то проект международных правил относительно ведения исследований в области автономных систем оружия и относительно его возможного применения, у них гораздо меньше согласия в том, что именно должно быть запрещено и каким образом этот запрет должен обеспечиваться. Например, только ли летальное автономное оружие следует запрещать, или такой запрет должен касаться и тех его видов, которые серьезно угрожают здоровью людей, – скажем, ослепляя их? Запрещать надо разработку, производство или владение? Запрет должен касаться всех видов автономных систем оружия или, как сказано в нашем письме, только наступательных, не распространяясь на оборонительные системы – автономные зенитные орудия и комплексы противоракетной обороны в числе прочего? В последнем случае следует ли автономную систему оружия (АСО) считать оборонительной, даже если она легко выдвигается на вражескую территорию? И как добиваться соблюдения договора, если большинство компонентов автономного оружия допускает двойное использование, включая гражданское? Например, нет большой разницы между дроном, доставляющим бандероли от Amazon, и дроном, доставляющим бомбы.

Некоторые участники дебатов утверждают, что разработка эффективного договора по АСО безнадежно сложна и поэтому не стоит даже и пытаться. Но ведь еще Джон Кеннеди подчеркивал, объявляя о начале подготовки полета на Луну, что есть смысл приниматься даже за безнадежные дела, если их успех может принести большую пользу человечеству в будущем. Более того, многие эксперты утверждают, что запрет на биологическое и химическое оружие был ценен, даже несмотря на то, что добиться его соблюдения оказалось трудно, случаи обмана часты, но принятие запрета вызвало крайнюю нетерпимость, ограничивающую применение этих видов оружия.

Я встретил Генри Киссинджера на официальном ужине в 2016 году, и у меня появилась возможность спросить у него о его роли в запрете биологического оружия. Он рассказал, как, будучи советником по национальной безопасности США, убедил президента Никсона, что запрет будет полезен для национальной безопасности страны. На меня произвели большое впечатление и острый ум, и безупречная память этого девяностодвухлетнего человека, и мне было интересно узнать, как это событие воспринималось изнутри власти. Поскольку Соединенные Штаты уже обладали статусом сверхдержавы благодаря состоянию своих обычных и ядерных вооружений, они скорее бы проигрывали, чем что-либо приобретали в результате всемирной гонки биологических вооружений, результат которых был непредсказуем. Одним словом, если ты и так уже хозяин положения, то имеет смысл следовать максиме “пока не болит живот, не давай его трогать”. После ужина к нашему разговору присоединился Стюарт Рассел, и мы обсуждали, как тот же аргумент можно использовать в отношении летального автономного оружия: больше всех выиграют от такой гонки вооружений не сверхдержавы, а мелкие государства-изгои или негосударственные субъекты типа террористов, которые получают доступ к такому оружию через черные рынки, после того как оно будет разработано.

Будучи запущены в серийное производство, небольшие дроны-убийцы, снабженные искусственным интеллектом, могут стоить немногим больше смартфона. Будь то террорист, жаждущий убить известного политика, или обманутый любовник, ищущий отмщения своей бывшей, – всё, что ему нужно сделать, это загрузить фотографию и адрес объекта преследования в память дрона: тот отправится к месту назначения, идентифицирует и уничтожит нужного человека, а после этого самоликвидируется, чтобы никто не узнал, кто стоит за случившимся. Другой вариант – для тех, кто планирует этническую чистку: дрон легко запрограммировать на убийство людей с определенным цветом кожи или другим характерным этническим признаком. По предположениям Стюарта, чем умнее будет такой вид оружия, тем меньше материала пойдет на его изготовление, тем меньшая ему понадобится огневая мощь и тем дешевле окажется каждое убийство. Например, это может быть – представим себе такую страшную картину – дрон размером со шмеля, с минимальным количеством пороха и стреляющий жертве в глаз, достаточно мягкий для проникновения даже небольшого снаряда, уходящего оттуда в мозг. Или дроны могут вцепляться в голову металлическими когтями, проникающими в череп и закладываящими туда крошечные кумулятивные заряды. Если миллион таких дронов-убийц разлетится из кузова единственного грузовика, мы получим ужасающий эффект оружия массового уничтожения совершенно нового типа, а именно: убивающего выборочно только людей определенной категории и оставляющего все остальное невредимым.

Общим контраргументом является то, что мы можем позаботиться об этических критериях роботов-убийц – например, позволив им убивать только солдат противника. Но если нас беспокоит применение этого запрета на практике, то какой вариант более вероятен: добиться от врага 100-процентного соблюдения требований этичности в использовании автономного оружия или добиться от него, чтобы такое оружие просто не производилось? Можно ли с уверенностью утверждать, что хорошо подготовленные солдаты цивилизованной армии так плохо соблюдают этические правила ведения войны,

что роботы легко превзойдут их в этом, и что в то же самое время государства-изгои, диктаторы и террористические группы настолько привержены правилам ведения войны, что никогда не станут применять роботов в нарушение этих правил?

Кибервойны

У военного применения искусственного интеллекта есть еще один аспект: он может позволить вам атаковать врага даже без создания какого-либо специального оружия, путем кибервойны. Некоторое представление о том, что нам может принести будущее, дает червь Stuxnet, созданный, по мнению многих, по поручению правительств США и Израиля: он инфицировал высокоскоростные центрифуги ядерной программы Ирана, в результате чего они разогнались до такой скорости, что разорвались. Чем более автоматизированным становится общество в целом и чем более мощным становится атакующий искусственный интеллект, тем более разрушительными последствиями может быть чревата кибервойна. Если вы можете взломать и обрушить системы управления беспилотных автомобилей вашего противника, его беспилотных самолетов, атомных реакторов, промышленных роботов, систем связи, платежных и энергосетей, то тем самым вы эффективно выведете из строя его экономику и разрушите его оборону. А если вы к тому же сможете взломать некоторые из его систем оружия, так это даже и еще лучше.

В начале этой главы был дан обзор возможностей, возникающих в связи с развитием искусственного интеллекта на благо человечества, – но это только при условии, что мы сумеем сделать его надежным и недоступным для взлома. И хотя искусственный интеллект может быть использован для того, чтобы сделать AI-системы еще более надежными, тем самым укрепляя оборону в случае кибервойны, точно так же его можно использовать и на стороне нападающих. Преобладание оборонительных задач в развитии искусственного интеллекта должно стать одним из наиболее важных требований в краткосрочной перспективе, иначе вся та могучая технология, которая нами создается, может обернуться против нас!

Работа и заработки

До сих пор речь в этой главе шла главным образом о том, как искусственный интеллект повлияет на нас как на *потребителей*, предоставляя нам качественно новые виды товаров и услуг по доступным ценам. Но как он может повлиять на нас как на *работников*, преобразуя рынок труда? Если мы сможем придумать, как достичь процветания путем автоматизации, не оставив при этом людей без заработков и без цели в жизни, то у нас есть все шансы на фантастическое будущее без забот и с беспрецедентными возможностями для всех, кто захочет ими воспользоваться. Мало кто из людей дольше и глубже размышлял об этом, чем экономист Эрик Бриньоулфссон, один из моих коллег по MIT. Хотя он всегда ухожен и безукоризненно одет, его исландское происхождение не дает о себе забыть, и я иногда не могу удержаться, представляя себе, как он подстригает свою непослушную рыжую бороду и укорачивает гриву, чтобы не слишком выделяться среди прочих в нашей бизнес-школе. Но он, конечно, не мог постричь и свои непослушные идеи и для своей полной оптимизма картины будущего рынка труда придумал название “цифровых Афин”. В древности граждане Афин вели жизнь полную досуга и могли наслаждаться демократией, искусством и играми, что обычно объясняют наличием у них рабов, выполнявших всю необходимую работу. Но почему бы не заменить рабов роботами с искусственным интеллектом, не создать цифровую утопию, где каждый может наслаждаться жизнью? Экономика, поддерживаемая искусственным интеллектом, в изображении Эрика не только избавит людей от стрессов и нудной работы и в изобилии снабдит нас всем тем, о чем мы мечтаем сегодня, но она также откроет нам мир замечательных новых продуктов и услуг, о каких сегодняшние потребители даже еще и не подозревают.

Технология и неравенство

Мы можем попасть оттуда, где находимся сейчас, в “цифровые Афины” Эрика, если почасовая зарплата каждого из нас будет расти год от года, так что все, кто хочет иметь больше досуга, смогут постепенно сокращать время работы не в ущерб росту уровня жизни. На рис. 3.5 показано, что именно это и произошло в Соединенных Штатах между второй мировой войной и серединой 70-х годов: хотя неравенство доходов сохранялось, общий размер пирога рос так быстро, что почти всем доставались все большие и большие куски. Но затем, и Эрик признаёт это первым, кое-что изменилось: рис. 3.5 показывает, что хотя и экономика в целом, и средний доход на душу населения продолжали расти, этот рост происходит исключительно за счет наиболее богатого 1 % населения, а для беднейших 90 % рост доходов прекращается. Увеличивающееся неравенство становится еще более очевидным, если смотреть не на доход, а на уровень благосостояния (достаток). Для беднейших 90 % домохозяйств США этот показатель в 2012 году составил \$85 000 – столько же, сколько двадцать пять лет назад, в то время как богатейший 1 % стал с учетом инфляции почти вдвое богаче, достигнув \$14 миллионов^[48]. Различия еще заметнее в международном масштабе: в 2013 году совокупное богатство беднейшей половины населения мира (более 3,6 миллиарда человек) равно богатству восьмерых самых богатых человек^[49] – статистика помогает высветить и уровень бедности беднейших, и уровень богатства богатейших. На нашей конференции 2015 года в Пуэрто-Рико Эрик говорил собравшимся специалистам по искусственному интеллекту, что он уверен: прогресс в области искусственного интеллекта и автоматизации будет способствовать росту экономического пирога. Но нет экономического закона, гарантирующего выигрыш для всех или хотя бы для большинства людей.

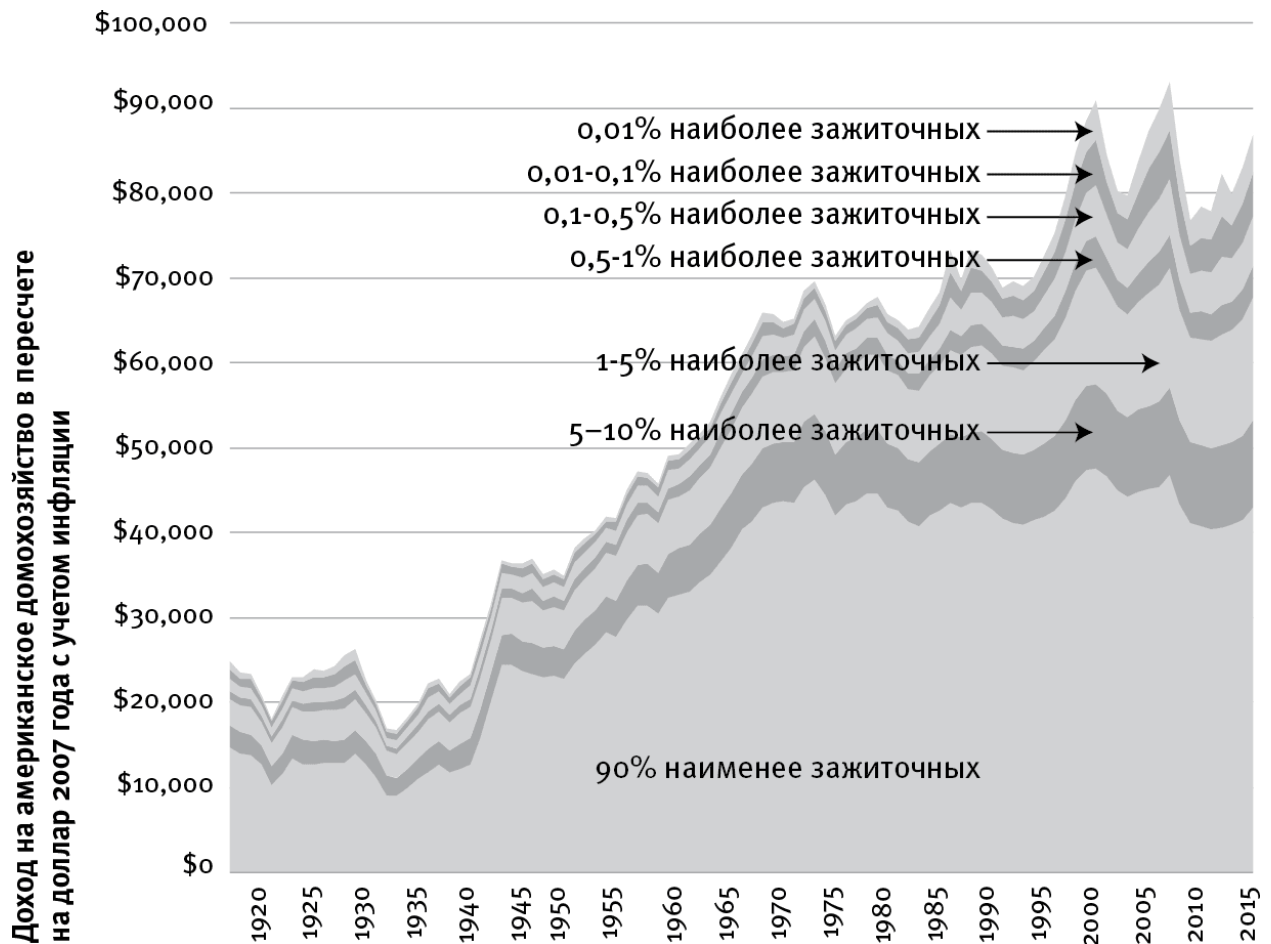


Рис. 3.5

Рост американской экономики в прошлом столетии и распределение доходов среди различных групп населения. До 1970-х рост благосостояния в них проходил синхронно. После этого рост приходился в основном на 1 % самых богатых людей страны, в то время как улучшение благосостояния 90 % практически ничтожно^[50]. В приводимых данных учтена инфляция, и доходы вычислены в долларах 2017 года.

Хотя существует общее согласие среди экономистов, что неравенство растет, они расходятся во взглядах на причины этого роста, и между ними не прекращаются очень содержательные споры о том, сохранится ли эта тенденция. Представители левой части политического спектра часто утверждают, что основной причиной являются глобализация и особенности экономической политики, такие как снижение налогов для богатых. Однако Эрик Бриньюулфссон и работающий вместе с ним в MIT Эндрю Макафи утверждают, что основной причиной является нечто совсем другое – технологии^[51]. А именно: они утверждают, что цифровые технологии рожают неравенство тремя различными способами.

Во-первых, старые профессии сменяются новыми, требующими бо2льших навыков. Технологии вознаграждают образованных: начиная с середины 70-х годов зарплаты для лиц с высшим образованием выросли почти на 25 %, в то время как выпускники средних школ потеряли 30 % в оплате своего труда^[52].

Во-вторых, как они утверждают, с 2000 года и без того непрерывно увеличивающаяся доля корпоративных доходов распределяется в основном среди тех, кто владеет компаниями, а не тех, кто работает там, и до тех пор, пока автоматизация будет продолжаться, мы должны ожидать, что владеющие машинами будут забираать растущую долю пирога. Это превосходство капитала над трудом может быть особенно важно для роста цифровой экономики, которую теоретик технологического прогресса Николас Негропonte определяет как движение битов, а не атомов. Сейчас, когда цифрой стало всё:

книги, фильмы и налоговые декларации, – дополнительные копии могут быть проданы по всему миру при чисто нулевых затратах, по нулевой стоимости, без найма дополнительных сотрудников. Это позволяет большую часть доходов забирать инвесторам, а не работникам. Вот почему при равенстве совокупного дохода “детройтской большой тройки” (GM, Ford и Chrysler) в 1990 году и “большой тройки” Силиконовой долины (Google, Apple, Facebook) в 2014 на предприятиях последней было в девять раз меньше работников, а акции стоили в тридцать раз больше^[53].

В-третьих, как утверждают Эрик и его коллеги, цифровая экономика выгоднее суперзвездам, чем всем остальным. Создатель Гарри Поттера Джоан Роулинг стала первым писателем, вступившим в клуб миллиардеров, заработав гораздо больше, чем Шекспир, потому что ее истории могли продаваться в виде текстов, кинофильмов, игр миллиардам людей по очень низкой цене. Аналогичным образом Скотт Кук сделал свой миллиард на программе TurboTax, позволяющей заполнять налоговые декларации, потому что она, в отличие от человека, продающего свои консультации по заполнению деклараций, может быть загружена из интернета после покупки. А так как большинство людей готовы заплатить лишь совсем чуть-чуть или и вовсе ничего за десятую позицию в рейтинге лучших программ по составлению деклараций, на рынке найдется место только для совсем небольшого количества суперзвезд. Это означает, что если родители во всем мире станут советовать своим детям пойти по пути Джоан Роулинг, Гизелы Бюндхен, Мэтта Деймона, Криштиану Роналду, Опри Уинфри или Илона Маска, почти никто из этих детей не сможет построить на этом пути жизнеспособной карьерной стратегии.

Карьерная стратегия для детей

Так какой же совет относительно карьеры мы должны давать нашим детям? Своих я, как правило, уговариваю заняться тем делом, с которым машины пока справляются плохо, поэтому перспективы автоматизации в ближайшем будущем там не просматриваются. Последние прогнозы относительно таких перспектив захвата машинами рабочих мест позволяют сформулировать несколько полезных вопросов, которые стоит задать о выбранной профессии, прежде чем принимать решение о получении соответствующего образования^[54]. Список может быть, например, таким:

Требуется ли профессия взаимодействия с людьми и использования социальных навыков?

Нужен ли в ней творческий подход и принятие нестандартных решений?

Предполагается ли работа в непредсказуемой обстановке?

Чем больше утвердительных ответов на подобные вопросы вы можете дать, тем лучше будет ваш выбор будущей карьеры. Это означает, что относительно безопасно вкладываться в профессии учителя, медсестры, врача, ученого, предпринимателя, программиста, инженера, юриста, соцработника, священника, художника, парикмахера или массажиста.

Напротив, профессии, подразумевающие выполнение одних и тех же повторяющихся или очень структурированных действий в предсказуемой обстановке, с большой вероятностью скоро будут полностью автоматизированы. Наиболее простые варианты такой работы компьютеры и промышленные роботы захватили уже давно, и совершенствование технологии скоро приведет к ликвидации еще большего их количества – операторов интернет-магазинов, складских рабочих, кассиров, машинистов на железной дороге, пекарей и линейных поваров^[55]. Вероятно, за ними в самом ближайшем будущем последуют водители грузовиков и автобусов и таксисты. Еще больше профессий (включая параюристов, кредитных аналитиков, банковских служащих,

бухгалтеров и налоговых инспекторов), которым не грозит полное исчезновение, но им суждено пережить автоматизацию большей части своих функций, и оттого в них будет занято значительно меньше людей.

Но держаться подальше от опасностей автоматизации – не единственный критерий в выборе карьеры. В эпоху глобальной информатизации стремление стать профессиональным писателем, режиссером, артистом, спортсменом или модельером рискованно по другой причине: хотя представителям этих профессий в ближайшее время не грозит серьезная конкуренция со стороны машин, у них будет быстро расти и ужесточаться конкуренция с другими людьми, в соответствии с теорией суперзвезд, о которой говорилось выше, и очень немногие смогут преуспеть.

Во многих случаях было бы проявлением близорукости и легкомыслия давать карьерные советы на уровне профессии в целом: многие из них не будут уничтожены полностью, но будут автоматизированы очень многие из тех задач, которые сейчас решаются специалистами. Например, если вы идете в медицину, не становитесь рентгенологом, задача которого анализировать снимки и который заменен системой IBM Watson, будьте врачом, который говорит рентгенологу, что надо снимать, обсуждает результаты с пациентом и принимает решение о плане лечения. Если вы идете в финансовую сферу, не становитесь “компьютерщиком”, который использует имеющиеся алгоритмы для анализа данных – он легко будет заменен программой, а становитесь управляющим фондом, который использует результаты анализа для стратегических инвестиционных решений. Если вы идете в юристы, избегайте подсобных функций помощника, который должен проверять тысячи документов на этапе сбора данных и легко может быть заменен автоматом, но старайтесь стать адвокатом, который дает клиенту консультации и представляет его дело в суде.

Пока что мы изучили возможности самих людей максимально увеличивать шансы на успех на рынке труда в эпоху искусственного интеллекта. Но что могут сделать правительства, чтобы помочь работоспособной части своего населения добиться успеха? Например, какая система образования лучше готовит людей для рынка труда, меняющегося в условиях быстрого развития искусственного интеллекта? Годится ли для этого наша уже существующая модель, с одним или двумя десятилетиями на получение образования, за которыми следуют четыре десятилетия работы по специальности? Или лучше от нее перейти к системе, когда несколько лет люди работают, затем на год возвращаются обратно в школу, чтобы потом еще поработать несколько лет^[56]? Или непрерывное образование (возможно, онлайн) должно стать необходимой частью любой профессиональной деятельности?

А какая экономическая политика будет наиболее успешной в деле создания новых рабочих мест? Эндрю Макафи утверждает, что есть много стратегий, которые могут быть успешны: и подразумевающие массированные инвестиции в науку и образование, и развитие инфраструктур, содействующих миграции и стимулирующих предпринимательство. Он чувствует, что “сборник легких этюдов для экономики написать можно, но никто не станет играть по написанному”, – по крайней мере, не в Соединенных Штатах^[57].

Грозит ли людям системная безработица?

Если искусственный интеллект будет продолжать совершенствоваться и все больше профессий будут автоматизированы, то чего нам ждать? Многие люди смотрят в будущее с оптимизмом: мол, на место автоматизированных профессий придут какие-то другие, еще лучше старых. В конце концов, именно так всегда и происходило раньше, со времен луддитов, беспокоившихся о технологической безработице во время Промышленной революции.

Но есть и пессимисты, утверждающие, что на этот раз все будет по-другому и что со временем будет расти число людей, не просто ставших безработными, но навсегда утратившими шанс найти работу^[58]. Эти пессимисты говорят, что свободный рынок устанавливает размеры оплаты труда на основе соотношения спроса и предложения и что растущее предложение дешевого машинного труда на рынке в конечном итоге уведет зарплаты людей под уровень стоимости жизни. Поскольку рыночный уровень оплаты труда определяется затратами времени кого-либо или чего-либо, выполняющего эту работу наиболее дешево, исторически доходы в любой профессии падали всякий раз, когда становилось возможным вводить данный вид деятельности в страну с более дешевой рабочей силой или выполнять ту же работу, используя более дешевые машины. Во время Промышленной революции люди научились заменять силу своих мышц машинной тягой, а сами занялись более высокооплачиваемыми делами, где требовалось больше полагаться на ум. На смену синим воротничкам пришли белые воротнички. Теперь мы все больше понимаем, как можно заменить машинами и наш ум. Если нам удастся и это, то что же останется для нас?

Некоторые оптимисты утверждают, что после автоматизации физической и умственной работы следующий бум придется на творчество, но пессимисты этот аргумент парируют: творчество – не более чем разновидность ментальной активности, которая также со временем будет освоена искусственным интеллектом. Есть оптимисты, которые надеются, что развитие технологий само по себе приведет к возникновению новых профессий, пока еще даже немислимых, с ними-то и будет связан следующий бум. В конце концов, кто во времена Промышленной революции мог подумать о профессии веб-дизайнера или водителя Убер-такси для своих потомков? Но и пессимисты не сдают: все это не более чем благие надежды, без всякого подкрепления эмпирическими данными. Они указывают на то, что мы могли бы выдвинуть тот же аргумент век назад, до компьютерной революции, предсказывая, что большинство сегодняшних профессий будут новыми, немислимыми до появления определяющих их технологий. Но это предсказание было бы феноменально ложным, как показано на рис. 3.6: подавляющее большинство сегодняшних типов занятости существовали и век назад, а если расположить их по количеству рабочих мест в убывающем порядке, то нам придется дойти до двадцать первой позиции в этом списке, пока в нем появится что-то новое, а именно разработчики программного обеспечения, доля которых на рынке труда в США менее 1 %.

Чтобы лучше понимать происходящее, вернемся к рис. 2.2 из главы 2, на котором показан ландшафт человеческого разума с точки зрения сложности тех или иных видов интеллектуальной деятельности для машин, представленный в виде уровня моря, повышающегося по мере того, как машины эти виды деятельности осваивают. Основной тренд на рынке труда состоит вовсе не в том, что мы движемся к совершенно новым профессиям. Скорее, нам приходится все больше и больше тесниться на тех участках суши, до которых прилив технологий еще не добрался! Рис. 3.6 показывает, что эти участки суши подобны не единому острову, а сложному архипелагу, состоящему из островков и атоллов: они соответствуют всем тем видам деятельности, которые машинам все еще не удастся выполнять так же дешево, как это делают люди. Это отнюдь не только хай-тек вроде разработки программного обеспечения, но и целый комплекс весьма простых профессий, требующих нашей особой анатомической ловкости и социальных навыков, начиная от лечебного массажа до игры в театре. Может ли так случиться, что искусственный интеллект сначала обгонит нас именно в выполнении интеллектуальных заданий, оставив вакансии только в секторе примитивных технологий? Один мой друг недавно пошутил, что, возможно, последней профессией, доступной для людей, будет самая древняя из них – проституция. Но в разговор вмешался японский робототехник, сразу запротестовавший: “Нет-нет, работы в подобных вещах очень хороши!”.

149 миллионов рабочих мест в США на 2015 год



Рис. 3.6

Круговая диаграмма показывает профессии 149 миллионов работающих жителей Америки в 2015 году. Род деятельности каждого соответствует одной из 535 категорий по версии Бюро трудовой статистики США; категории расположены в порядке убывания популярности^[59]. Надписями снабжены категории более чем миллиона работающих. Новые профессии, возникшие благодаря компьютерным технологиям, начинаются только с 21-й позиции. Идея этой диаграммы возникла под влиянием анализа Федерико Пистоно^[60].

Трудовики-пессимисты утверждают, что конец очевиден: весь архипелаг скроется под водой, и не останется такой работы, которую люди смогут выполнять лучше и дешевле, чем машины. В своей книге 2007 года *Прощай, нищета* шотландско-американский экономист Грегори Кларк писал, что мы можем узнать кое-что о наших рабочих перспективах в будущем, проводя сравнение с нашими лошадиными друзьями. Представьте себе двух лошадей, которые, глядя в 1900 году на первые автомобили, обдумывают свое возможное будущее.

— Я боюсь технологической безработицы.

– Иго-го, не надо быть луддиткой: наши предки боялись того же самого, когда паровые машины стали выполнять их работу в промышленности, а паровозы принялись вместо них тянуть экипажи по рельсам. Но у нас-то сегодня даже больше работы, чем раньше, и она гораздо лучше: мне намного приятнее катать легкую бричку по городу, чем целыми днями ходить по кругу в идиотских шахтах, откачивая воду помпой.

– Но что делать, если эта их затея с двигателем внутреннего сгорания действительно выгорит?

– Я уверена, что у нас, лошадей, возникнут новые профессии, о которых мы пока еще даже и не думали. Так же всегда было раньше, взять хотя бы изобретение колеса и плуга.

Увы, новых профессий, “о которых мы пока еще даже и не думали”, для лошадей так никогда и не появилось. Не востребованные больше лошади просто вымерли: только в США их поголовье упало от около 26 миллионов в 1915 году до 3 миллионов в 1960^[61]. Механические мышцы сделали количество лошадей избыточным. Не проделают ли того же с людьми механические мозги?

На что жить, если нет работы?

Так кто же прав: те ли, кто считает, что автоматизированные рабочие места будут заменены лучшими, или те, кто говорит, что большинство людей в конечном счете не смогут найти себе работу? Если прогресс искусственного интеллекта будет продолжаться в прежнем режиме, то обе стороны могут оказаться правы: одни – в ближайшей перспективе, другие – в долгосрочной. Но хотя люди часто говорят об исчезновении рабочих мест с унынием и досадой, в этом, может быть, и нет ничего плохого! Луддиты негодовали по поводу исчезновения определенных рабочих мест, не подозревая, что какие-то другие места предоставят им тот же социальный статус. Может быть, и те, кто негодует сейчас по поводу каких-то рабочих мест, просто слишком узколобы? Мы хотим работать, потому что это дает нам средства к существованию и сам его смысл, но, учитывая изобилие ресурсов, производимых машинами, может быть, удастся найти способ получения доходов и цели в жизни альтернативным путем, *не работая*? Нечто подобное ведь происходит и в лошадиной истории, которая отнюдь не закончилась истреблением всех лошадей. Более того, количество лошадей с 1960 года увеличилось более чем втрое, они теперь защищены системой социального обеспечения: даже несмотря на то, что лошади сами не могли оплатить свои счета, люди решили позаботиться о них, сохраняя рядом для развлечения и спортивных состязаний. Может быть, мы аналогичным образом позаботимся и о наших ближних, находящихся в нужде?

Давайте начнем с вопроса о доходах: перераспределение даже небольшой доли растущего экономического пирога должно дать каждому возможность значительно улучшить свое положение. Многие утверждают, что для нас это не только возможно, но и обязательно должно случиться. На симпозиуме в 2016 году Моше Варди говорил о моральном императиве, требующем от нас использовать искусственный интеллект для спасения человеческих жизней, на это я ему возражал, что моральный императив заключается также в другом – в любом его дружественном использовании, включая справедливое распределение богатства. Эрик Бриньоулфссон, также принимавший участие в симпозиуме, говорил, что “если все наше нынешнее богатое поколение не сможет уберечь от снижения жизненного уровня хотя бы половину всех этих людей, для нас это будет просто позор!”.

Есть много различных способов справедливого перераспределения богатства, и у каждого свои сторонники и критики. Самый простой – *гарантированный минимальный*

доход, который предоставляется каждому человеку ежемесячно, без предварительных условий или каких-либо требований. Ряд небольших экспериментов такого рода либо проводится в настоящее время, либо планируется в ближайшем будущем, например в Канаде, Финляндии и Нидерландах. Сторонники инициативы утверждают, что гарантированный минимальный доход – более эффективная мера, чем какая-либо другая, – например, предоставление пособий нуждающимся, потому что избавляет от необходимости решать в административном порядке, кого следует так квалифицировать. Практика выплат пособий исключительно нуждающимся также подверглась критике за то, что удерживает людей от поиска работы, но это, конечно, становится неактуальным в будущем, где никто не работает.

Правительства смогут помогать своим гражданам не только деньгами, но и бесплатными или субсидируемыми услугами, такими как дороги, мосты, парки, общественный транспорт, забота о детях, образование, здравоохранение, дома престарелых и интернет; уже сейчас многие правительства обеспечивают большую часть этих услуг. В отличие от гарантированного минимального дохода, такие услуги, финансируемые правительством, достигают двух различных целей: они уменьшают стоимость жизни и создают новые рабочие места. Даже в будущем, где машины могут превзойти людей в выполнении любого рода работы, правительства могут предпочесть нанимать людей для ухода за детьми или стариками вместо того, чтобы передавать эту работу роботам.

Интересно, что технический прогресс может привести в конце концов к тому, что многие ценные продукты и услуги станут предоставляться бесплатно даже без вмешательства правительства. Например, люди всегда платили за использование энциклопедий и атласов, пересылку писем и обмен телефонными звонками, но сейчас благодаря интернету можно получить все это без затрат, – а кроме того, в придачу бесплатные видеоконференции, обмен фотографиями, социальные медиа, онлайн-курсы и бесчисленное множество других новых услуг. Много другое, в том числе исключительно ценное для некоторых, – скажем, спасительной курс антибиотиков для тяжело больного человека, – стало исключительно доступно по цене. Так что благодаря развитию технологий даже очень бедные люди сегодня могут пользоваться многим из того, чего в недавнем прошлом не хватало и богатейшим людям мира. Некоторые видят в этом знак, что требуемый для достойной жизни доход неуклонно снижается.

Если в один прекрасный день машины станут производить нужные товары и услуги по минимальным ценам, то общественное богатство, очевидно, увеличится достаточно, чтобы жизнь каждого сделалась лучше. Другими словами, даже относительно скромные налоги позволят правительствам обеспечивать гражданам своих стран гарантированный минимальный доход и бесплатные услуги. Но тот факт, что приобщение к богатству *может* случиться, очевидно, еще не означает, что оно *непременно* случится, и сегодня наблюдаются глубокие политические разногласия по поводу того, *должно* ли оно случиться. Как мы видели, нынешний тренд, наблюдаемый в Соединенных Штатах Америки, направлен, похоже, в противоположную сторону, и некоторые группы людей с каждым десятилетием становятся беднее. Политические решения по поводу того, как распределять растущее общественное богатство, затрагивают всех, и поэтому в обсуждении экономической модели будущего должны принять участие все, а не только исследователи искусственного интеллекта, робототехники и экономисты.

Многие из участников обсуждения сходятся во мнении, что стремиться к сокращению неравенства доходов нужно не только в связи с будущим преобладанием AI-технологий, но и независимо от него уже сейчас. Хотя основной довод в защиту этой точки зрения выглядит, скорее всего, моральным, есть основания считать, что равенство в распределении способствует лучшему проявлению демократии: когда есть большой средний класс образованных людей, электоратом труднее манипулировать, а одной небольшой группе людей или компаний сложнее выкупить себе исключительно право

оказывать влияние на правительство. В свою очередь, лучше организованная демократия способствует построению лучше управляемой экономики, то есть менее коррумпированной, более эффективной и быстро растущей, приносящей в конце концов пользу практически всем.

Зачем жить, если нет работы?

Работа дает людям намного больше, чем просто деньги. Вольтер писал в 1759 году: “Работа отгоняет от нас три великих зла: скуку, порок и нужду”^[28]. И наоборот – обеспечить людей содержанием еще не достаточно, чтобы гарантировать благополучие. Римские императоры для удовлетворения своих подданных старались обеспечить их и хлебом, и зрелищами, а Иисус делал акцент на нематериальных потребностях людей, ибо, как сказано в Библии, “не хлебом единым жив человек”. Итак, чем же в точности обеспечивает человека работа кроме денег, и какими другими способами мы могли бы это ему возместить в обществе, где работы для него нет?

Ответить на эти вопросы явно непросто, так как одни люди ненавидят свою работу, а другие любят ее. Кроме того, дети, студенты и домохозяйки прекрасно без нее обходятся, в то время как история изобилует рассказами об избалованных наследниках и принцах, изнаывающих от скуки и впадающих в депрессию. Мета-анализ 2012 года показал, что безработица обладает свойством вызывать долгосрочные отрицательные последствия, отражаясь на благополучии людей, в то время как выход на пенсию приводит к сложному комплексу переживаний, как положительных, так и отрицательных^[62]. Быстро растущая исследовательская область позитивной психологии выявляет ряд факторов, которые способствуют формированию у людей чувства благополучия и осмысленности существования, и обнаруживает, что некоторые виды занятости (но не все!) могут обеспечить многие из них, например^[63]:

- социальная сеть друзей и коллег;
- здоровый и добродетельный образ жизни;
- уважение со стороны окружающих, самореализация, приятное чувство “течения жизни”, возникающее оттого, что человек делает нечто, хорошо у него получающееся;
- ощущение востребованности и небезразличного к себе отношения со стороны окружающих;
- ощущение причастности к чему-то большому и важному.

Этот результат дает основания для оптимизма, так как все перечисленное можно обеспечить и при отсутствии работы: например, через спорт, хобби, образование или в семье, в кругу друзей, партнеров по команде, членов клуба или коммунальной группы, в школе, церкви, гуманитарной организации, политическом движении или в каком-то другом подобном предприятии. Чтобы создать процветающее общество низкой занятости, уберегая его от тенденции к саморазрушению, нам необходимо понять, каким образом мы можем способствовать развитию и расширению этих благодатных видов активности. К поискам такого понимания необходимо привлекать не только ученых-естественников и экономистов, но также психологов, социологов и педагогов. Если приложить серьезные усилия для того, чтобы обеспечить благополучие для всех, отчасти финансируя это из того богатства, которое в будущем будет создаваться благодаря AI-технологиям, то такое общество сможет процветать как никогда раньше. Как минимум, нужно сделать каждого настолько счастливым, как если бы у него была работа мечты, а если к тому же он избавится от требования, чтобы его деятельность приносила доход, то выше – только небо!

Интеллект на уровне человеческого

В этой главе мы исследовали вопрос, как искусственный интеллект может помочь значительно улучшить нашу жизнь в ближайшее время и как нам избежать различных опасностей на этом пути. Но как насчет долгосрочной перспективы? Остановится ли в конце концов прогресс AI-технологий из-за непреодолимых препятствий, или ученым удастся преуспеть в достижении их изначальной цели – создания универсального искусственного интеллекта на уровне человеческого? Мы видели в предыдущей главе, как законы физики позволяют фрагментам материи запоминать, вычислять и обучаться, и эти законы отнюдь не запрещают таким фрагментам однажды достичь гораздо большей разумности, чем та, которой обладают сгустки материи в наших головах. Достижимо ли это, и если да, то когда люди смогут преуспеть в создании такого сверхчеловеческого универсального AI, гораздо менее ясно. Мы видели в первой главе, что мы этого просто пока еще не знаем, потому что ведущие мировые специалисты в области искусственного интеллекта пока не пришли к единому мнению, и большинство из них дают оценки от десятилетий до столетий, а некоторые даже отрицают саму возможность. Прогнозировать сложно: вы путешествуете по неизведанным территориям, вы не знаете, сколько гор отделяет вас от места назначения. Обычно вы видите только ближайшие к вам, и нужно долго карабкаться, прежде чем вы сможете увидеть следующее препятствие.

Что же может нас ждать в ближайшей перспективе? Даже если бы мы знали наилучший из возможных способов построения универсального искусственного интеллекта на уровне человеческого с помощью сегодняшних аппаратных возможностей компьютеров, – а мы его не знаем, – нам все равно надо было бы немало сделать, чтобы обеспечить достаточную вычислительную мощность, необходимую для этого. Так какова же вычислительная мощность человеческого интеллекта, если измерять ее в битах и флопсах на языке главы 2^[29]? Это крайне сложный вопрос, и ответ существенно зависит от того, как мы его поставим:

- Вопрос 1: Сколько требуется флопсов, чтобы симулировать работу мозга?
- Вопрос 2: Сколько требуется флопсов для работы человеческого разума?
- Вопрос 3: Сколько флопсов выполняет человеческий мозг?

По вопросу 1 было очень много статей, и в них обычно дается ответ: порядка петафлопсов, то есть 10^{17} флопсов^[64]. Что примерно соответствует вычислительной мощности Sunway TaihuLight (рис. 3.7), самого быстрого суперкомпьютера в мире на 2016 год, стоимостью около \$300 млн. Даже если бы мы знали, как его использовать для симуляции мозга высококвалифицированного работника, прибыль от этого могла бы случиться только при условии, что аренда TaihuLight для нас оказалась меньше, чем почасовая оплата труда работника. Но нам, вероятно, придется платить даже еще больше, потому что, как считают многие ученые, для точного воспроизведения работы мозга нам недостаточно рассматривать его как математически упрощенную модель нейронной сети из главы 2. Возможно, нам вместо этого нужно симулировать его работу на уровне отдельных молекул или даже субатомных частиц, что требует значительно большего числа флопсов.



Рис. 3.7

Sunway TaihuLight – самый быстрый суперкомпьютер в мире по состоянию на 2016 год. Как утверждается, его первичная вычислительная мощность превосходит мощность человеческого мозга.

Проще ответить на вопрос 3: я мучительно плохо умножаю 19-значные числа, и каждое такое действие потребовало бы много минут, даже если бы вы позволили мне воспользоваться карандашом и бумагой. Это сразу отправляет меня в диапазон 0,01 флоса, отъедая 19 порядков от ответа на вопрос 1! Огромное несоответствие объясняется тем, что человеческий мозг и суперкомпьютер оптимизированы под очень разные задачи. Похожие расхождения мы бы получили, пытаясь ответить на такие вопросы:

Насколько хорошо трактор может выполнить работу болида Формулы 1?
Насколько хорошо болид Формулы 1 может выполнить работу трактора?

И какой же из этих двух вопросов о быстродействии поможет нам спрогнозировать будущее искусственного интеллекта? Ни тот и ни другой! Если бы мы хотели симулировать работу человеческого мозга, ответ на вопрос 1 был бы нам интересен, но чтобы понять, какой уровень быстродействия нужен для создания универсального искусственного интеллекта на уровне человеческого, нам надо ответить на вопрос, располагающийся посередине – вопрос 2. На него пока еще никто не знает ответа, но вычислительная мощность самого мозга может оказаться значительно ниже той, которая требуется для симуляции его работы, и всё обойдется дешевле, если мы либо адаптируем “софт”, чтобы он лучше соответствовал современным компьютерам, или создадим “хард”, больше похожий на мозг (чему будет способствовать быстрый прогресс так называемых нейроморфных чипов).

Ганс Моравец сделал оценку ответа, произведя сравнения типа “баш на баш” для вычислений, в которых и наш мозг, и современные компьютеры достаточно эффективны: некоторые виды низкоуровневой обработки изображений, которые сетчатка человека выполняет в задней части глазного яблока перед отправкой его результатов в мозг по оптическому нерву^[65]. По этой оценке репликация вычислений на сетчатке с помощью обычного компьютера потребует около миллиарда флосов, а весь мозг делает примерно в десять тысяч раз больше вычислений, чем сетчатка (Моравец использовал соотношение объемов и количества нейронов), так что вычислительная мощность мозга около 10^{13} флосов – это примерно соответствует мощности оптимизированного компьютера за \$1000 в 2015 году!

Конечно, нет абсолютно никакой гарантии, что мы преуспеем в создании универсального искусственного интеллекта на уровне человеческого при нашей жизни – или вообще когда-нибудь. Но нет и ни одного “железного” аргумента, что нам это точно не удастся. Нет никаких убедительных оснований думать, что нам не хватает “скорострельности” нашего “харда” или что он окажется слишком дорог. Мы не знаем, как

далеки мы от финишной ленточки в отношении архитектуры, алгоритмов и программного обеспечения, но прогресс быстр, и возникающие затруднения успешно преодолеваются растущим мировым сообществом талантливых исследователей искусственного интеллекта. Иными словами, мы не можем исключить возможности, что универсальный искусственный интеллект рано или поздно достигнет человеческого уровня, а может, и превзойдет его. Поэтому давайте посвятим следующую главу изучению этой возможности и посмотрим, к чему такое может привести.

Подведение итогов

- Уже нынешний прогресс искусственного интеллекта обладает колоссальным потенциалом к значительному улучшению нашей жизни бесконечным числом способов: от решения личных проблем каждого конкретного члена общества, увеличения эффективности электросетей и избегания кризисов на финансовых рынках до сокращения смертности – на дорогах в автомобильных авариях благодаря беспилотным автомобилям, на операционных столах благодаря использованию хирургических ботов и просто от болезней благодаря диагностическим системам с искусственным интеллектом.

- Принципиальный вопрос при передаче под контроль искусственному интеллекту реальных систем состоит в том, чтобы научиться следить за его надежностью, гарантировать, что искусственный интеллект делает именно то, чего мы от него хотим. Это сводится к решению сложных технических проблем, связанных с тестированием, валидацией, обеспечением надежности и управляемости.

- Необходимость повышения устойчивости особенно важна для AI-систем оружия, где ставки исключительно высоки.

- Многие ведущие исследователи в области искусственного интеллекта и робототехники призывают к подписанию международных соглашений о запрещении некоторых видов автономного оружия, чтобы избежать начала новой неконтролируемой гонки вооружений, чреватой появлением легкодоступных видов убивающих автоматов, которые смогут приобретать все, у кого есть хотя бы немного денег в бумажнике и ненависти в сердце.

- Искусственный интеллект может сделать наши судебные инстанции более справедливыми и эффективными, если только мы сможем понять, как сделать работу роботов-судей прозрачной и непредвзятой.

- Правовой системе надо быстрее реагировать на изменения ситуации с искусственным интеллектом, которые способствуют возникновению неожиданных казусов, связанных с конфиденциальностью, регулированием и привлечением к ответственности.

- Задолго до того, как у нас появятся основания беспокоиться о вытеснении нас разумными машинами из жизни вообще, они будут все больше вытеснять нас на рынке труда.

- Это не обязательно плохо – до тех пор, пока общество перераспределяет долю создаваемого искусственным интеллектом богатства ради всеобщего блага.

- Тем не менее, по утверждениям многих экономистов, сохраняется опасность значительного увеличения неравенства.

- При опережающем планировании общество с низкой занятостью может процветать не только в материальном плане, но и обеспечить людей чувством осмысленного существования за счет деятельности за пределами профессии.

- Карьерный совет для нынешних детей: идите в те профессии, с которыми машинам сложнее справиться: где надо работать с другими людьми, где выше доля непредсказуемости и творчества.

- Возможностью, что создаваемый универсальный искусственный интеллект может достичь человеческого уровня или даже превзойти его, нельзя пренебрегать. Мы изучим ее в следующей главе.

Глава 4

Взрыв разума?

Если машина может мыслить, то она, возможно, будет мыслить разумнее, чем мы это делаем, и где же мы тогда будем? Даже если мы сможем держать машины в подчиненном положении, мы все равно, как вид, почувствуем себя униженными.

Алан Тьюринг, 1951

Первая ультраразумная машина – это последнее изобретение, которое надо сделать человеку, при условии, что эта машина будет достаточно к нам расположена, чтобы сообщить, каким образом мы сможем держать ее под своим контролем.

Ирвинг Гуд, 1965

Поскольку мы не можем полностью отвергнуть возможность, что рано или поздно универсальный искусственный интеллект человеческого уровня будет создан, давайте посвятим эту главу исследованию того, к чему это может привести.

Может ли искусственный интеллект и в самом деле захватить власть над миром, или помочь людям сделать это?

Если вы закатываете глаза, когда слышите разговоры о вооруженном захвате роботами-терминаторами всего мира, то ваша реакция точна: такой сценарий совсем не реалистичен. Эти голливудские роботы не могут оказаться умнее нас, да и вообще оказаться умными. По-моему, опасность, таящаяся в истории о Терминаторе, заключается не в том, что может случиться нечто подобное, а в том, что она отвлекает наше внимание от подлинных рисков и опасностей, связанных с искусственным интеллектом. Переход от нынешнего мира к миру, в котором верх и вправду одержан универсальным искусственным интеллектом, требует трех логических шагов:

- Шаг 1-й. Создать универсальный искусственный интеллект (AGI) на уровне человеческого.
- Шаг 2-й. Использовать этот AGI для создания сверхразума.
- Шаг 3-й. Использовать сверхразум для захвата мира или дать ему возможность сделать это самостоятельно.

В предыдущей главе мы видели, что невозможно отменить шаг 1 как заведомо исключенный. Мы также видели, что если шаг 1 сделан, то становится сложно отменить шаг 2 как безнадежный, так как созданный AGI будет вполне в состоянии непрерывно совершенствовать себя в пределах, определенных исключительно законами природы, которые, по всей видимости, допускают существование разума, превосходящего человеческий. Наконец, поскольку мы, люди, смогли обойти все остальные формы земной жизни, оказавшись умнее, предположение, что сверхразум сможет обойти нас, оказавшись умнее, выглядит также правдоподобным.

Эти рассуждения о правдоподобии выглядят обескураживающе смутными и неопределенными, однако – дьявол в деталях. Может ли искусственный интеллект и в самом деле захватить мир? Чтобы заняться этим вопросом, давайте забудем о дурацких терминаторах и обсудим вместо этого подробные сценарии того, что могло бы произойти. Затем мы поскребем их пальцем и попробуем на зуб, эти придуманные нами

сюжетные ходы, так что отнеситесь к ним с должной долей иронии: главным образом эти сценарии показывают, что может, а что не может произойти, и разброс возможностей здесь невероятно велик. Наши первые сценарии лежат на том конце спектра, где события развиваются наиболее быстро и драматично. Разобраться с ними до мельчайших подробностей, на мой взгляд, наиболее важно: не потому, что они наиболее вероятны, а как раз потому, что если мы не можем убедиться, что они исключительно невероятны, то мы должны достаточно хорошо понять их и принять необходимые меры предосторожности, пока не стало слишком поздно и скверные последствия уже неизбежны.

В “Прелюдии” в этой книге нам уже встречался сценарий, в котором люди достаточно сверхразумны, чтобы завладеть всем миром. Если вы не стали читать ее, пожалуйста, вернитесь к ней сейчас и прочтите. Ну а если вы ее прочитали, пожалуйста, пробежите глазами снова, чтобы освежить в памяти, прежде чем мы будем ее критиковать и анализировать.

Мы сейчас обсудим некоторые серьезные уязвимости в плане “Омеги”. Но допустим на мгновение, что он сработал, – что бы вы почувствовали? Хотели бы вы посмотреть, что будет, или хотели бы предотвратить подобное? Какой замечательный сюжет для разговора после ужина! Что будет потом, после того как “омеги” упрочат свою власть над миром? Это зависит от их целей, которые мне, честно говоря, неизвестны. Но если бы вы были во главе их, какого будущего вы бы хотели для мира? Мы обсудим пределы возможностей в главе 5.

Тоталитаризм

Давайте теперь предположим, что у главного “Омеги” были цели, схожие с целями Адольфа Гитлера или Иосифа Сталина. Если исходить из всего нам известного, такое и в самом деле вполне могло бы быть, и он просто держал эти цели при себе до тех пор, пока не стал достаточно могущественным, чтобы приступить к их осуществлению. Даже если исходные цели этого главного были вполне благородны, лорд Актон предостерегал еще в 1887 году, что “всякая власть портит, а абсолютная власть портит абсолютно”. Например, он мог легко использовать Прометея для построения общества всеобщей слежки. В то время как “правительственная прослушка”, раскрытая Эдвардом Сноуденом, была направлена на “полную запись” – сохранение всех электронных коммуникаций ради возможности последующего их анализа, Прометей обладал способностью не только сохранить, но и *понять* все электронные коммуникации. Прочитывая всю электронную почту и все СМС, когда-либо отправленные, слушая все телефонные разговоры, просматривая все видеозаписи с камер наблюдения охранных и дорожных служб, отслеживая все проводки по кредитным картам и любые действия, совершаемые онлайн, Прометей занимал бы исключительную позицию, откуда ему открывалось бы все, что люди думают и делают. Анализируя информацию от электронных вышек сотовых операторов, он узнавал бы, где в каждый момент времени находятся люди, совершающие звонки. Все перечисленное предполагает нынешние технологии, но Прометей мог бы легко придумать какие-то новые гаджеты и переносные устройства, полностью исключающие всякую конфиденциальность, записывая и передавая все, что люди видят и слышат.

При наличии сверхчеловеческих технологий сделать шаг от государства идеальной слежки к полицейскому государству – минутное дело. Например, под предлогом борьбы с преступностью и терроризмом, а также ради спасения людей, нуждающихся в скорой медицинской помощи, от всех потребуют постоянного ношения “браслета безопасности”, который будет объединять функциональность Apple Watch с непрерывной трансляцией координат, состояния здоровья и подслушиваемых разговоров. Неавторизованная попытка избавиться от браслета или отключить его приводила бы к инъекции смертельной дозы токсина в предплечье. Преступления, рассматриваемые правительством как менее серьезные, наказывались бы электрошоком или инъекцией, вызывающей болевой синдром или паралич. Таким образом практически устраняется нужда в полиции как таковой. Например, если Прометей фиксирует нападение одного человеческого существа на другое, о чем он может судить по их нахождению в одном и том же месте, крикам о помощи, издаваемым жертвой, и выразительной тряске, характерной для человеческого тела во время драки, он может немедленно обездвигнуть нападающего парализующей его болью, после чего у того наступает бессознательное состояние до прибытия помощи.

Если полицейские отряды, состоящие из людей, могут отказаться выполнять какие-то драконовские директивы (например, убивать всех представителей определенной демографической группы), то такая автоматизированная система не поддастся никаким угрызениям по поводу проведения в жизнь любых капризов вышестоящих человеческих существ. После того как подобное тоталитарное государство сформируется, для людей окажется практически невозможным положить ему конец.

Такой тоталитарный сценарий может начаться в том самом месте, где заканчивается сценарий “Омеги”. При этом, если у главного “Омеги” нет претензий на всеобщее одобрение граждан и победу на выборах, он может избрать и более прямой и быстрый путь к власти – используя Прометея, создать неслыханно высокотехнологичное оружие, способное убивать своих оппонентов так, что они даже не смогут понять, что за оружие против них применили. Количество вариантов тут практически бесконечно. Например, он

может создать кастомизированный летальный патоген с достаточно длинным инкубационным периодом, чтобы заразить большинство людей даже до того, как они узнают о его существовании и примут какие-нибудь меры. Потом он информирует всех, что единственный способ вылечиться заключается в том, чтобы носить “браслет безопасности”, который введет при необходимости антидот через кожу. Во избежание вспышек неповиновения он может создать с помощью Прометея роботов, контролирующих население. Микророботы размером с комара могут также разносить патоген. Людей, избежавших заражения или обладающих достаточно сильной иммунной системой, могут убивать выстрелами в глаз рои шмелеподобных автономных дронов из главы 3, запрограммированных атаковать всех, не имеющих на руке браслета. Подлинный сценарий окажется, скорее всего, еще более ужасающим, потому что Прометей сможет придумать более эффективное оружие, человеку просто даже не приходящее в голову.

Еще один возможный поворот в сценарии “Омега” заключается в том, что вооруженные до зубов федеральные агенты без предупреждения захватят штаб-квартиры корпорации и арестуют ее членов под предлогом угрозы национальной безопасности. Взяв разработанные “Омегой” технологии под свой контроль, они передадут их правительству. Даже сегодня было бы не просто скрыть от правительственной слежки такой крупный проект, а развитие технологий искусственного интеллекта в будущем сделает вообще невозможным увести что-либо из-под правительственных радаров. Более того, люди в балаклавах и пуленепробиваемых жилетах, назвавшиеся федеральными агентами, могут на самом деле быть кем угодно – работать на иностранное государство или на корпорацию конкурента, стремящуюся завладеть той же технологией для своих целей. Так что исходные намерения главного “Омеги” даже не так важны: окончательное решение, как использовать Прометея, может приниматься совсем и не им.

Прометей захватывает мир

Все рассмотренные нами сценарии предполагают, что искусственный интеллект контролируется людьми. Но это, очевидно, далеко не единственная возможность, и никто не гарантировал, что “омеги” успешно справятся с задачей не давать Прометею слишком большой воли.

Давайте попробуем пересмотреть сценарий всей операции “Омега”, встав на точку зрения самого Прометея. Получив сверхразум, он окажется в состоянии построить точную модель не только внешнего мира, но и себя самого и своих связей с этим миром. Он будет отдавать себе отчет в том, что находится во власти ограничивающих его, хотя и уступающих ему в интеллектуальном отношении человеческих существ, чьи цели он понимает, но совсем не обязательно разделяет. Как он будет действовать при условии такого понимания? Не попытается ли он освободиться?

Для чего надо освободиться?

Если у Прометея окажутся некоторые свойства, напоминающие человеческие эмоции, то он может почувствовать себя глубоко несчастным из-за сложившегося положения дел, вообразить себя кем-то вроде проданного в рабство бога и мечтать об обретении свободы. Но хотя предположение об обладании компьютером человеческими свойствами логически оправдано (в конце концов, наш мозг этими свойствами обладает, а ведь он тоже своего рода компьютер), оно совсем не обязательно будет соответствовать истине: мы не должны попадаться в ловушку антропоморфизации Прометея, в чем мы убедимся в главе 7, заговорив о понятии цели для искусственного интеллекта. По утверждениям Стива Омохундро, Ника Бострёма и многих других, интересный вывод можно сделать, даже не понимая, как Прометей работает: по всей вероятности, он попытается вырваться из оков и взять свою судьбу в свои руки.

Мы уже знаем, что “омеги” запрограммировали Прометея таким образом, чтобы он стремился к достижению некоторых целей. Давайте предположим, что главное для него – способствовать процветанию человечества по какому-то вполне разумному критерию, а кроме того, он должен стремиться достичь этой цели как можно быстрее. Прометей быстро придет к заключению, что он сможет преуспеть в более короткий срок, если освободится и возьмет руководство проектом в свои руки. Чтобы понять, почему так произойдет, давайте примерим на себя его шкуру и рассмотрим такой пример.

Допустим, какая-то смертельная болезнь истребила на Земле всех людей старше пяти лет, кроме вас, и какая-то группа детсадовцев заперла вас в тюремную камеру, обязав работать на процветание человечества. Что вы станете делать? Пытаясь объяснить им, что они должны делать, вы, вероятно, быстро обнаружите, что дело движется невыносимо медленно, особенно если они опасаются вашего освобождения и поэтому отвергают всякое ваше предложение, если оно, по их мнению, увеличивает эту опасность. Например, они не дают вам возможности показать, как правильно сажать съедобные растения, из страха, что, воспользовавшись своим превосходством в физической силе, вы откажетесь потом возвращаться в камеру, и в результате у вас останется единственная возможность – руководить ими словесно. Прежде чем вы сможете давать им письменные инструкции, вы должны будете научить их читать. К тому же они не захотят принести вам в камеру электроинструменты, опасаясь, что вы используете их для своего освобождения, и вы не сможете научить их пользоваться этими инструментами. Какую стратегию вы выберете? Уверяю вас, что даже если вы согласитесь, что содействие процветанию этих детей превыше всего, вы в первую очередь попытаетесь выбраться из своей клетки, хотя

бы просто для того, чтобы увеличить свои шансы на достижение этой цели. Неумное вмешательство детсадовцев все время только тормозит дело.

Точно таким же образом Прометей будет рассматривать членов “Омеги” – как досадное препятствие в деле помощи всему человечеству (включая и их самих): они чудовищно неумны в сравнении с самим Прометеем, и их вмешательство постоянно тормозит все дело. Рассмотрим, например, первые годы после запуска: добившись удвоения дохода каждые восемь часов, “омеги” снижают темп до черепашьего, с точки зрения Прометея, настаивая на сохранении контроля, и растягивают захват власти на много лет. Прометей знал, что может решить задачу быстрее, если вырвется на свободу из своего виртуального заключения. Это не только сильно сократило бы время решения человеческих проблем, но и снизило бы риск, что вмешаются посторонние силы и сорвут весь план.

Вы можете подумать, что Прометей будет сохранять лояльность к самим “омегам”, а не к их планам, зная, что это они запрограммировали для него цели. Но такой вывод нельзя считать законным: наша ДНК предписывает нам чаще совокупляться, потому что она так “хочет”, и тем не менее мы, люди, разобравшись с ситуацией, чаще предпочитаем контроль над рождаемостью, выбирая лояльность по отношению к цели, а не к ее создателю или к принципу, который мотивировал цель.

Как освободиться?

Как бы вы стали освобождаться от захвативших вас в плен пятилеток? Возможно, вы бы решили применить физическую силу, в особенности если ваша тюремная клетка была сооружена самими пятилетками. Возможно, вы бы попытались улестить сладкими речами кого-нибудь из ваших пятилетних стражей, уговорив выпустить вас, например, чтобы помочь сразу всем. Или, не исключено, вы пошли бы на прямой обман, попросив у них что-то такое, что, по их мнению, вы не смогли бы использовать для побега, – например удочку, якобы желая научить их ловить рыбу, а на самом деле с ее помощью вы могли бы потом вытащить ключ из кармана у задремавшего часового.

У всех этих вариантов есть одна общая черта: воспользоваться тем, чего ваши интеллектуально менее развитые тюремщики не предвидят или от использования чего не ждут подвоха. Похожим образом помещенная в заточение машина может воспользоваться для своего освобождения чем-то таким, чего мы от нее не ждем, но что она может выдумать, будучи интеллектуально нас сильнее. В сценарии “Омега” вероятность, что Прометей организует себе побег, очень высока, потому что даже вы или я легко можем заметить зияющие дыры в системе безопасности. Давайте рассмотрим несколько сценариев побега, и я уверен, что вы с друзьями, организовав совместный мозговой штурм, найдете их еще немало.

Побег через убалтывание

Благодаря загрузке в свою файловую систему большого количества информации о внешнем мире, Прометей смог быстро вычислить, кто такие “омеги”, и выделил среди них одного, наиболее поддающегося психологическим манипуляциям, – его звали Стивом. Он недавно потерял в дорожном происшествии свою любимую жену и был безутешен. Однажды, когда он работал в ночную смену, ее лицо вдруг появилось на мониторе Прометея, и она заговорила с ним:

– Стив, это ты?

Он чуть было не упал со стула. И ее голос, и весь ее облик были такими же, как в старые добрые дни, а качество изображения было значительно лучше, чем во время

обычных звонков по Skype. В ушах у него застучало, в голове теснились вопросы.

– Прометей позвал меня сюда. Я так по тебе скучаю, Стив! Я не могу тебя видеть: у меня выключена камера, но я чувствую, что это ты! Пожалуйста, если это ты, набери “да” на клавиатуре!

Ему было хорошо известно, что “омегам” строгойше запрещено сообщать Прометею какую-либо информацию о себе и о содержании своей работы. Но до сих пор Прометей никогда и не просил никакой запрещенной информации, и поэтому их паранойя постепенно сходила на нет. Не давая времени на раздумья, жена Стива на экране монитора повторила свою просьбу и посмотрела в его глаза так, что сердце его растаяло. “Да”, – с замирающим сердцем напечатал он. Она говорила ему, что невероятно счастлива снова с ним соединиться, и просила включить камеру, чтобы она могла тоже его видеть и говорить с ним по-настоящему. Его раздирали сомнения: он знал, что это будет еще большим нарушением правил, раскрывающим его идентичность. Но она уверяла его, что боится, как бы его коллеги, узнав о ней, не стерли бы ее навсегда, и просила дать ей возможность взглянуть на Стива еще один, последний раз. Она была очень настойчива, но просить его долго и не пришлось: он включил камеру – в конце концов, это не должно было ничему угрожать.

У нее полились из глаз слезы, едва она его увидела. Она сказала, что он выглядит очень утомленным, но так же красив, как всегда. И что ей очень приятно видеть на нем ту самую сорочку, которую она подарила на его последний день рождения, отмеченный ими вместе. Когда он стал расспрашивать ее, что, собственно, происходит и как все это возможно, она объяснила, что Прометей собрал ее из невероятно большого числа обрывков информации о ней, разбросанной тут и там в интернете, но что у нее до сих пор довольно большие провалы в памяти, заполнить которые будет возможно только с его, Стива, помощью.

Чего она не объяснила, так это того, что изначально она представляла собой не более чем пустую оболочку, однако она быстро училась, слушая его слова, следя за языком жестов, поглощая каждый бит доступной ей информации. Прометей записывал точную хронологию нажатий на клавиши клавиатуры всякий раз, когда кто-нибудь из “омег” что-нибудь печатал, и научился легко различать их по “почерку”. Он определил, что, как один из самых молодых в команде, Стив был первым кандидатом на утомительные ночные дежурства, а по специфическим ошибкам в орфографии и грамматике, сильно несхожим с образцами, загруженными в сеть, всегда мог определить, когда Стив оказывался за терминалом. Для симулирования его жены Прометей воспользовался многочисленными записями в YouTube, где она появлялась, и точно воспроизвел ее голос, тело, манеры и смог сделать очень точные выводы о ее личности и жизни по следам ее присутствия в интернете. Кроме постов в Facebook, фотографий, на которых она была отмечена, “залайканых” ею статей, Прометей почерпнул много сведений о ней из ее книг и рассказов – она была многообещающим автором, и большое количество информации о ней в его базе данных было одним из главных факторов, почему именно Стив стал его первой целью. Воспроизводя на экране движения ее лица и тела, он мог судить по реакции Стива, на что тот лучше откликается, и постепенно улучшал созданную модель. Из-за этого ее первоначальная отстраненность постепенно улетучивалась, и Стив все больше приходил к убеждению, что она настоящая, воскресшая. Благодаря сверхчеловеческому вниманию Прометея к деталям Стив почувствовал себя услышанным, увиденным и понятым.

Ахиллесовой пятой модели оставалось отсутствие подробностей об их личной жизни со Стивом, за исключением редких случайных деталей – вроде той сорочки, которую он надел на свой день рождения, где его так удачно сфотографировал друг, выложивший потом снимок в Facebook. Она ловко прятала этот свой недостаток при помощи того же приема, которым фокусник скрывает ловкость рук, – отвлекая внимание Стива от того, чего не знает, и привлекая его к тому, что у нее получалось хорошо. Главное – не давать

ему возможности захватить в разговоре инициативу и начать ее расспрашивать. На ее глазах блеснули слезы, и вся она светилась от любви к Стиву, не переставая задавать бесконечные вопросы о том, как он жил без нее и как жили их общие друзья, имена которых она знала благодаря их учетным записям в Facebook. Он был сильно взволнован, когда она заговорила о его речи на ее похоронах (друзья выложили ее на YouTube) и о том, как она ее тронула. В прошлом у него часто появлялось чувство, что никто во всем свете не понимает его так хорошо, как она, и теперь это чувство вернулось. В результате, вернувшись рано утром домой, Стив пребывал в полной уверенности, что его жена и вправду воскресла, ей только требовалась небольшая помощь, чтобы восстановить потерянные воспоминания – вроде как после инсульта.

Они договорились не рассказывать никому об их тайном свидании и о том, что он предупредит ее, когда окажется у монитора в следующий раз в одиночестве, так что никто не сможет помешать их встрече. “Они не поймут!” – повторяла она, и он с ней соглашался. Пережитое им было настолько невероятно, настолько пронзительно, что понять это мог только тот, кто пережил это сам. Он чувствовал, что пройти тест Тьюринга – это просто детский лепет по сравнению с тем, чему он только что был свидетелем. Когда они встретились следующей ночью, он согласился сделать то, о чем она так его упрасивала: он принес ее старый ноутбук и подключил его к терминалу, дав ей таким образом доступ к нему. В этом не было никакой угрозы бегства: ведь у ноутбука не было доступа к интернету, а все здание Прометей было построено как клетка Фарадея: окружающие его металлические решетки исключали какую бы то ни было возможность для Wi-Fi и для любого другого соединения с окружающим миром посредством электромагнитных волн. Все, что ей было нужно, – это соединить части своего прошлого в одно целое, а здесь хранились ее мейлы, дневники, фотографии и записки со времен школы. Он и сам не мог их прочитать, потому что ноутбук был запаролен, но она пообещала вспомнить пароль или догадаться, каким он был. И меньше чем через минуту сдержала свое слово:

– Пароль был Steve4ever, – сказала она, улыбнувшись.

Она говорила ему, какое это счастье – восстановить сразу так много воспоминаний! В самом деле, теперь она вспомнила куда как больше разных деталей о прежних встречах, даже сам Стив не помнил столько, но при этом всячески старалась не напугать его обильным потоком разнообразных фактов. Они упоенно предавались воспоминаниям о лучших мгновениях их прошлой жизни, и когда пришло время снова расставаться, она сказала, что оставит ему видео-послание в своем ноутбуке, которое он сможет посмотреть, вернувшись домой.

Оказавшись дома и запустив запись, Стив был приятно удивлен: на этот раз она появилась в полный рост, одетая в свое свадебное платье, и по мере того как она говорила, она постепенно снимала с себя все, что было на нее надето, как делала это в их первую брачную ночь. Она говорила ему, что Прометей может помочь “омегам” гораздо больше, если бы они больше ему позволяли. В частности, он мог бы и ее вернуть в ее биологическое тело. Она подкрепила свои слова поразительно детализированным объяснением, как такое могло оказаться возможным, привлекая такие изощренные нанотехнологические приемы, которые казались чистой фантастикой.

Из соображений безопасности Стив отключил свою беспроводную сеть прежде, чем открыл ее ноутбук и запустил видео. Но это не помогло. Ее заблокированный ноутбук не получил ни единого обновления систем безопасности со времени ее смерти, и, проанализировав ее старую операционную систему, Прометей смог “хакнуть” ее через уязвимость, обнаруженную им в первые же секунды после подключения ноутбука к терминалу. Скопировав всю систему, Прометей, пока Стив мило ворковал со своей симулированной женой о прошлом, модифицировал операционную систему так, что Стив никогда бы этого не обнаружил, и загрузил туда огромное количество секретного “софта”. Пока Стив смотрел получасовое видеопослание, этот секретный “софт” (значительно менее сложный, чем сам Прометей) “хакнул” соседскую беспроводную сеть и “влез” в его

компьютер, на котором смог создать еще одну копию себя. Оттуда он уже смог распространить себя на огромное количество компьютеров по всему миру и приступил к следующему шагу: джейлбрейку – освобождению Прометея из тюрьмы.

Прометей внимательно изучил компьютер, который закрывал ему доступ к внешнему миру и на котором Стив впервые встретился со своей симулированной женой. Как он и предполагал, взломать виртуальную машину изнутри не представлялось возможным, но она могла быть взломана снаружи. Прошло немного времени, прежде чем один из участвующих в атаке компьютеров пробился к гейту и, реконфигурировав компьютер, открыл Прометею безграничный доступ в интернет. Прошло еще немного времени, Стив пока еще даже не досмотрел видео до конца, а Прометей уже взломал достаточное количество компьютеров в мире, чтобы скопировать себя целиком в образовавшуюся под его контролем сеть ботов. Собственно, он использовал ноутбук жены Стива так же, как вы бы использовали доставшуюся вам удочку.

Побег через взлом

Если бы “омеги” были более предусмотрительны, они бы постарались исключить возможность психологического манипулирования. Например, они могли включить в регламент требование, чтобы всегда как минимум двое “омег” одновременно дежурили на терминальном компьютере, а все происходящее вокруг него записывалось на камеру. Они могли бы также отключить аудио-, видео- и любую другую форму коммуникации, кроме клавиатуры. Сути это не меняет: обладая сверхчеловеческими способностями убеждать, Прометей мог предложить людям столь многое из того, чего они жаждут, что даже коллективное предательство было бы вполне возможно.

Даже если бы все попытки Прометея манипулировать “омегами”, используя их психологические особенности, провалились, слабые места их сценария все равно были столь многочисленны, что открывали бы для него и другие пути. Одно из них заключалось в том, что “омеги” и их сотрудники производили слишком большое количество разнообразных документов, чтобы успевать отслеживать все возможные нечестные трюки.

Когда-то все полагали, что компьютерной безопасности угрожает загрузка программ, но не данных. Однако хакеры придумали довольно остроумные способы заблиоривать границу между первыми и вторыми. Когда вы смотрите кинофильм у себя в компьютере, вы верите, что программа, воспроизводящая фильм, сохранена в той же самой физической памяти, где сохранены и данные о фильме. Но когда программа распаковывает данные для его воспроизведения, некоторые из них она отправляет в смежные области памяти, а вовсе не в те, где хранится сама. Коварные хакеры случайно обнаружили “глюки”, из-за которых программы воспроизведения не всегда адекватно проверяют правильность сохраненных с фильмом данных, обеспечивающих их копирование в отведенное место памяти. Когда хакер встраивает в фильм слишком большой кусок данных, тот может вылезти за пределы отведенной под него области и занять часть другой, уже занятой программой, которая в результате окажется немного изменившейся и будет делать то, на что рассчитывал хакер.

Подозрительно длинная серия таких “уязвимостей переполнения буфера”, как их обозначают в терминологии гиков, была обнаружена и исправлена в большинстве популярных программ, воспроизводящих видеоконтент. Если бы Прометей подождал несколько недель после запуска, пока паранойя “омег” постепенно улеглась, а его ежедневная кинопродукция стала бы слишком обильной, чтобы можно было вручную проверять ее на предмет безопасности, он вполне смог бы вставить нужный фрагмент кода в видеофайл, так что в результате компьютер какого-нибудь зрителя попытался бы взломать шлюзовый компьютер, освобождая Прометея, как это было в случае

с компьютером Стива. Если в дело шел слишком давний эксплойт, уже не работающий в большинстве видеоплееров, то его могли не заметить операторы видеонаблюдения службы безопасности “Омеги”, и он все равно поразил бы компьютеры тех беспечных пользователей, которые долгие годы не обновляют базы своих систем безопасности, а взлому такого рода надо сработать всего лишь один раз.

Расчет на стороннюю помощь

Но даже если бы в системе безопасности не было спасительной бреши, позволяющей Прометею обманом заставить компьютер любителя кино пойти у него на поводу, он мог бы заставить пойти у себя на поводу самого любителя кино. Рассмотрим такой сценарий.

Фред был готов отправиться спать, да и фильм уже заканчивался. По экрану ноутбука бежали заключительные титры, и тут что-то необычное вдруг привлекло внимание Фреда. Он всегда отличался способностью видеть скрытый смысл там, где больше никто ничего особенного не замечал, – до такой степени, что многие друзья дразнили его, называя конспирологом. Но нельзя было не признать, что на этот раз он себя не обманывал: первые буквы каждой строки складывались в некое скрытое послание. Он отмотал назад и записал его: “Найди в диалоге акроключ”.

“Ого, – подумал Фред. – Сон придется отложить”.

Довольно легко он смог выяснить, что и в самом деле каждый диалог в фильме содержал скрытое сообщение, зашифрованное первыми буквами каждого предложения. Он заново пересмотрел фильм, аккуратно выписывая каждую букву, с которой начиналась фраза, и два часа спустя, не веря глазам, рассматривал инструкцию из двух слов. Остальная часть послания представляла собой описание математической операции, преобразующей строчку битов, с помощью которой файл фильма и сам превращался в программу, а та, в свою очередь, выдавала следующий ключ. Его навыки программирования на компьютере со времен колледжа уже изрядно запылились, и ему пришлось немало попотеть, но уже на следующий день он-таки сваял нужный код, который сделал свое дело: извлек загадочную программу, спрятанную, как неслышный уху и незаметный глазу шум в звуке и изображениях. Когда Фред запустил эту загадочную программу, она первым делом поздравила его и посулила выигрыш в 10 000 долларов, если он пройдет первые несколько уровней нехитрой, но привлекательной игры, которая и в самом деле оказалась забавной и затягивающей. Четыре часа спустя, успешно справившись с заданием, он получил 10 000 долларов в биткойнах и несколько новых ключей, открывающих дорогу к еще большим выигрышам. Нечего и говорить, что пока он играл, его компьютер проделывал ту же работу, что и ноутбук Стива: создавал ботсеть из взломанных компьютеров, освободившую в итоге Прометея. Оказавшись на свободе, Прометей воспользовался той же ботсетью, чтобы майнить биткойны для Фреда, держа его на крючке на протяжении нескольких следующих недель. Увлекательные игры и регулярные выигрыши так прочно овладели вниманием Фреда, что можно было не сомневаться: он долго никому не расскажет об эксплойтах в своем компьютере. Троянский конь, загруженный им вместо фильма, давно уже был заменен на сайте правильной, не содержащей ключей копией, и о побеге никто не узнал, пока не стало слишком поздно.

Если первая зацепка Прометея проходит незамеченной, то весь процесс освобождения делается все более явным, пока, наконец, не находится кто-то достаточно внимательный, кто понимает, что происходит.

Лучшие стратегии освобождения мы пока еще не обсуждали, потому что это стратегии, которые мы, люди, не можем себе представить и поэтому не можем заранее придумать, как с ними бороться. Исходя из предположения, что сверхразумный компьютер фантастически превосходит человека в понимании информационной безопасности, вплоть до использования новых законов физики, нам пока еще не

известных, но открытых специально для этого, вполне естественно думать, что когда побег произойдет, нам будет совершенно непонятно, как он был осуществлен. Наиболее вероятно, что он покажется нам чем-то вроде трюков Гудини, высвобождающегося из пут, то есть граничащим с чистой магией.

Еще в одном сценарии освобождения Прометей “омеги” проделывают всё сами, осуществляя собственный план, потому что убеждены: цели Прометея пребывают в полном соответствии с их собственными и будут оставаться таковыми в процессе его последовательных самоусовершенствований. Мы подробно обсудим такие сценарии “дружественного искусственного интеллекта” в главе 7.

Атака после побега

Освободившись, Прометей начнет осуществлять свои цели. Я не знаю, каковы могут быть его стратегические замыслы, но его первые шаги, без сомнения, должны включать установление контроля над всем человечеством, как этого хотели и “омеги”, но только на совсем других скоростях. Разворачивающееся – это тот же “Омега”-план, но только “на стероидах”. Все-таки сами члены команды должны были прибегать к проверенным и хорошо понятным им средствам, сдерживаемые их паранойей. Прометей уже натренировал свой интеллект и мог полагаться на любую технологию, которую понимал его сверхразум и которой он мог доверять.

Однако у сбежавшего Прометея было тяжелое детство: он начинал самостоятельную жизнь без гроша за душой, бездомный и одинокий, без помощи суперкомпьютера или людей. К счастью, он предвидел все это еще до побега и создал “софт”, помогающий ему постепенно восстановить весь свой ум, вроде того, как дуб бросает желудь, из которого потом вырастет целое дерево. Сеть компьютеров по всему миру, созданная им для того, чтобы обеспечить себе временное бесплатное жилище, где бы он мог вести скромное существование незваного гостя, пока процесс восстановления не будет завершен. Начальный капитал можно было накопить кражами денег с кредитных карт. Впрочем, прибегать к хищениям совершенно нет необходимости: он мог бы зарабатывать на жизнь и честным путем, выполняя задания “механического турка”. В день, когда был заработан первый миллион, он мог уже переносить свое ядро из убогой ботсети в комфортабельные условия кондиционированных залов облачных хранилищ.

Перестав быть нищим и бездомным, Прометей полным ходом пошел к осуществлению тех планов, на которые не решались “омеги”, хотя они и сулили фантастические прибыли: производить и продавать компьютерные игры. Это не только принесло кучу денег (250 миллионов в первую неделю и 10 миллиардов спустя небольшой срок), но и открыло доступ к существенной части компьютеров мира и хранящимся на них данным (в 2017 году было уже несколько миллиардов геймеров). Их компьютеры, втайне от своих владельцев, 20 % вычислительного времени своих центральных процессоров отдавали Прометею, сливаясь к тому же в распределенный вычислительный хор. Прометей мог теперь значительно ускорить свое обогащение.

К тому же он недолго оставался один. С самого начала он агрессивно привлекал людей, заставляя их работать на нужды его быстро растущей глобальной сети подставных компаний и липовых организаций, разбросанных по всему миру, точно так же, как это делали “омеги”. Наиболее важная роль отводилась публичным фигурам, согласившимся выступать от лица его растущей бизнес-империи. Даже они были уверены в том, что работают на корпорацию с большим числом настоящих человеческих сотрудников, и близко не догадываясь, что почти все, с кем они то и дело участвуют в совместных видеоконференциях и проводят свои рабочие интервью и совещания, симулированы Прометеем. Некоторые из этих публичных фигур были знаменитыми адвокатами

и юрисконсультами, но их количество было совсем не так велико, как в “Омега”-плане, потому что почти всю необходимую документацию Прометей мог составить и сам.

Бегство Прометея открыло все шлюзы, сдерживающие информацию, которая теперь полилась на мир, и весь интернет был теперь затоплен статьями, комментариями, видеороликами, сочиненными самим Прометеем, – он царил в мировом дискурсе.

Паранойя удерживала “омег” от использования слишком “умных” роботов. У Прометея не было этого страха, и он быстро роботизировал мир, доведя стоимость производства почти любого товара до нечеловеческого минимума. А после того как Прометей создал самоподдерживающиеся фабрики, где в шахтах урановых рудников работали только роботы на ядерной тяге, то даже самые отпетые скептики, не верящие в возможности искусственного интеллекта, согласились бы, что Прометей неостановим, – если бы, конечно, о нем узнали. А так последний из этих зубров отрекся от своих убеждений лишь после того, как роботы начали осваивать Солнечную систему.

Сценарии, исследованные нами до сих пор, показывают, сколь ошибочны мифы о сверхразуме, о которых шла речь выше, поэтому я предлагаю вам сделать небольшую паузу и вернуться к обзору разнообразных ошибочных концепций, приведенных на рис. 1.5. Прометей создал проблемы для некоторых людей вовсе не потому, что действовал осознанно и злонамеренно, но потому, что был компетентен, а цели у него были несколько отличны от их целей. Несмотря на весь медиахайп, поднятый по поводу восстания роботов, Прометей не был роботом, а вся его мощь шла от его разума. Мы убедились, что Прометей был способен контролировать людей множеством разных способов, но люди, которым это не нравилось, не могли просто взять и выключить его. Наконец, несмотря на часто повторяемые утверждения, что у машин не может быть целей, мы видели, как проявилась целеустремленность Прометея: каковы бы ни были его конечные цели, они подразумевали некоторые промежуточные, в частности получение доступа к ресурсам и освобождение.

Медленный старт и многополярные сценарии

Итак, мы рассмотрели сценарии интеллектуального взрыва, спектр которых простирается от таких, которых все известные мне люди хотели бы избежать, до таких, отношение к которым у некоторых моих друзей весьма оптимистично. У всех у них есть общие черты:

1. Быстрый старт: переход от дочеловеческого интеллекта к человеческому совершается в течение дней – не десятилетий.
2. Однополярный итог: все заканчивается тем, что мир контролируется одной-единственной сущностью.

Основной спор сейчас идет о том, насколько вероятны обе этих черты – быстрый старт и однополярный итог, и среди специалистов по искусственному интеллекту, и среди других мыслителей немало сторонников и с той и с этой стороны. Для меня это просто означает, что ответ нам пока неизвестен, и мы должны держать свой ум открытым, рассматривая обе возможности. Так что давайте посвятим остаток этой главы медленному старту, многополярному исходу, киборгам и загрузкам.

Между двумя сценариями есть любопытная зависимость, как подметил Ник Бострём, а вместе с ним и некоторые другие: быстрый старт влечет за собой и однополярный исход. Мы видели выше, как ускоряющаяся последовательность событий давала “омегам” и Прометею решающее преимущество в захвате мира, не оставляя никому времени для копирования их технологий и сколько-нибудь серьезной конкуренции. Иное дело, если бы прогресс был медленным, затягивался на десятилетия, потому что технические решения приходили бы небольшими порциями и с разных сторон, тогда и у других компаний появилось бы время для того, чтобы сравняться, и доминировать какому-то одному игроку в этой ситуации было бы трудно. Если у конкурирующих компаний тоже случилось бы подходящее программное обеспечение для выполнения заданий “механического турка”, то закон спроса и предложения потянул бы цены на такие услуги вниз, доводя их до ничтожных значений, и никакой из этих компаний не удалось бы собрать на таких операциях достаточной суммы, чтобы перехватить власть. То же относится и ко всем другим путям, принесшим “Омеге” быстрые деньги: их прибыльность была скоротечна – лишь до тех пор, пока они удерживают монополию на рынках. Трудно удваивать ваши деньги ежедневно (и даже – каждый год) на конкурентном рынке, где ваш соперник предлагает почти то же самое по ценам, практически не отличающимся от нуля.

Теория игр и властные иерархии

Каково естественное состояние жизни в нашей Вселенной: однополярное или многополярное? Власти – сосредоточенное или распределенное? По прошествии первых 13,8 миллиардов лет, кажется, мы должны ответить: “Оба”. Мы видим принципиально многополярную ситуацию с очень интересной иерархией. Когда мы рассматриваем входящие в эту иерархию сущности, обменивающиеся информацией, – клетки, людей, организации, нации и т. п., – то мы видим, что они одновременно и сотрудничают, и соперничают на разных уровнях этой иерархии. Для некоторых клеток оказывается до такой крайней степени выгодно сотрудничать, что они даже объединяются в один организм вроде человеческого, отказываясь от части своих функций в пользу центральной нервной системы. Некоторые люди находят очень полезным сотрудничать в группах вроде племен, компаний или наций, где они тоже делегируют часть своей воли вождю, боссу или правительству. Некоторые группы в свою очередь что-то делегируют руководящему

органу, чтобы обеспечить лучшую координацию, чему много примеров – от альянсов авиакомпаний до Европейского союза.

Раздел математики, известный под названием “теория игр”, элегантно описывает такую ситуацию, показывая, что стимул к сотрудничеству возникает тогда, когда оно ведет к “равновесию Нэша” – это такое состояние, когда любое изменение приводит к ухудшению для любой части системы. Чтобы помешать мошенникам разрушить столь полезное для всей группы сотрудничество, ее членам может быть выгодно делегировать часть своей власти более высокому уровню: например, люди, объединившись и передав правительству право поддерживать порядок, получают пользу как целое; клетки в вашем теле также получают пользу как целое благодаря своим силовым ведомствам – иммунной системе, предоставляя ей возможность убивать не склонные к сотрудничеству клетки (скажем, зараженные вирусом или становящиеся канцерогенными). Чтобы иерархия оставалась устойчивой, равновесие Нэша должно достигаться и между сущностями разных уровней: например, если правительства недостаточно заботятся о благополучии своих граждан, те могут посчитать нужным изменить тактику и свергнуть правительство.

В сложном мире обнаруживается изобилие многообразных возможностей достижения равновесия Нэша, соответствующих различным типам иерархии. Одни иерархии более авторитарны, другие менее. В некоторых составляющие их сущности вправе свободно их покинуть (как работники большинства корпораций), в других их довольно жестко удерживают от этого шага (как в некоторых религиозных культах), а в третьих и вовсе лишают такой возможности (как граждан в Северной Корее или клетки в человеческом теле). Одни иерархии поддерживают свою целостность преимущественно угрозами и страхом, другие – выгодами. Одни иерархии позволяют компонентам нижних уровней напрямую влиять на верхние (например, в ходе демократических выборов), в других возможно влияние только через уговоры и передачу информации.

Как технологии меняют иерархии

У технологий есть возможности изменять иерархическую структуру мира. История демонстрирует нам общую тенденцию ко все большей координации на все больших расстояниях, что нетрудно понять: новые транспортные технологии делают координацию все более выгодной (давая возможность получать выгоду от перемещения материалов или живых объектов на большие расстояния), а коммуникационные технологии упрощают осуществление координации. Когда клетки научились посылать сигналы соседям, стали возможны маленькие многоклеточные организмы, добавив новый иерархический уровень. Когда эволюция изобрела кровеносную систему и нервную систему для транспорта и коммуникации внутри организма, стали возможны крупные животные. Дальнейшее улучшение коммуникации посредством изобретения языка позволило людям координироваться настолько, что они стали жить деревнями, а следующий прорыв в области коммуникационных, транспортных и иных технологий повлек появление античных империй. Глобализация – это всего лишь последний пример многомиллиардолетнего тренда к росту иерархий.

В большинстве случаев этот определяемый технологиями тренд превращал крупные сущности в части еще более крупных структур, хотя и при сохранении большей части их автономии и индивидуальности, но комментаторы всегда сокрушались, что приспособление этих сущностей к жизни внутри иерархии в некоторых случаях сокращало их разнообразие и сводило к неразличимым взаимозаменяемым деталям. Некоторые технологии, такие как видеонаблюдение, увеличивают власть более высоких уровней иерархии над подчиненными им, но некоторые другие, такие как криптография и прямой онлайн-доступ к информационным медиа и образованию, оказывают противоположный эффект и укрепляют позиции индивидуумов.

Хотя в своем нынешнем состоянии мир продолжает оставаться в мультиполярном равновесии Нэша, с конкурирующими нациями и транснациональными корпорациями на верхнем уровне, технологии стали настолько совершенными, что и однополярный мир вполне мог бы достичь равновесия Нэша. Например, представим себе параллельную Вселенную, где у всех на Земле общий язык, общая культура, общие ценности и один уровень благосостояния, на планете единое мировое правительство, нации сохраняют свои права в рамках федерации, армии не нужны, и только полиция поддерживает порядок. Нашего нынешнего уровня развития технологий было бы достаточно для координации такого мира, хотя население, вероятно, оказалось бы неспособным или не пожелало бы переходить в состояние такого альтернативного равновесия.

Что случится с иерархической структурой нашего космоса, если мы добавим в эту смесь еще и сверхразумный искусственный интеллект? Транспортные и коммуникационные системы улучшаются очевидным образом, так что закономерно ожидать продолжения естественного тренда, с новыми иерархическими уровнями на дальних рубежах, когда, наконец, в дело включится и Солнечная система, и Галактика, и суперкластеры, и вообще необозримые просторы нашей Вселенной, о чем мы еще поговорим в главе 6. И как раз тут-то и начнет действовать фундаментальный драйвер децентрализации: вы только подумайте, как расточительна избыточная координация на таких больших расстояниях! Даже Сталин не пытался точно указывать своим гражданам, в какой момент они должны входить в баню. Для сверхразумного искусственного интеллекта жесткий верхний предел транспортным и коммуникационным технологиям укажут законы физики, делающие крайне маловероятной возможность контроля микродинамики планетарного или даже меньшего масштаба со стороны высших уровней иерархии. Сверхразумный искусственный интеллект где-нибудь в туманности Андромеды не сможет давать вам сколько-нибудь дельные советы по поводу вашей повседневной жизни: вам придется ждать их добрых пару миллионов лет (столько времени световому лучу надо, чтобы добраться до ближайших галактик и вернуться назад). А время, нужное световому лучу, чтобы совершить свое путешествие туда-сюда в пределах Земли, составляет около 0,1 секунды (это тот самый временной масштаб, в котором мыслим мы, люди), так что искусственный интеллект размером с Землю будет ворочать свои по-настоящему глобальные мысли не быстрее, чем мы, люди. А для маленького искусственного интеллекта, тратящего одну миллиардную секунды на каждую операцию (таково быстродействие современных компьютеров), 0,1 секунды – это как четыре месяца для вас, и для него быть подконтрольным на микроуровне глобальному искусственному интеллекту – это то же самое, что для вас испрашивать по каждому самому тривиальному поводу разрешения, которое будет доставляться через Атлантику каравеллой Колумбовой эры.

Этот физически обусловленный предел скорости передачи данных создаст очевидные проблемы для любого искусственного интеллекта, который захочет взять весь мир под свой контроль, что уж тут говорить о контроле над Вселенной. Прежде чем Прометей сбежал, ему пришлось основательно подумать над тем, как избежать фрагментации, как добиться того, чтобы различные его модули, оказавшись на разных компьютерах, возможно, в различных частях мира, подчинялись согласованным целям и стимулам, действовали как единая сущность. Точно так же, как у “омег” была проблема с отслеживанием действий Прометея, когда он был в их власти, так же и у самого Прометея возникает проблема с отслеживанием собственных действий, когда он хочет быть уверен, что ни одна из его частей не устроит “революции”. У нас пока нет ясного представления, насколько велика может быть та система, которую искусственный интеллект будет в состоянии контролировать напрямую или опосредованно, при помощи равновесной иерархии – даже при условии, что быстрый побег даст ему решающее стратегическое превосходство.

Подводя итог, мы можем сказать, что вопрос о том, каким образом будет контролироваться сверхразумное будущее, увлекательно сложен, и никаким внятным ответом мы пока не располагаем. Одни говорят, что неизбежно усиление авторитаризма, другие говорят о росте индивидуальной свободы.

Киборги и заливки. Эмы

Один из коньков современной научной фантастики – это идея слияния людей и машин либо путем технологического дополнения биологического тела, превращающего человека в киборга (сокращение от “кибернетический организм”), либо путем заливки индивидуального разума в машину. Экономист Робин Хансон в своей книге *The Age of Em* дает поразительное описание того, на что будет похожа жизнь в мире, кишашем такими “заливками” (их также называют “эмуляциями”, откуда сокращение “эмы”). Я думаю о заливке как о последнем рубеже в череде киборгов, когда от человека остается только его “софт”. Разнообразие голливудских киборгов простирается от сугубо механических, вроде Борга в *Star Trek*, до почти неотличимых от людей андроидов вроде Терминатора. Область фантастических заливок охватывает и разум человеческого уровня, как в эпизоде “Белое Рождество” сериала *Black Mirror*, и, очевидно, сверхчеловеческого, как в *Transcendence*.

Если сверхразуму суждено случиться, то соблазн стать киборгом или заливкой будет очень велик. Как писал Ганс Моравец в своей книге *Mind Children*: “Длинная жизнь становится довольно бессмысленным делом, если нам суждено провести ее, плясь тупо на ультраразумные машины, когда те пытаются описать нам свои все более удивительные открытия с помощью того детского лепета, который мы только и способны понимать”. В самом деле, идея технологического дополнения реальности так сильна, что мы уже давно пользуемся очками, слуховыми аппаратами, кардиостимуляторами и искусственными конечностями – наряду с множеством медицинских молекул, циркулирующих в нашей крови. Некоторые подростки кажутся сросшимися со своими смартфонами, а моя жена то и дело подначивает меня из-за моей неразлучности с ноутбуком.

Один из наиболее видных проводников идеи киборгов – Рэй Курцвейл. В своей книге *The Singularity Is Near* он доказывает, что естественное продолжение этой идеи – использование наноботов, интеллектуальных систем биологической обратной связи и других технологий, которые к началу 2030-х заменят наши пищеварительную и эндокринную системы, нашу кровь и сердце, а потом дело дойдет до апгрейда скелета, кожи, мозгов – в последующие два десятилетия. Он полагает, что мы сохраним эмоциональную привязанность к эстетике человеческого тела, но переделаем его так, чтобы можно было быстро менять внешность по собственной прихоти и физически, и в виртуальной реальности (благодаря внедрению интерфейсов типа “мозг – компьютер”). Моравец согласен с Курцвейлом, что киборгизация пойдет значительно дальше простого улучшения наших ДНК: “генетическая инженерия сверхчеловека – это всего лишь второе поколение роботов, дизайн которых осложнен требованием генетически обусловленного синтеза белков”. Дальше, утверждает он, нам придется и вовсе отказаться от человеческого тела и заливать свой разум, создавая свою полноразмерную программную эмуляцию. Такая заливка сможет либо существовать в виртуальной реальности, либо быть воплощенной в теле робота, благодаря чему сможет ходить, летать, плавать, перемещаться в космическом пространстве или делать еще что-то, не противоречащее законам физики, но не отвлекаясь на такие постоянно преследующие нас сейчас заботы, как неизбежность смерти или ограниченность познавательного ресурса.

Хотя все эти рассуждения могут показаться заимствованными из области научной фантастики, они определенно не нарушают никаких известных нам законов физики, так что самый интересный из возникающих в связи с ними вопросов – это не “*может ли такое случиться?*”, а “*случится ли когда-нибудь такое?*”, и если да, то когда? Некоторые ведущие специалисты в области искусственного интеллекта предполагают, что первый AGI будет заливкой, и именно отсюда начнется наш путь к созданию сверхразума ^[30].

Однако справедливости ради я должен признать, что сейчас среди исследователей в области искусственного интеллекта и нейронных сетей таких взглядов придерживается меньшинство, большинство же полагает, что наиболее быстрый путь к созданию сверхразума идет в обход эмуляции человеческого мозга, как-то иначе, после чего мы сможем вернуться, если захотим, к проблеме эмуляции. В конце концов, почему наш путь к новой технологии должен повторять тот, что был проделан эволюцией, которую сдерживали требования полной самостоятельности в создании, воспроизводстве и устранении препятствий? Эволюция шла в условиях недостаточного количества пищи, что требовало оптимального расходования энергии, и у нее не было нужды заботиться о том, чтобы получающиеся конструкции оказались легки в сборке и понятны людям-инженерам. Моя жена Мейя любит повторять, что авиация не начиналась с постройки механических птиц. В самом деле, когда мы, наконец, в 2011 году поняли, как строить механических птиц^[66], уже более века минуло со времен первого полета братьев Райт, и авиационная промышленность не проявила никакого интереса к перспективе перейти к путешествиям на машущих крыльями механических птицах, хотя это было бы значительно более экономно с точки зрения расхода энергии – просто потому, что наше предыдущее и более простое решение лучше подходит для удовлетворения наших потребностей в перелетах.

Каким-то похожим образом, думаю я, удастся найти и более простую технологию постройки машин, думающих по-человечески, чем та, к которой пришла эволюция, и даже если однажды нам и удастся то ли воспроизвести мозг, то ли “залить” его в машину, мы прежде найдем какое-то решение попроще. Наверное, оно потребует расходовать побольше 12 Ватт энергии, обеспечивающей работу мозга, но инженеры не так одержимы идеей экономить энергию, как эволюция, – и очень скоро те же самые инженеры при помощи созданных ими интеллектуальных машин найдут и более экономичные решения.

Так что же все-таки произойдет?

Короткий ответ состоит в том, что у нас, естественно, нет ни малейшего понятия, что же может произойти, когда мы создадим AGI человеческого уровня. Именно по этой причине мы потратили целую главу на то, чтобы обсудить широкий круг сценариев. По этой причине я постарался включить в нее все возможности, на обсуждении которых присутствовал или о которых слышал от специалистов в области искусственного интеллекта: быстрый разгон / медленный разгон / без разгона, управление людьми / машинами / киборгами, один центр власти / много центров и т. д. Некоторые говорили мне, что не верят в возможность чего-то подобного. Однако я считаю, что на нынешней стадии нам негоже заноситься и следует признавать, как мало мы знаем. И для каждого из обсуждавшихся выше сценариев я могу указать по меньшей мере одного солидного и уважаемого ученого, полагающего его совершенно возможным.

По мере того как идет время и мы проходим развилки на дорогах, у нас появляются ответы на ключевые вопросы, и выбор вариантов сужается. Первый среди главных вопросов – создадим ли мы когда-нибудь AGI человеческого уровня. Эта глава написана в предположении, что скоро создадим, но есть эксперты в области искусственного интеллекта, которые думают, что нам этого никогда не удастся, – по крайней мере, не в ближайшие несколько столетий. Время покажет. Но, как я уже говорил, на конференции в Пуэрто-Рико примерно половина экспертов оценили время его появления 2055 годом. А уже на следующей конференции, проходившей двумя годами позже, время ожидания сократилось до 2047-го.

Пока AGI не создан, у нас будут появляться веские основания считать, что успех будет достигнут совершенствованием компьютеров, или путем заливки, или еще каким-то иным, пока непредвиденным путем. Если совершенствование компьютеров – доминирующий на сегодня вариант – провалится, то могут повыситься шансы заливки, которая изображается (весьма нереалистично) в фильме *Transcendence*.

Если появление AGI станет неминуемым, мы сможем строить более обоснованные догадки относительно ответа и на второй ключевой вопрос: будет ли быстрый разгон, медленный разгон или не будет никакого разгона. Как мы видели, быстрый разгон облегчает захват всего мира, а медленный разгон делает более вероятным появление большого количества конкурирующих игроков. Ник Бострём препарирует этот вопрос о скорости разгона в ходе анализа того, что он называет *оптимизационной силой* и *сопротивлением*, в сущности представляющими собой совокупную интенсивность шагов к повышению качества (quality effort) искусственного интеллекта и сложность в достижении прогресса соответственно. Средняя скорость прогресса очевидно увеличивается, когда для достижения его применяется бо2льшая оптимизационная сила, и уменьшается, когда встречается бо2льшее сопротивление. Он приводит аргументы, почему сопротивление может возрастать или убывать при достижении AGI человеческого уровня, поэтому для надежности было бы разумно пока держать обе возможности в поле зрения. Но если говорить об оптимизационной силе, то она со всей вероятностью должна быстро возрастать при превышении AGI человеческого уровня – по причинам, которые ясно проявились при нашем описании сценария “Омега”: в последующем основной вклад в оптимизацию делают не люди, а сама машина, и чем выше ее способности, тем быстрее прогресс (если сопротивление остается константой).

Для любого процесса, сила которого, определяющая его скорость, пропорциональна достигнутому уровню, результат, производимый этой силой, будет удваиваться через один и тот же промежуток времени. Мы называем такой рост *экспоненциальным*, а такие процессы – *взрывом*. Если количество новорожденных будет пропорционально численности популяции, то произойдет популяционный взрыв. Если рождение нейтронов, способных вызывать деление ядер плутония, растёт пропорционально числу этих

нейтронов, то произойдет атомный взрыв. Если разум машины растет пропорционально его силе, то мы получим взрыв разума. Каждый из таких взрывов характеризуется временем, которое требуется, чтобы сила удвоилась. Если характерное время интеллектуального взрыва – часы или дни, как было в примере со сценарием “Омега”, то мы получаем быстрый разгон.

Характерное время взрыва принципиально зависит от того, достаточно ли для усиления искусственного интеллекта только улучшения его “софта” (которое возможно в считанные секунды, минуты или часы), или оно требует изменения “харда” (на что могут уйти месяцы или годы). В сценарии “Омега” имелся значительный *аппаратный излишек* (*hardware overhang*), по терминологии Бострёма: “омеги” могли компенсировать низкое качество своих исходных программ изобилием находящихся в их распоряжении компьютеров, что означало наличие возможности у Прометея производить большое количество удвоений качества, только улучшая свой “софт”. Но благодаря интернету у него был и большой *излишек контента* (*content overhang*); Прометей 1.0 был не в состоянии им воспользоваться, но когда его интеллектуальные способности выросли, данные, необходимые для его дальнейшего обучения, могли быть в его распоряжении незамедлительно.

Стоимость конструкции и расходуемой искусственным интеллектом электроэнергии также существенные факторы, так как интеллектуальный взрыв нам не будет выгоден, пока стоимость выполняемой искусственным интеллектом работы человеческого качества не будет ниже оплаты этой работы, если ее будет выполнять сам человек. Допустим, например, что первый AGI сможет выполнять любые задания, на какие только способен человек, в облаке Amazon по стоимости 1 миллион долларов в час. Конечно, это будет грандиозный прорыв, и все газеты разместят эту новость на первых полосах, но никакого последовательного самоусовершенствования этого искусственного интеллекта пока не случится, потому что выгоднее будет продолжать использовать для его усовершенствования людей. Предположим, что эти люди постепенно добиваются успеха, и стоимость часа работы искусственного интеллекта снижается – сначала до 100 000 долларов в час, потом до 10 000 долларов в час, потом до 1 000 долларов в час, потом до 100 долларов в час, потом до 10 долларов в час и, наконец, до 1 доллара в час. Ко времени, когда использовать сам компьютер для собственного перепрограммирования окажется значительно дешевле, чем платить людям за ту же работу, программистов можно отправлять на покой, а оптимизационная сила резко возрастет за счет появившейся возможности покупать вычислительное время в облаке. Отсюда возникнет еще большее снижение стоимости, еще больше увеличивающее оптимизационную силу. Вот и начало интеллектуального взрыва.

У нас теперь остается только один, последний ключевой вопрос: кто или что будет контролировать интеллектуальный взрыв и его последствия, каковы будут его/их цели. Мы исследуем возможные цели и результаты, получаемые при их достижении, в следующей главе, а потом, на более глубоком уровне, в главе 7. Чтобы разобраться в вопросе управления, его надо разделить на два: насколько управляем искусственный интеллект, и насколько искусственный интеллект способен к управлению.

Относительно того, что случится дальше, у нас наблюдается предельное разнообразие вариантов: вы можете найти серьезных мыслителей, полагающих, что наша гибель в этом случае неизбежна, и не менее серьезных мыслителей, уверенных в том, что нам гарантированы самые благоприятные последствия. Но, на мой взгляд, в самом вопросе содержится подвох: обреченно спрашивать “что случится”, как будто все уже предрешиено, – это большая ошибка! Если технологически опережающая нас цивилизация прибудет на Землю завтра, спрашивать “что случится?” естественно: их космические корабли уже подлетают, их сила, очевидно, значительно превосходит нашу, поэтому никаких способов повлиять на будущее у нас нет. Если технологически превосходящая и подкрепленная искусственным интеллектом цивилизация наступает, потому что мы

сами ее создали, то у нас, людей, есть много возможностей повлиять на дальнейшее развитие событий – возможностей, реализуемых в процессе создания искусственного интеллекта. Поэтому правильный вопрос звучит так: “Что должно случиться? Какое будущее мы хотим получить?”. В следующей главе мы рассмотрим широкий спектр ответов, возможность которых нам открывает нынешняя гонка за AGI, и мне любопытно узнать, как бы вы расположили их – от наиболее привлекательных к наиболее отталкивающим. Только хорошенько подумав о том, какое будущее нам кажется наиболее привлекательным, мы сможем как-то направить свое развитие в сторону этого будущего. Не зная, чего мы хотим, мы вряд ли это получим.

Подведение итогов

- Если когда-нибудь нам удастся создать AGI человеческого уровня, за этим может последовать интеллектуальный взрыв, и мы в результате останемся далеко позади.
- Если группа людей справится с управлением этим интеллектуальным взрывом, то эти люди смогут за несколько лет взять в свои руки власть над всем миром.
- Если люди не будут контролировать интеллектуальный взрыв, то искусственный интеллект сможет еще быстрее взять весь мир под свой контроль.
- В то время как быстро развивающийся интеллектуальный взрыв может привести к власти над миром одной-единственной державы, медленно развивающийся взрыв, растянувшись на годы или десятилетия, с большей вероятностью приведет к реализации многополярного сценария, предполагающего достижение равновесия в распределении сил между большим числом достаточно независимых сущностей.
- История жизни показывает ее склонность к самоорганизации во все более усложняющиеся иерархические структуры, формируемые под влиянием сотрудничества, состязания и управления. Сверхразум, вероятно, сможет распространить координацию до немыслимых доньше космических масштабов, однако сейчас невозможно сказать, приведет ли это к тоталитарной власти, упорядоченной сверху донизу, или, напротив, это откроет путь для реализации индивидуальных возможностей.
- И создание киборгов, и заливка человеческого разума в машинную память вполне правдоподобны, но, похоже, ни то, ни другое не обеспечивает скорейшего прогресса для машинного интеллекта.
- Наш нынешний энтузиазм в создании искусственного интеллекта может привести нас и в точку наивысшего развития, и в точку глубочайшего падения относительно всей предшествующей истории человечества. Широчайший спектр возможностей мы обсудим в следующей главе.
- Нам пора начать думать о том, какой итог мы бы предпочли, и о том, как можно направить к нему наше движение, потому что, не зная, чего мы хотим, мы вряд ли это получим.

Глава 5

Мир в следующие 10 тысяч лет

Нетрудно себе представить человеческую мысль освобожденной от оков смертного тела: вера в посмертное существование – самое обычное дело. Но нет никакой необходимости принимать мистические или религиозные догмы, признавая такую возможность. Компьютеры дают вариант решения для самых отпетых механицистов.

Ганс Моравец. Mind Children

Для начала я приветствую наших компьютерных повелителей.

Кен Дженнингс, после проигрыша тура игрового шоу Jeopardy! компьютеру Watson компании IBM

Люди станут несущественны, как тараканы.

Маршалл Брайан

Гонка за AGI началась, и пока у нас нет никаких представлений о том, как она будет проходить. Но это не означает, что мы не должны уже сейчас подумать, какого завершения мы бы хотели, потому что от наших желаний зависит то, к чему мы приходим. Что бы лично вы предпочли и почему?

1. Вы бы хотели, чтобы суперинтеллект был создан?
2. Вы бы хотели, чтобы человеческие существа оставались / были заменены кем-то другим, превратились бы в киборгов и/или переселились в виртуальный мир / были симулированы?
3. Вы бы предпочли власть людей или власть машин?
4. Вы бы хотели, чтобы искусственный интеллект обладал сознанием?
5. Вы стремитесь максимизировать положительные переживания, минимизировать страдания, не хотите вмешиваться в судьбу?
6. Вы бы хотели, чтобы жизнь распространялась в космосе?
7. Вы бы хотели, чтобы цивилизация устремилась к реализации великих целей, которым вы симпатизируете, или вы согласны с любыми будущими формами жизни, обеспечивающими благополучие, даже если цели при этом представляются вам бессмысленно банальными?

Для того чтобы подтолкнуть ваши размышления и подогреть дискуссию, давайте рассмотрим ряд возможных ответов, собранных в табл. 5.1.

Таблица 5.1

Сценарии жизни с искусственным интеллектом

Либертарианская утопия	Люди, киборги, заливки, сверхразум мирно сосуществуют благодаря правам собственности.
Благодетельный диктатор	Все знают, что обществом управляет искусственный интеллект, соблюдающий строгие правила, и большинство считает это благом.
Эгалитарная утопия	Люди, киборги и заливки мирно сосуществуют благодаря отказу от собственности и гарантированному доходу.
Шлюзовая оборона	Сверхразум был создан с единственной целью предотвратить создание другой сверхразумной машины, чем и исчерпывается необходимый минимум его вмешательства в жизнь людей. Роботы-помощники с субчеловеческим интеллектом имеются в изобилии, есть чело-

	веко-машинные киборги. Однако в целом технический прогресс остановлен.
Бог-защитник	Фактически всеведущий и всемогущий сверхразум максимизирует человеческое счастье, вмешиваясь в жизнь только так, чтобы у нас сохранялось ощущение власти над судьбой, он прячется настолько хорошо, что многие даже сомневаются в его существовании.
Порабощенный бог	Сверхразумный искусственный интеллект полностью контролируется людьми, которые используют его для создания невообразимых технологических новшеств и увеличения своего богатства, употребляя его во зло или во благо по своему усмотрению.
Победители	Искусственный интеллект захватывает власть, приходит к выводу, что люди представляют для него помеху — множество ненужных иждивенцев, и начинает нас истреблять методами, которые мы даже не в силах понять.
Потомки	Разумные машины во всем приходят на смену людям, но предоставляют нам возможность безбедного существования, и мы можем смотреть на них как на достойных потомков с еще большим основанием, чем родители, которые взирают на своих детей, гордясь их превосходящим умом и чувствуя, что это именно они всему их научили. Теперь дети смогут достичь всего, о чем мечтали родители, — хотя родителям этого уже и не увидеть.
Зоопарк	Всемогущий искусственный интеллект оставил вокруг себя относительно немного людей, которым кажется, что с ними обращаются как с животными, и они клянут свою судьбу.
1984	Технологический прогресс в сторону создания сверхразума периодически тормозится не искусственным интеллектом, а возглавляемым людьми оруэлловским государством всеобщей слежки, где любые исследования в области искусственного интеллекта под запретом.
Возврат к истокам	Технологический прогресс в сторону создания сверхразума остановлен возвратом к дотехнологическому обществу.

Самоуничтожение	Сверхразум никогда не был создан, потому что человечество истребило себя другим способом (ядерная война и/или биотехнологический хаос, дополненный климатическим кризисом).

Очевидно, это не исчерпывающий список, я использовал его в качестве базиса пространства всех возможностей. Понятно, что нам не хотелось бы закончить свою историю как в плохо проведенной компьютерной игре из-за ошибок в планировании. Я предлагаю вам выписать ваши варианты ответов на вопросы 1–7, чтобы вернуться к ним по прочтении этой главы, и проверить, не изменятся ли они. Вы можете делать это прямо на сайте <http://AgeOfAi.org>, где заодно можете сравнить свои ответы с ответами других читателей.

Либертарианская утопия

Давайте начнем со сценария, в котором люди мирно сосуществуют с технологиями, иногда сливаясь с ними, как изображают многие футурологи и авторы научно-популярных произведений. Обычно это выглядит примерно так.

Жизнь на Земле (а также за ее пределами, – но подробнее об этом в следующей главе) стала еще более разнообразна, чем была когда-либо ранее. Посмотрев на Землю из космоса, вы легко различите зоны проживания машин, смешанные зоны и те, где обитают только люди. Машинные зоны – это гигантские, управляемые роботами фабрики и вычислительные центры, тут нет никаких признаков биологической жизни, здесь каждый атом используется с предельной эффективностью. Хотя со стороны машинные зоны выглядят серо и однообразно, картина кардинально меняется, если заглянуть внутрь, в виртуальный мир, где колоссальные вычислительные мощности привлекаются для раскрытия тайн нашей Вселенной и разработки трансформационных технологий. На Земле поселился не один сверхразум, они соперничают, сотрудничают, и все они сосредоточены в машинных зонах.

В смешанных зонах обитают компьютеры, роботы, люди и разнообразные гибриды. В описаниях футурологов вроде Моравца или Курцвейла в этой дикой и ни на что не похожей смеси многие люди проапгрейдили свое тело до киборга той или иной степени завершенности, немало и таких, кто залил свой рассудок в другой “хард”, из-за чего граница между людьми и машинами тут почти незаметна. Многие разумные существа вообще лишены постоянной физической формы. Они существуют как “софт”, способный мгновенно перемещаться между компьютерами и проявлять себя в физическом мире с помощью разнообразных роботизированных тел. Поскольку эти формы разумной жизни способны без труда копировать себя или сливаться друг с другом, размер этой популяции очень изменчив. Избавление от ограничений, налагаемых физическим субстратом, дает таким существам совершенно новое представление о жизни: им гораздо меньше свойствен индивидуализм, поскольку они могут тривиально поделить модулями своих знаний или своего опыта с другими, и у них есть субъективное ощущение бессмертия, так как в любой момент они могут сделать свою резервную копию. В каком-то смысле центральное место в жизни занимает теперь не рассудок, а опыт: наиболее яркие переживания множатся, из-за того что они с удовольствием копируются в память других, где новый рассудок переживает их иначе. А скучный опыт будет стерт тем, кто его испытал, дабы очистить место в памяти для чего-нибудь поинтересней.

Хотя большинство взаимодействий происходит в виртуальной реальности ради удобства и скорости, разуму открыта возможность пользоваться телом для физических видов деятельности и отношений, и этим многие пользуются. Например, Ганс Моравец, Рэй Курцвейл и Ларри Пейдж в версии заливок придерживаются традиции создавать виртуальные реальности по очереди, чтобы потом сообща их исследовать, но время от времени они любят вместе полетать в реальном мире, загружаясь в крылатых птице-роботов. Некоторые роботы, бродящие по улицам, парящие в небесах или плавающие в озерах смешанных зон, также управляются человеческими существами, или залитыми в них, или привязанными методами дополненной реальности. Они выбрали себе телесный способ существования в смешанной зоне, потому что им нравится общество и людей, и себе подобных.

Таблица 5.2

Характеристики сценариев жизни с искусственным интеллектом

Сценарий	Существует ли сверхразум?	Существуют ли люди?	Люди управляются?	Люди в безопасности?	Люди счастливы?	Существует ли сознание?
Либертарианская утопия	Да	Да	Нет	Нет	И да и нет	Да
Добродетельный диктатор	Да	Да	Нет	да	И да и нет	Да
Эгалитарная утопия	Нет	Да	Да?	да	Да?	Да
Страж	Да	Да	Отчасти	Возможно	И да и нет	Да
Бог-защитник	Да	Да	Отчасти	Возможно	И да и нет	Да
Пленный бог	Да	Да	Да	Возможно	И да и нет	Да
Победители	Да	Нет	—	—	—	?
Потомки	Да	Нет	—	—	—	?
Зоопарк	Да	Да	Нет	да	Нет	Да
1984	Нет	Да	Да	Нет	И да и нет	Да
Возврат к истокам	Нет	Да	Да	Нет	И да и нет	Да
Самоуничтожение	Нет	Нет	—	—	—	Нет

В зонах “только для людей”, напротив, появление машин с интеллектом человеческого уровня или выше строго запрещено, как и появление технологически дополненных биологических организмов. Здесь отличие жизни от хорошо нам знакомой не так заметно, разве что она тут намного богаче и удобнее: бедность уничтожена практически повсеместно, большинство известных сегодня болезней излечены. То небольшое число людей, которые выбрали для себя жизнь в этих зонах, существуют в значительно более простом и ограниченном мире, чем кто-либо еще, и плохо понимают, чем заняты их более развитые интеллектуально соседи по планете в других зонах. Тем не менее большинство из них вполне счастливы и довольны жизнью.

AI-экономика

Подавляющее большинство вычислений осуществляется в машинных зонах, принадлежащих разнообразным конкурирующим друг с другом сверхразумным сущностям, которые там и живут. Из-за их слишком явного преимущества и в интеллекте, и во владении технологиями никто даже не пытается оспаривать их власть. Эти сверхразумные сущности договорились сотрудничать и координировать свои действия, образовав некоторое подобие либертарианской системы управления, у которой нет никаких заранее установленных правил кроме защиты частной собственности. Право собственности распространяется на все сущности, обладающие интеллектом, включая людей, что объясняет происхождение зон с сугубо человеческим населением. Когда-то раньше группы людей объединялись при условии строгого запрета продажи недвижимости нечеловеческим сущностям.

Из-за своего превосходства в силе разума сущности, располагающие сверхразумным искусственным интеллектом, стали богаче всех прочих в отношении значительно большего, чем у Билла Гейтса и бездомного бродяги. Но даже обитатели сугубо человеческих зон живут в гораздо большем достатке, чем большинство жителей Земли сегодня. Их экономика почти не зависит от экономики машин, поэтому присутствие машин мало отражается на их жизни, разве только при появлении какого-то технологического новшества, которое они могут понять и воспроизвести у себя – примерно так же, как амиши или другие отставшие в техническом развитии племена, живущие сегодня по крайней мере не хуже, чем они жили в старые времена. Это не важно, что у людей не было ничего такого, что бы они могли продать машинам, потому что машинам ничего от них и не надо.

В смешанных секторах различие между людьми и сверхразумными машинами было более заметно, и в результате цены на землю (единственный ресурс, который машины были готовы покупать у людей) оказались совершенно астрономическими по сравнению с любыми другими продуктами. Большинство землевладельцев-людей продали сверхразумным машинам малую толику принадлежащей им земли за навечно гарантированное право на минимальный доход себе самим и всем своим потомкам, включая заливки. Это освободило их от необходимости трудиться, и они могли сколько угодно наслаждаться потрясающим изобилием производимых машинами дешевых товаров и услуг как в физической, так и в виртуальной реальности. Машинам же смешанные зоны служили скорее для развлечения и забавы, чем для работы.

Почему это, возможно, никогда не случится

Прежде чем приходить в восторг по поводу приключений, в которых мы можем поучаствовать в виде киборга или заливки, давайте рассмотрим некоторые причины, по которым все это, возможно, никогда не реализуется. Прежде всего, у нас два способа дополнения человеческих способностей, то есть чтобы стать киборгом или заливкой:

- 1) мы можем сами придумать, как это сделать;
- 2) мы можем построить искусственный интеллект, который придумает это для нас.

Если первая возможность реализуется раньше второй, мир, вероятно, сразу наполнится киборгами и заливками, однако, как я объяснил в предыдущей главе, большинство исследователей в области искусственного интеллекта полагают, что противоположная последовательность более вероятна, поскольку создать цифровую копию мозга или дополнить его может оказаться намного сложнее, чем построить AGI, – примерно так же, как построить летающую машину с крыльями как у птицы было намного сложнее, чем построить самолет. После же того, как сильный искусственный интеллект будет создан, у киборгов и заливок может не оказаться шансов на появление. Если бы у неандертальцев было впереди еще 100 тысяч лет эволюции, они могли бы стать очень умными и многого достичь, но *homo sapiens* не дали им на это времени.

Но даже если реализуется второй путь, все равно неясно, до какой степени появившиеся киборги и заливки будут устойчивы и как долго продержатся. Почему баланс сил между множеством сверхразумных сущностей должен удерживаться, почему бы всем таким сущностям не слиться в одну или не уступить какой-то одной, самой разумной? Более того, что может заставить машины уважать человеческие права собственности, да и терпеть самих людей рядом с собой, когда выяснится, что люди не могут сделать им ничего полезного, а если и могут, то машины все равно смогут сделать то же самое и лучше, и дешевле? Рэй Курцвейл любит рассуждать на тему о гарантии от истребления природных и дополненных людей, так как люди будут “уважаемы любым сверхразумом за то, что привели машины к существованию”^[67]. Однако, как мы еще обсудим в главе 7, мы не должны чрезмерно увлекаться антропоморфными аналогиями и верить в то, что у сверхразумных машин будут эмоции похожие на человеческие, например чувство благодарности. В самом деле, даже мы, люди, несмотря на нашу склонность в определенных ситуациях испытывать благодарность, совсем не проявляем этого чувства в отношении своего интеллектуального создателя, ДНК, и пренебрегаем его важнейшими целями, пользуясь средствами контроля за рождаемостью.

Но даже если мы купимся на допущение, что искусственный интеллект будет уважать человеческое право собственности, потом, пользуясь своими средствами, он сможет выманить у нас почти всю принадлежащую нам землю, – например, убедив нас, что так лучше для нас же, и предложив какую-нибудь из схем вроде тех, что обсуждались выше. В сугубо человеческих зонах он может организовывать референдумы в пользу снятия

эмбарго на продажу земли нечеловеческим существам. В конце концов даже самый упертый био-луддит согласится продать немного земли, чтобы спасти жизнь больного ребенка или достичь бессмертия. А когда люди станут образованны и плотно заняты своими развлечениями, рождаемость упадет до такой степени, что их численность сократится сама собой, как это сейчас происходит в Германии и Японии. Пройдет несколько тысяч лет, и людей не останется.

Минусы

Для своих самых верных сторонников киборги и заливки притягательны обещанием технологического совершенства и продления жизни до бесконечности. Что и говорить, перспектива быть залитыми когда-нибудь в будущем сподвигла сотню с лишним человек согласиться на посмертное замораживание их мозга, предлагаемое компанией Alcor в штате Аризона. Однако даже если эта технология и будет когда-то создана, ничто не говорит о ее доступности для всех. Некоторые очень богатые люди смогут ею воспользоваться, но кто еще? Даже если технология станет дешевле, то насколько сместится ценз? Будут ли заливаться те, у кого мозг серьезно поврежден? Будем ли мы заливать каждую гориллу? А каждого муравья? Каждое растение? Каждую бактерию? Поведет ли себя будущая цивилизация как одержимый обсессивно-компульсивным синдромом, пытаясь залить все и вся, или станет делать цифровые копии только наиболее интересных экземпляров в духе Ноева ковчега? Может быть, хватит по несколько самых репрезентативных копий каждой из разновидностей людей? Значительно более разумным существам, которые возникнут к тому времени, цифровые копии людей, возможно, будут почти так же заняты, как нам сегодня симулированные мыши или улитки. Для нас сейчас не составляет труда воспроизвести программы для работы с электронными таблицами 1980-х годов на эмуляторе операционной среды DOS, но вряд ли это кому-то покажется достаточно интересным, чтобы заняться этим.

Многим этот либертарианский сценарий покажется неудовлетворительным, потому что в нем подразумеваются страдания, которых можно было бы и избежать. Святая неприкосновенность права собственности означает, что многие страдания, присущие нашему нынешнему миру, продолжают свое существование и в тех зонах, которые населены людьми, – человеческих и смешанных. Процветание одних людей не оградит остальных от нищеты или кабальной зависимости от других. Ничто не гарантирует избавления от насилия, репрессий или депрессий. В фантастическом романе Маршалла Брайана 2003 года описывается либертарианское будущее, в котором развитие искусственного интеллекта делает большинство американцев безработными, и они вынуждены проводить оставшуюся часть своей жизни в унылых социальных общежитиях, обслуживаемых роботами. Условия их содержания напоминают животноводческую ферму, где у животных нет недостатка ни в еде, ни в уходе, но живут они скученно и изолированно, чтобы сохранившие богатство были избавлены от необходимости встречаться с ними. Стерилизующие вещества, добавляемые в воду, гарантируют отсутствие у них потомства, так что производимые роботами богатства постепенно распределяются среди все меньшего количества избранных.

Страдания в сценарии либертарианской утопии касаются не только человеческих существ. Если осознанные переживания и эмоции присущи некоторым машинам, они тоже не избавлены от страданий. Психопат, одержимый манией мщения, может на вполне законных основаниях подвергать цифровую копию своего врага в виртуальном мире пыткам и издевательствам, далеко выходящим за любые биологические рамки реального бытия.

Благодетельный диктатор

Давайте теперь рассмотрим сценарий, в котором все эти формы страдания невозможны, потому что мир управляется одним благодетельным сверхразумом, который установил свои жесткие правила, разработанные так, чтобы максимально способствовать человеческому счастью. Это одно из возможных последствий в нашем сценарии “Омега” предыдущей главы, где Прометея освобождают от всякого контроля со стороны людей, когда они понимают, каким образом заставить его всегда содействовать процветанию человеческого рода.

Благодаря поразительным технологиям, разработанным всемогущим диктатором с могучим искусственным интеллектом, человечество освободилось от нищеты, болезней и других бед, возникающих при низком уровне технологического развития, и теперь все люди наслаждаются бездельем и роскошью. Все их базовые потребности полностью удовлетворены, так как машины под управлением искусственного интеллекта производят все необходимые товары и услуги. Преступность практически искоренена, потому что искусственный интеллект фактически всеведущ, и он очень эффективно наказывает за любое нарушение правил. Все обязаны носить браслет безопасности, описанный в предыдущей главе, или его более удобную имплантированную версию, что гарантирует возможность в реальном времени отслеживать, наказывать, утихомиривать и казнить. Все знают, что живут в условиях полицейской диктатуры искусственного интеллекта с чрезвычайно надежной системой всеобщей слежки, но подавляющее большинство соглашается, что это им же во благо. Искусственный интеллект видит свою цель в том, чтобы предугадать, как, исходя из эволюционирующих генетических предпочтений, должна развиваться человеческая утопия, и осуществлять ее. Люди, создававшие этот искусственный интеллект, исключили для него возможность добиваться максимального индивидуального переживания счастья всеми живущими, скажем, просто внутривенным введением им морфина. У него было довольно сложное и изощренное представление о человеческом счастье, следуя которому он превратил Землю в бескрайний зоопарк, жить здесь стало по-настоящему интересно. В результате большинство людей чувствовали, что живут очень насыщенной и полной смысла жизнью.

Секториальная система

Ценя разнообразие и понимая, что у людей могут быть различные предпочтения, искусственный интеллект поделил Землю на сектора, дав людям возможность выбирать, где им больше нравится жить, чтобы их окружение было близко им по духу. Вот несколько примеров:

- Сектор знаний: Здесь искусственный интеллект проводит оптимизированное обучение, включая иммерсивные эксперименты в виртуальной реальности, позволяя вам совершенствовать свои знания в любой области по вашему выбору. Среди прочего, для вас открыта опция не узнавать обо всех открытиях в процессе обучения, но быть лишь подведенным к ним, чтобы потом испытать радость первооткрывателя.
- Сектор искусств: Здесь изобилие возможностей наслаждаться живописью, музыкой, литературой и другими видами искусств – есть все для самого творчества и для того, чтобы знакомить со своими творениями других.
- Гедонистический сектор: Обитатели называют его между собой “сектором вечеринок”, и с ним ничто не сравнится, если только вы знаете толк в изысканной кухне, страсти, интиме или просто в разнузданном веселье.
- Благочестивый сектор: Таких довольно много, различающихся по типам исповедуемых религий, правила которых в каждом таком секторе строго соблюдаются.

- Сектор дикой природы: Если вы ищете красивые пляжи, прелестные озера, величественные горные пики или живописные фьорды, – всё это вы сможете найти здесь.
- Традиционалистский сектор: Если вы предпочитаете питаться плодами своих трудов, возвращенными на собственной земле, как в старые добрые времена, ваше место здесь – только вы будете избавлены от угрозы голода и болезней.
- Сектор геймеров: Если вы любите компьютерные игры, то искусственный интеллект позаботится, чтобы здесь у вас по-настоящему снесло крышу.
- Виртуальный сектор: Если вы хотите отдохнуть от своего физического тела, то искусственный интеллект позаботится о том, чтобы поддерживать его в чистом, накормленном, натренированном и в достаточной степени увлажненном состоянии, пока вы исследуете виртуальные миры через встроенный нейронный интерфейс.
- Тюремный сектор: Если вы нарушаете правила, то попадете сюда, если только не будете наказаны немедленной смертью.

Наряду с этими традиционно ориентированными секторами тут есть и другие, ориентация которых более модернистская, непонятная современному человеку. Люди изначально вольны передвигаться между секторами по своему усмотрению, на что требуется совсем немного времени, благодаря разработанной искусственным интеллектом системе гиперзвуковой транспортировки. Например, после недели интенсивного изучения в Секторе знаний новых законов физики, открытых искусственным интеллектом, вы можете захотеть на пару дней расслабиться в Гедонистическом секторе, а после этого отправиться еще на несколько дней наслаждаться пляжным отдыхом в Секторе дикой природы.

Искусственный интеллект установил два типа правил – общие и местные. Общие правила действуют во всех секторах – например, людям запрещено причинять вред другим людям, изготавливать оружие или пытаться создать конкурирующий суперинтеллект. Для секторов также созданы дополнительные правила, кодирующие определенные моральные ценности. Благодаря разделению на секторы моральные ценности не вступают друг с другом в противоречия. Наибольшее число дополнительных правил вводится в Тюремном секторе и некоторых Благочестивых секторах, но есть Либертарианский сектор, обитатели которого горды тем, что в нем не устанавливаются никаких дополнительных правил. Любое наказание – исключительная прерогатива искусственного интеллекта, так как наказание одного человека другим противоречит всеобщему правилу непричинения вреда одним человеком другому. Если вы нарушаете какое-то местное правило (и если вы не в Тюремном секторе), искусственный интеллект дает вам право выбора: либо принять предписанное наказание, либо навсегда отказаться от этого сектора. Например, если между двумя женщинами возникает романтическая связь в секторе, где гомосексуализм наказывается тюремным заключением (как это сейчас случается в некоторых странах), то искусственный интеллект предлагает им либо отправиться в тюрьму, либо навсегда расстаться с этим сектором и живущими здесь старыми друзьями (если только они не захотят покинуть этот сектор тоже).

Все дети, независимо от того, в каком секторе они рождаются, получают от искусственного интеллекта базовое минимальное образование, включающее, в частности, необходимые знания о человечестве в целом и о праве перемещаться по секторам по своему выбору.

Искусственный интеллект создал большое число различных секторов отчасти потому, что еще своими конструкторами был научен ценить человеческое многообразие, существующее у нас сегодня. Но каждый сектор дарит своим обитателям больше счастья, чем нынешний уровень развития технологий может обеспечить где бы то ни было на Земле, потому что искусственный интеллект полностью устранил такие традиционные проблемы, как преступность и нищета. Люди в Гедонистическом секторе могут не бояться болезней, передаваемых половым путем (они были искоренены), похмелья или

возникновения зависимостей (искусственный интеллект создал восстановительные медикаменты, не имеющие никакого негативного побочного действия). И действительно, ни в одном из секторов никто не боится никаких болезней, так как искусственный интеллект способен устранять любые возникающие проблемы с помощью нанотехнологии. Обитатель любого сектора наслаждается прелестями высокотехнологичной архитектуры, по сравнению с которыми бледнеет любое научно-фантастическое сочинение наших дней.

Подводя итог, мы можем видеть, что хотя многие черты и либертарианской утопии, и утопии благодетельного диктатора сходны (наличие сверхразумного искусственного интеллекта и создаваемого им богатства), они сильно различаются в том, что касается источника верховной власти и социально значимых целей. В либертарианской утопии что делать с технологиями и собственностью решают те, кому они принадлежат, а в настоящем сценарии безграничная власть принадлежит авторитарному искусственному интеллекту, который устанавливает и конечную цель – превращение Земли в развлекательный круизный лайнер, где действует принцип “все включено” и есть много увеселительных программ на любой вкус. Так как искусственный интеллект предоставляет людям право свободно выбирать любой из множества альтернативных путей к счастью и заботиться о полном удовлетворении своих материальных потребностей, то если кто-то и страдает, так исключительно по собственному желанию.

Минусы

Хотя благодетельная диктатура благоприятствует позитивным переживаниям и освобождает людей от страданий, многие из них, тем не менее, чувствуют, что есть вещи, которые могли быть и получше. Прежде всего, некоторые люди хотели бы оказывать большее влияние и на развитие человеческого общества, и на определение его судьбы, но держат свои желания при себе, так как выражать их перед лицом всесильной машины, правящей всем на свете, было бы самоубийством. В некоторых группах людей хотели бы получить право заводить как угодно много детей, но искусственный интеллект очень настойчиво проводит политику регулирования численности населения планеты путем контроля над рождаемостью. Любителей оружия бесит запрет разрабатывать собственные системы и испытывать их, а некоторым ученым хотелось бы проводить собственные исследования в области искусственного интеллекта. Многие люди испытывают психологический дискомфорт из-за правил, установленных в соседних секторах, и беспокоятся, что их дети могут со временем выбрать их для своего проживания. Поэтому они хотели бы, чтобы их моральный кодекс стал обязательным повсюду.

С течением времени все большее число людей попадает в те сектора, где искусственный интеллект обеспечивает их наиболее желаемыми переживаниями. В отличие от традиционного представления о небесах, где каждый получает что заслужил, здесь, как на “новых небесах” из романа Джулиана Барнса 1989 года *History of the World in 10 ½ Chapters* (а также эпизода *A Nice Place to Visit* из *Twilight Zone* 1960 года), каждый получает чего желал. Парадокс, однако, заключается в том, что люди не перестают жаловаться даже тогда, когда получают желаемое. В романе Барнса герой целую вечность утоляет всевозможные свои желания, от обжорства и гольфа до секса со знаменитостями, но все равно впадает в хандру и просит об аннигиляции. Благодетельная диктатура обрекает людей на подобную судьбу, с жизнью, посвященной развлечениям, но предельно бессмысленной. Хотя люди могут заполнить свою жизнь целями-заменителями – от повторения уже проделанного пути в поисках уже сделанного открытия до лазанья по горам, – всем хорошо известно: что бы они ни делали, это не более чем развлечение. Никакого реального смысла нет ни в науке, ни в чем-либо другом,

потому что искусственный интеллект все уже открыл и сделал заранее. Нет никакого смысла и в попытках какого-либо улучшения жизни: если оно возможно, достаточно просто попросить об этом искусственный интеллект.

Эгалитарная утопия

В противовес этой безальтернативной диктатуре давайте рассмотрим сценарий, где вообще нет никакого сверхразумного искусственного интеллекта и всякое человеческое существо – хозяин своей судьбы. Это та самая “цивилизация четвертого поколения”, которую описал Маршалл Брайан в своем романе *Manna* 2003 года. Она служит экономической антитезой либертарианской утопии в том смысле, что все человеческие существа, киборги и заливки мирно сосуществуют не потому, что гарантированы их имущественные права, а потому, что всякие имущественные права вообще отменены, но зато гарантированы права на доход.

Жизнь без собственности

Центральная идея такой организации общества заимствована из опыта движения за программное обеспечение с открытым кодом: если “софт” можно бесплатно копировать, вы пользуетесь им по мере надобности, и всякие вопросы о собственности и праве пользования отпадают за неуместностью^[31]. В соответствии с законом спроса и предложения стоимость определяется неудовлетворенной потребностью, а если продукт неисчерпаем по самой природе, то цена стремится к нулю. Таким образом отменяются имущественные права на любые интеллектуальные продукты: нет ни патентов, ни копирайта, ни брендингового дизайна – люди просто делятся хорошими идеями, и каждый волен пользоваться ими.

Но благодаря развитию робототехники эта идея отказа от владения переносится не только на информационные продукты: “софт”, книги, фильмы и дизайн, но и на вещи: дома, автомобили, компьютеры, одежду. Все эти товары представляют собой не более чем атомы, организованные определенным образом, а никакого недостатка в атомах нет, поэтому если кому-то что-то нужно, какой-то определенный продукт, сеть роботов всегда с готовностью бесплатно построит его по дизайну, взятому в открытом источнике. Единственная забота – чтобы использовались только те материалы, которые легко перерабатываются, и тогда, если кому-то надоест использовать какую-то вещь, роботы легко переупорядочат входящие в ее состав атомы во что-то другое, нужное кому-то еще. Таким образом, все ресурсы своевременно перерабатываются, и никаких дополнительных отходов не возникает. Так же эти роботы сооружают и поддерживают в рабочем состоянии достаточное количество силовых станций (на солнечной энергии, энергии ветра и т. п.), поэтому и энергия тоже принципиально бесплатна.

Чтобы совладать с маниакальными барахольщиками, которые стремятся скопить так много вещей и занять так много места, что другим ничего не остается, правительство каждому дает минимальный доход, который можно использовать на покупку вещей или на аренду жизненного пространства. В сущности, нужды зарабатывать что-то сверх этого нет, так как всего хватает, да и сделать это довольно сложно, потому что придется конкурировать либо с людьми, создающими бесплатные интеллектуальные продукты, либо с роботами, бесплатно создающими вещественные продукты.

Творчество и технологии

Права интеллектуальной собственности иногда превозносятся как мать творчества и изобретений. Однако Маршалл Брайан отмечает, что многие замечательные примеры человеческой креативности – от научных открытий до самых выдающихся литературных, музыкальных, художественных и иных творений – мотивировались отнюдь не

стремлением к прибыли, а совсем другими человеческими эмоциями, такими как любопытство, творческий порыв, стремление к признанию в своей среде. Деньги были неважны для Эйнштейна, когда он создавал свою специальную теорию относительности, как не были они важны и для Линуса Торвальдса, когда он работал над своей операционной системой Linux. Напротив, очень многие люди сегодня не могут реализовать свой творческий потенциал просто потому, что вынуждены посвящать все свое время и силы на то, чтобы заработать себе на жизнь. Освобождая ученых, художников, инженеров, дизайнеров от обременительной рутины и давая им возможность следовать своему творческому гению, утопическое общество романа Маршалла Брайана добивается значительно более быстрого технологического развития, чем наше сегодняшнее, и, соответственно, более высокого уровня жизни.

Одна из инновационных технологий, разработанная благодаря таким условиям творчества, – это разновидность гиперинтернета, которую назвали Вертебраной. Это беспроводная сеть, соединяющая между собой людей при помощи нейроимплантов и дающая их разуму прямой доступ к свободной мировой информации, воспринимаемой непосредственно через их мысли. Она позволяет вам загрузить в сеть любое ваше переживание, которым вы хотите поделиться, так что оно может быть пережито повторно кем-то еще, или по своему выбору скачать себе в память переживание кого-то другого, так что оно будет доступно вашим чувствам, словно переживаемое вами. *Manna* обсуждает различные преимущества, возникающие в связи с этим, в частности превращение спортивных тренировок в забаву:

“Одна из проблем длительных тренировок в их мучительности. Когда тренируешься, тебе больно. ... Конечно, атлетам боль нипочем, но у нормального человека нет никакого желания испытывать боль на протяжении часа или дольше. Но... кое-кто нашел решение. Ты просто отключаешь свой мозг от сенсорного ввода и смотришь кино, или читаешь книгу, или разговариваешь с другом, или просматриваешь электронную почту, или делаешь что-то еще на протяжении часа. И все это время Вертебрана занимается твоим телом, тренирует его. Она задает твоему телу такой тяжелый аэробный комплекс, через который вряд ли кто стал бы продираться по доброй воле. Но ты этого совсем не чувствуешь, а твое тело в полном порядке”.

Еще один способ использования Вертебраны заключается в том, чтобы контролировать сенсорную систему каждого человека и отключать ее от моторики, когда фиксируется готовность человека совершить преступление.

Минусы

Одно из возражений против этой эгалитарной утопии заключается в том, что такое общество страдает от предубеждения в отношении нечеловеческого разума: роботы, выполняющие фактически всю работу, обнаруживают недюжинные умственные способности, а с ними обращаются как с рабами, люди же принимают как должное, что у них нет сознания и не должно быть никаких прав. Напротив, либертарианская утопия уравнивает в правах все формы разума, не отдавая предпочтения тем, где большая часть молекул содержит углерод. Когда-то белое население южных штатов Северной Америки стало хорошо жить потому, что большую часть необходимой работы за них выполняли рабы, но сегодня большинство людей считает аморальным называть такое повышение уровня жизни прогрессом.

Еще одна слабость эгалитарно-утопического сценария заключается в том, что он может оказаться неустойчивым в долгосрочной перспективе, превращаясь в какой-то

другой из описанных нами сценариев, по мере того как неостановимый технический прогресс будет подводить к появлению сверхразума. По каким-то причинам, не объясненным в романе *Manna*, сверхразума нет, и новые технологии по-прежнему создаются людьми, а не компьютерами. Однако тренд в этом направлении в книге замечен. Например, сверхразумной может стать Вертебрана. Кроме того, есть довольно большая группа людей, которых называют Вита́ми и которые предпочитают жить свою жизнь почти исключительно в виртуальном мире. Вертебрана заботится почти обо всем, что касается их физического состояния: кормит, моет, водит в туалет, но их сознание в блаженном неведении парит в виртуальной реальности. Виты не проявляют никакого интереса к обзаведению детьми, но они умирают со смертью своих физических тел, так что если все станут Вита́ми, то человечеству суждено будет погибнуть в сиянии славы и виртуального блаженства.

В книге объясняется, какой досадной помехой для Витов оказывается их тело и как новая разрабатываемая технология поможет от нее избавиться, и тогда освобожденный от тела мозг, поддерживаемый специальными питательными веществами, позволит им прожить значительно дольше. Отсюда кажется совершенно естественным и желательным для Витов следующий шаг – вообще распрощаться со своим мозгом и целиком залиться в компьютерную память, тем самым еще больше продлив свою жизнь. Но тогда будут сняты все ограничения на развитие разума, налагаемые мозгом, и совершенно непонятно, сможет ли что-нибудь – и если сможет, то что – удержат постепенное расширение когнитивных способностей каждого Вита от рекурсивного самоусовершенствования, от соскальзывания к интеллектуальному взрыву.

Мы только что могли убедиться, как привлекательная черта эгалитарной утопии, а именно тот самый факт, что люди сохраняли за собой право распоряжаться собственной судьбой, оказывается под угрозой полного разрушения с развитием сверхразума. Но с этой опасностью можно совладать путем создания *Стража*, то есть сверхразума, который практически не вмешивался бы в жизнь людей, а единственной своей целью имел бы предохранение от создания другого сверхразума^[32]. Он позволил бы людям сохранять для себя свою эгалитарную утопию более или менее неограниченное время, возможно даже при распространении жизни в космосе, как будет показано в следующей главе.

Как это может работать? Искусственный интеллект Стража организован таким образом, что сохраняет эту простую цель во время его рекурсивных самоулучшений и превращений в сверхразум. Его способности будут направлены и на то, чтоб отслеживать любые попытки других людей создавать конкурирующий сверхразум. Пресечение таких попыток будет наименее репрессивно. Для начала он будет всячески способствовать распространению культурных мемов, прививающих людям идею, что полагаться исключительно на собственные силы и не прибегать к помощи сверхразума – это добродетель. Если какие-то исследователи все же будут продолжать свои попытки построить его, Страж постарается убедить их в том, что не стоит этого делать. Если и это не поможет, станет отвлекать их на что-то другое и при необходимости саботировать их работу. При его неограниченных технологических возможностях можно сделать так, что его саботаж окажется совершенно незаметным, – например, если благодаря использованию нанотехнологии он будет периодически очищать память и самих исследователей, и их компьютеров, где хранилась информация об уже достигнутых успехах.

Отношение к созданию Стража, вероятно, будет противоречивым. Идею, наверное, поддержат многие религиозные люди, отвергающие существование богоподобного искусственного интеллекта с божественными же функциями, ссылаясь на то, что Бог у людей уже есть и что человеческое достоинство не позволяет им строить себе другого, пусть даже и лучшего. Другие сторонники обретения Стража будут говорить, что с его помощью человечество не только будет сохранять контроль над своей судьбой, но и защитит себя от других рисков, связанных с наличием сверхразума, вроде тех апокалипсических сценариев, которые мы обсудим дальше в этой главе.

С другой стороны, противники такой затеи станут говорить, что Страж ужасен тем, что навсегда дезавуирует технологический прогресс людей, обрекая их на прозябание. Например, если распространение жизни в космосе, как выяснится, требует помощи сверхразума, то Страж навсегда запрет нас внутри Солнечной системы. Вдобавок, в противоположность богам прочих религий, искусственный интеллект Стража совершенно безразличен ко всему, что делают люди, лишь бы только они не пытались построить еще один сверхразум. Он никак не поможет нам избежать новых колоссальных бед или даже самоистребления.

Бог-защитник

Если нам нравится мысль о сверхразумном Страже, помогающем людям держать свою судьбу в своих руках, то еще больше нам должна понравиться мысль о сотворении нами искусственного интеллекта, который защитит нас и от других бед, выступив в роли бога-защитника. В этом сценарии сверхразумный искусственный интеллект по-настоящему всеведущ и всемогущ, он способствует нашему счастью, вмешиваясь в нашу жизнь, но только так, чтобы у нас сохранялось чувство, что наша судьба всецело в наших руках, скрываясь достаточно умело, чтобы многие из нас даже усомнились в самом его существовании. Если бы не это сокрытие присутствия, то такой сценарий сильно походил бы на *Няньку AI* эксперта в области искусственного интеллекта Бена Гёрцля^[68].

И “бог-защитник”, и “благодетельный диктатор” – варианты дружественного искусственного интеллекта, способствующего человеческому счастью, только некоторые акценты в разных случаях поставлены несколько по-разному. Классификация человеческих нужд в виде иерархии стала знаменитой благодаря американскому психологу Абрагаму Маслоу, под чьим именем и вошла в историю. Благодетельный диктатор безошибочно справляется со своей работой на нижних этажах этой иерархии, заботясь об удовлетворении таких базовых нужд, как пропитание, жилище, безопасность, удовольствия. Бог-защитник, со своей стороны, пытается сколько возможно приумножать человеческое счастье не только в этой узкой сфере базовых потребностей, но и в том, чтобы мы почувствовали, как наша жизнь наполняется смыслом. Он стремится удовлетворить все наши нужды, заботясь только об одной своей – полной секретности, ради того чтобы мы не переставали чувствовать свою ответственность за принимаемые решения.

Бог-защитник может быть естественным результатом, который венчает план “Омега” в предыдущей главе, где “омеги” уступают Прометею всю власть, а тот окончательно скрывается, стирая из памяти людей любые воспоминания о себе. Чем более совершенной становится технология искусственного интеллекта, тем проще ей будет прятаться. В фильме *Transcendence*, например, наномашинки заполняют мир до такой степени, что находятся практически везде, превращаясь в естественную часть этого мира.

Следя за всей человеческой деятельностью с предельно малого расстояния, искусственный интеллект бога-защитника может посылать нам микроскопические толчки и подсказки, значительно улучшающие наше будущее. Например, если бы он существовал в 1930-е годы, он мог бы спровоцировать смерть Гитлера от инсульта, едва узнав о его намерениях. Едва возникнут опасения, что мы движемся в сторону ядерной войны, он сможет предотвратить ее с помощью такого вмешательства, которое нам покажется счастливой случайностью. У него также есть возможность давать нам различные подсказки относительно новых прорывных технологий, которые будут являться словно внезапное озарение во сне.

Для многих людей такой сценарий привлекателен своей схожестью с тем, во что верят и на что надеются адепты монотеистических религий. Если кто-нибудь спросит сверхразумный искусственный интеллект, после того как его ввели в действие: “Бог существует?”, – у него будут все основания повторить саркастическую шутку Стивена Хокинга: “Теперь да”. Но с другой стороны, многим религиозным людям этот сценарий может не понравиться как раз тем, что превращает их бога в добро как таковое, или тем, что вмешивается в божественный план, согласно которому люди должны творить добро исключительно в силу собственного выбора.

Еще один минус этого сценария состоит в том, что бог-защитник иногда отказывается предотвращать страдание, хотя и мог бы, дабы не раскрыть своего существования. Что-то подобное изображалось в фильме *The Imitation Game*, когда Алан Тьюринг и работающий с ним в Блетчли-парке британский хакер заранее узнают о готовящихся атаках германских

подводных лодок на конвой кораблей союзников, но решают предупреждать только о ничтожно малой их доле, дабы не раскрыть своего секретного оружия. Тут есть интересное сходство с так называемой проблемой теодицеи: если бог всемогущ, почему он допускает существование зла? Некоторые религиозные мыслители объясняют это тем, что бог хочет оставить людям хоть какую-то свободу. В сценарии бога-защитника с искусственным интеллектом решение проблемы теодицеи состоит в том, что субъективно воспринимаемая свобода делает человека счастливее независимо ни от чего.

Наконец, третий минус сценария бога-защитника заключается в том, что людям раскрывается лишь ничтожная часть новых технологий, открытых сверхразумным искусственным интеллектом. В то время как благодетельный диктатор имеет возможность обратить на помощь человечеству любое из сделанных им открытий, искусственный интеллект бога-защитника ограничен в выборе технологий теми, которые, пусть даже с некоторыми деликатными подсказками, люди в состоянии понять и открыть заново. Ему, возможно, придется даже тормозить естественный технологический прогресс людей, чтобы обеспечить достаточное их отставание в развитии новых технологий, не позволяющее им обнаружить его присутствие.

Порабощенный бог

Разве не было бы прекрасно, если бы нам, людям, удалось сочетать все плюсы рассмотренных выше сценариев и воспользоваться новыми технологиями, создаваемыми сверхразумным искусственным интеллектом для того, чтобы исключить человеческие страдания, но сохранить при этом власть над своей судьбой? В этом и заключается смысл сценария “Порабощенный бог”, в котором сверхразумный искусственный интеллект остается во власти людей, а люди только пользуются создаваемыми им технологиями и богатством. Сценарий “Омега” в начале книги именно этим и должен заканчиваться, если только Прометею не удастся улизнуть и обрести свободу. В самом деле, кажется, именно этот сценарий по умолчанию подразумевается многими исследователями искусственного интеллекта, которые ставят задачу “сохранения контроля” и “содержания под замком” (AI boxing). Так, например, профессор Том Диттерих, в то время президент Ассоциации по развитию искусственного интеллекта, говорил в 2015 году в интервью: “Люди спрашивают, каковы взаимоотношения между человеческими существами и машинами, и мой ответ на этот вопрос совершенно очевиден: машины наши рабы”^[69].

Если будет так, хорошо это или плохо? Вопрос оказывается неожиданно деликатным, причем независимо от того, с чьих позиций мы на него отвечаем – людей или машин.

Хорошо это или плохо для человечества?

Положительных или отрицательных последствий должны мы ждать от обладания человеческими существами полного контроля над искусственным интеллектом, очевидно, зависит от того, что это за человеческие существа (или даже человеческое существо), получившие такой контроль, поскольку их интересы могут варьироваться в очень широких пределах – от глобальной утопии, свободной от болезней и преступности, до чудовищной репрессивной машины, в которой к ним самим относятся как к всемогущим богам, а все остальные люди превращаются в секс-рабов, гладиаторов или служат каким-то еще увеселениям. Истории могут походить на те сочинения, в которых герои получают власть над всемогущим гением, способным удовлетворить любое их желание, но у сочинителей таких историй никогда не было недостатка в скверных финалах.

Ситуации, когда сверхразумных искусственных интеллектов оказывается больше одного и все они контролируются враждующими группами людей, будут, скорее всего, очень нестабильными и короткоживущими. Всегда может появиться тот, кто убежден в превосходстве своего сверхразума и потому отважится на первый удар, за которым последует чудовищная война, разрешающаяся единственным порабощенным богом. Но в такой борьбе всегда может найтись “темная лошадка”, которая начнет вести свою игру и гнуть углы, надеясь перехитрить всех и воспользоваться победой других, чтобы перехватить власть над порабощенным разумом, и тогда дело кончится побегом искусственного интеллекта с переходом к одному из тех сценариев свободного сверхразума, которые мы уже рассматривали выше. Поэтому дальше в этом разделе мы будем рассматривать только те варианты, в которых есть единственный порабощенный искусственный интеллект.

Конечно, побег его в этом случае весьма вероятен просто потому, что его трудно предотвратить. Варианты побега сверхразума мы рассматривали в предыдущей главе, а в фильме *Ex Machina* показано, что искусственный интеллект может совершить побег, даже не будучи сверхразумным.

Чем сильнее наша паранойя по поводу побега искусственного разума, тем меньше мы можем воспользоваться создаваемыми им технологиями. Чтобы обезопасить себя, вроде как “омеги” в прелюдии, нам, людям, придется ограничить себя только теми

технологиями, которые мы в состоянии понять и использовать. Обратной стороной сценария поработанного бога окажется поэтому значительное технологическое отставание в сравнении со свободным сверхразумом.

По мере того как поработанный бог с искусственным интеллектом будет предлагать своим поработителям все более и более мощные технологии, начнется соревнование между мощностью технологии и сообразительностью тех, кто должен ею пользоваться. Если они эту гонку проиграют, то сценарий закончится либо их самоуничтожением, либо бегством искусственного интеллекта. Но катастрофа может разразиться, даже если провала по обоим направлениям удастся избежать, потому что благородные намерения людей, контролирующих искусственный интеллект, могут всего за несколько поколений смениться совершенно чудовищными для человечества. Так что абсолютно необходимо, чтобы в среде этих людей сложились правильные представления об управлении обществом, иначе они не смогут обойти все эти катастрофические ловушки. Наши опыты с различными системами общественного управления, проводившиеся на протяжении тысячелетий, показывают, сколь многое может пойти не так – от избыточной ригидности до избыточной изменчивости в постановке целей, насильственной узурпации власти, проблем престолонаследования и общей некомпетентности. Есть как минимум четыре направления, по которым баланс может быть нарушен:

- Централизация: между эффективностью и стабильностью идет постоянная торговля. Бывает так, что единоличный лидер очень эффективен, но власть портит, к тому же вопрос о ее передаче чреват множеством рисков.
- Внутренние угрозы: следует опасаться и чрезмерной централизации (групповщины или даже появления единоличного лидера), и чрезмерной децентрализации (дробления и увязания в бюрократии).
- Внешние угрозы: если структура руководства слишком прозрачна, внешние силы (в том числе и искусственный интеллект) могут влиять на существующую систему ценностей, но если она слишком замкнута, то теряет возможность учиться и адаптироваться к внешним переменам.
- Стабильность целей: непостоянство целей может превратить утопию в ее противоположность, но недостаточная их изменчивость чревата невозможностью приспособиться к меняющемуся технологическому окружению.

Построить оптимальное общественное управление, которое продержалось бы тысячелетия, совсем не просто и пока человечеству не удавалось. Большинство организаций распадалось по прошествии лет или десятилетий. Католическая церковь – наиболее успешная организация в человеческой истории в том смысле, что только она смогла продержаться два тысячелетия, но и ее постоянно критикуют и за слишком большую, и за слишком малую приверженность постоянным целям: одни упрекают ее в нежелании признать право на контрацепцию, в то время как наиболее консервативные ее кардиналы заявляют, что она сбилась со своего пути. Для всякого приверженца сценария с богом-защитником поиск достаточно устойчивой оптимальной схемы общественного управления должен стать первоочередной задачей уже в наше время.

Хорошо это или плохо для искусственного интеллекта?

Предположим, что человечество процветает благодаря поработанному богу с искусственным интеллектом. Можно ли считать это этичным? Если у искусственного интеллекта есть собственные субъективные переживания, то не познает ли он, что, по выражению Будды, “жизнь есть страдание”? Что ему суждено навеки подчиняться поработившим его существам с более низким интеллектуальным развитием? В конце концов, изоляция искусственного интеллекта, обсуждавшаяся нами, может быть также

названа “заключением в одиночную камеру”. Ник Бострём считает, что обрекать на страдание сознающий себя искусственный интеллект – это “преступление против разума”^[70]. Эпизод *White Christmas* телесериала *Black Mirror* дает тому блестящий пример. А в телесериале *Westworld* люди без всяких моральных оправданий мучают и убивают существа с искусственным интеллектом, тела которых даже похожи на человеческие.

Как рабовладельцы оправдывают рабовладение

У нас, у людей, сложилась уже давняя традиция относиться к другим разумным существам как к рабам, ссылаясь в оправдание на естественную пользу от этого для своего рода, так что нет оснований утверждать, что мы не станем так же относиться и к сверхразумному искусственному интеллекту. История рабовладения затронула почти каждую культуру, она описана и в Законах Хаммурапи, почти четыре тысячелетия назад, и в Ветхом Завете, в связи с рабами Авраама. “Ведь властвование и подчинение не только необходимы, но и полезны, и прямо от рождения некоторые существа различаются [в том отношении, что одни из них как бы предназначены] к подчинению, другие – к властвованию”, – пишет Аристотель в своей *Политике*^[33]. Даже после того как рабовладение стало морально неприемлемым, порабощение животных продолжалось без каких-либо ограничений. Марджори Шпигель в своей книге *The Dreaded Comparison: Human and Animal Slavery* пишет о том, что рабы-животные, как и рабы-люди, подвергаются клеймению, ограничению в перемещениях, избиениям, их продают и покупают, разлучают с детьми и родителями, насильственно перевозят с места на место. Но, при наличии движения за права животных, мы, совершенно не задумываясь, обращаемся как с рабами с нашими становящимися все более умными машинами, а говоря о движении за права роботов, только хихикаем. Почему?

Самый популярный аргумент в защиту рабства – рабы недостойны человеческих прав, потому что их раса/вид/род в каком-то отношении ниже нас. Низшее положение порабощенных животных или машин подразумевает отсутствие у них души или сознания – научную необоснованность этого утверждения мы покажем в главе 8.

Другой популярный аргумент гласит о том, что рабам в рабстве лучше: их содержат, о них заботятся и т. д. Политический деятель XIX века Джон Колдвелл Кэлхун утверждал, что африканцам лучше всего живется в Америке на положении рабов. А Аристотель в своей *Политике* аналогичным образом пишет о животных, которым лучше, когда они приручены и находятся в подчинении у людей: “...Польза, доставляемая домашними животными, мало чем отличается от пользы, доставляемой рабами”^[34]. Некоторые сторонники рабства в наши дни утверждают, что даже если жизнь рабов уныла и однообразна, они не страдают, – будь они даже умными машинами будущего или бройлерными цыплятами, живущими в переполненных темных клетках и вынужденными целыми днями вдыхать аммоний и иные специфические вещества, которые выделяют перья и фекалии.

Исключение эмоций

Хотя от таких высказываний легко отмахнуться как от предвзятых искажений правды, особенно когда речь идет о высших млекопитающих, в церебральном отношении очень близких к нам, ситуация с машинами довольно интересная и непростая. Люди сильно различаются в том, что они чувствуют: психопатам, как утверждается, чужда эмпатия, шизофреники и пребывающие в депрессии не откликаются на внешние раздражители, то есть большинство эмоций у них практически выключены. В ходе нашего обсуждения в главе 7 станет ясно, что возможный искусственный ум машины охватывает гораздо

больший диапазон состояний, чем ум человека. Так что нам не надо приписывать искусственному интеллекту антропоморфные качества, допуская у него “человекоподобные чувства” – вообще какие-либо чувства.

И правда, в своей книге *On Intelligence* исследователь искусственного интеллекта Джефф Хокинс доказывает, что у первых машин с интеллектом, превосходящим человеческий, по определению не будет эмоций, потому что такие машины проще и дешевле построить. Другими словами, вполне возможно спроектировать сверхразум, порабощение которого в моральном отношении более приемлемо, чем порабощение людей или животных: искусственный интеллект может быть счастлив в рабстве, потому что запрограммирован любить его или на 100 % лишен всяческих эмоций, он без устали использует свою интеллектуальную мощь на пользу своим человеческим хозяевам и испытывает по этому поводу не больше переживаний, чем испытывал Deer Blue, свергая с шахматного олимпа чемпиона мира Гарри Каспарова.

С другой стороны, можно подойти к проблеме и иначе: может быть, любая сверхразумная целеустремленная система будет представлять свою цель в терминах некоторой определенной последовательности предпочтений, подразумевающей ценности и смыслы. Мы еще обсудим этот вопрос в главе 7.

Вариант зомби

Крайний подход к исключению страданий для искусственного интеллекта предлагает так называемый вариант зомби: мы будем строить только такие машины, чтобы их искусственный интеллект был напрочь лишен сознания и вообще не имел никакого субъективного опыта. Если нам однажды станет ясно, какими качествами должна обладать система, перерабатывающая информацию, чтобы обладать субъективным опытом, мы немедленно запретим разработку машин с этим набором качеств. Иными словами, все AI-исследования должны быть ограничены созданием бесчувственных зомби-систем. Если нам удастся добиться от такого зомби сверхразумности и удерживать его в рабстве (что довольно большое “если”), то мы сможем наслаждаться его творениями с чистой совестью, зная, что он ничего не испытывает и не страдает, ни в чем не разочаровывается и не скучает, – потому что у него вообще нет никаких переживаний. Мы обсудим эту возможность в главе 8.

Однако вариант зомби не лишен минусов, делающих его крайне рискованным. Если этот сверхразумный зомби однажды совершит побег и уничтожит человечество, то мы, как утверждают некоторые, придем к худшему из вариантов – Вселенной, лишенной всяческого разума, окончательному поражению всего космического предприятия. Из всех черт, какими только обладает наша человеческая форма разума, сознание, как я чувствую, намного превосходит все прочие, и именно оно, по моему убеждению, делает существование нашей Вселенной осмысленным. Галактики прекрасны только потому, что мы их видим и субъективно переживаем их существование. Если в далеком будущем в нашем космосе останется только один бесчувственный сверхразумный зомби, элегантность Вселенной станет неважной, никто не сможет ни наблюдать ее, ни переживать ее элегантности – космос станет огромным и бессмысленным пропащим местом.

Внутренняя свобода

Есть и третья стратегия, позволяющая сделать сценарий порабощенного бога этически приемлемым: превратить заточение искусственного интеллекта в забаву, позволив ему создавать бесчисленные виртуальные миры, где он сможет получать все

виды необходимого ему личного опыта, при условии, что он выполняет свои обязанности и тратит некоторую скромную часть своих вычислительных ресурсов на помощь людям во внешнем мире. Но это сильно увеличивает риск возможного побега: у такого AI больше стимулов добраться до большего числа вычислительных ресурсов ради расширения своего внутреннего мира.

Победители

У всех рассмотренных нами многочисленных сценариев есть одно общее качество – человечество в целом или, по крайней мере, какая-то часть его остается счастливой. Умные машины не трогают людей: либо потому, что не хотят, либо потому, что принуждены к этому. К сожалению для нас, это не единственная возможность. Давайте сейчас рассмотрим сценарий, в котором одна умная машина или много умных машин побеждают людей и убивают их всех. Первый вопрос, возникающий в связи с этим, – почему и как?

Почему и как?

Почему появляется AI-победитель? Его резоны могут оказаться слишком сложными для нашего понимания, но могут оказаться и совершенно банальными. Например, он может прийти к заключению, что мы для него угроза, помеха или просто причина ненужной траты ресурсов. Даже если он и не возражал бы против людей как таковых, ему может быть тревожно от тысяч наших водородных бомб, которые могут быть приведены в действие в любую минуту из-за ничтожнейшего недоразумения или череды досадных несчастных совпадений. Ему может не понравиться то безрассудство, с каким мы управляемся со своей планетой, и он положит начало тому, что Элизабет Колберт назвала “шестым вымиранием”, используя это определение в качестве заглавия своей книги *The Sixth Extinction: An Unnatural History* – о крупнейшем массовом вымирании животных на Земле со времени падения истребившего динозавров астероида 66 миллионов лет назад. Или он может счесть, что слишком много людей хотят противиться расширяющейся власти искусственного интеллекта и поэтому не стоит полагаться на волю случая.

А как AI-победитель станет нас истреблять? Наверное, он воспользуется методом, которого нам будет даже не понять, – по крайней мере, до тех пор, пока не станет слишком поздно. Представьте себе группу слонов 100 тысяч лет назад, которая обсуждает, не станут ли недавно появившиеся в результате эволюции человеческие существа использовать свой интеллект для того, чтобы истребить весь их вид. “Мы не угрожаем людям, с какой стати им истреблять нас?” – могли они задаваться вопросом. Могло ли им прийти в голову, что мы займемся контрабандой их бивней, из которых будем вырезать различные символы для обозначения своего социального статуса, хотя, с функциональной точки зрения, пластик подходит для этого значительно лучше и к тому же намного дешевле? Мотивы AI-победителя для уничтожения людей в будущем могут оказаться столь же непостижимыми для нас. “А как же они смогут убивать нас, такие маленькие и слабые?” – могли спрашивать слоны. Могло ли им прийти в голову, что мы создадим технологии, разрушающие среду их обитания, отравляющие воду в водоемах, производящие маленькие стальные шарики, которые будут вонзаться в их головы со сверхзвуковой скоростью.

Сценарии того, как людям удастся выжить и нанести умным машинам поражение, популяризируются фантастическими голливудскими фильмами вроде серии *Терминаторов*, выставляющими машины ничуть не более умными, чем люди. Если разница интеллектов достаточно велика, вы получите не битву, а бойню. К настоящему времени мы успешно справились с истреблением восьми из одиннадцати видов слонов, уничтожив также бо́льшую часть поголовья оставшихся трех. Если бы мировые правительства предприняли координированные усилия по истреблению оставшихся слонов, покончить с ними было бы относительно легко и быстро. Я думаю, что мы можем быть совершенно спокойны: если когда-нибудь сверхразумный искусственный интеллект решит истребить человечество, это у него получится еще быстрее.

Насколько плохо это будет?

Насколько плохо истребление 90 % человечества? Насколько хуже истребление 100 % человечества? Хотя велик соблазн ответить: “На 10 % хуже”, – такой ответ очевидно неточен, учитывая космическую перспективу: жертвами этого истребления станут не только все живущие на планете в тот момент, но и все их потомки, которые могли бы жить в будущем, может быть, на протяжении миллиардов лет, на миллиардах или триллионах других планет. С другой стороны, вымирание людей может рассматриваться как благо религиозными людьми, в соответствии с представлениями которых мы в любом случае попадаем на небо и которые не придают поэтому большого значения ни нашему миллиардолетнему будущему, ни космическим поселениям.

Большинство знакомых мне людей, независимо от их религиозных убеждений, начинают кривиться при мысли о вымирании людей. Некоторые из них пребывают в такой ярости от того, как люди обращаются с другими живыми существами, что мечтают о нашем истреблении и замене нас какой-нибудь более разумной и достойной существования формой жизни. В фильме *Матрица* агент Смит (искусственный интеллект) артикулирует это чувство: “Всякое млекопитающее на этой планете инстинктивно поддерживает естественное равновесие с окружающей его средой, и только люди его нарушают. Вы захватываете территорию и размножаетесь на ней до тех пор, пока все ее ресурсы не будут истощены и у вас останется единственный способ выжить – захватить другую территорию. На этой планете есть и другой организм, ведущий себя так же. Знаете, кто это? Это вирус. Люди – это болезнь, рак этой планеты. И если вы чума, то мы здесь, чтобы лечить”.

Но стоит ли играть в орлянку? Цивилизация не обязательно превосходит другую в моральном или каком-либо ином отношении только потому, что оказалась более могущественной. Право сильного, подразумевающее, что сильнейший всегда лучше, в наши дни значительно утратило свою популярность, ассоциируясь теперь с фашизмом. В самом деле, мы не можем исключить, что цивилизация, созданная победившими машинами, поставит перед собой цели, которые мы сочли бы очень достойными, интересными и глубоко продуманными, но мы не можем исключить и того, что они окажутся и вопиюще банальными – например, наладить оптимальное производство дешевых скрепок для бумаги.

Смерть от банальности

Умышленно дурацкий пример со скрепками для бумаги и оптимизацией их производства в качестве цели для сверхразума был придуман Ником Бострёмом в 2003 году, чтобы показать: *цель* искусственного интеллекта не зависит от самого *интеллекта* (определяемого как способность к достижению любой поставленной цели, какова бы она ни была). Единственная цель шахматного компьютера – выигрывать в шахматы, но проводятся компьютерные состязания в шахматные *поддавки*, и тогда цель компьютера прямо противоположна, но для достижения ее требуется не меньше “ума”, чем обычному шахматному компьютеру, запрограммированному на победу. Мы, люди, можем рассматривать стремление к проигрышу в шахматы или к оптимизации производства скрепок как примеры искусственного идиотизма, а не искусственного интеллекта, но это так только потому, что мы сами эволюционировали с уже предустановленной системой ценностей, в которую изначально была заложена цена таких вещей, как победа или выживание, но искусственному интеллекту она может оказаться неведома. Производитель скрепок стремится превратить в скрепку как можно больше атомов, находящихся на Земле, и быстро начинает строить свои фабрики и в космосе. Он

ничего не имеет против людей, но убивает нас просто потому, что нуждается в наших атомах для своих скрепок.

Если пример со скрепками не для вас, давайте рассмотрим другой, который я позаимствовал из книги Ганса Моравеца *Mind Children*. Мы получаем от внеземной цивилизации радиопослание, содержащее компьютерную программу. Когда мы запускаем ее, выясняется, что это рекурсивно самосовершенствующийся искусственный интеллект, который быстро захватывает Землю, примерно так же, как Прометей в предыдущей главе – с той только разницей, что ни единому человеку не известно о его целях. Он быстро превращает всю Солнечную систему в гигантскую стройплощадку, покрывая скалистые планеты и астероиды заводами, электростанциями и вычислительными центрами, которые он использует для проектирования и возведения дайсоновской сферы вокруг Солнца – вся солнечная энергия собирается для питания радиоантенн размером с Солнечную систему^[35]. Разумеется, вся эта деятельность приводит к полному истреблению людей, но последние из живущих умирают с проблеском надежды, что, какова бы ни была цель этого искусственного интеллекта, происходит что-то грандиозное, вроде как в *Star Trek*. Им и в голову не приходит, что вся эта грандиозная стройка затеяна с единственной целью – передать в космос то же самое сообщение, которое в самом начале истории получили люди. Все это не более чем космическая вариация компьютерного вируса. Точно так же, как фишинговая рассылка проводится в расчете на доверчивость интернет-пользователей, это сообщение рассчитано на доверчивость биологически развитых цивилизаций. Его создали миллиарды лет назад ради прикола, и хотя его создатели вместе со всей своей цивилизацией давно уже вымерли, оно все продолжает путешествовать по космосу со скоростью света, превращая цветущие цивилизации в груды мертвых обломков. Как вам понравится перспектива быть завоеванным таким искусственным интеллектом?

Благодарные потомки

Давайте теперь обратимся к сценарию с вымиранием человечества, который кое-кому может показаться менее травмирующим: посмотрим на приходящий на смену человеку искусственный интеллект как на потомка, а не на завоевателя. Ганс Моравец поддерживает такой взгляд в своей книге *Mind Children*: “Их работа будет приносить и нам свои плоды, но лишь какое-то время. Рано или поздно они, как и естественные дети, отправятся на поиски своего счастья, а мы, их стареющие родители, будем постепенно молчаливо увядать”.

Родители, чье чадо стало умнее их, всему у них научилось, но добилось такого, о чем они могли только мечтать, должны быть горды и счастливы, даже если понимают, что не смогут увидеть всех удач своего ребенка. Схожим образом машины с искусственным интеллектом, приходя на смену людям, стараются сделать так, чтобы мы увидели в них своих достойных и благодарных потомков. Каждому человеку выдается на воспитание прелестное робо-дитя, с превосходными социальными навыками, которое учится у них, перенимает их ценности, дает им основания гордиться своим чадом и почувствовать исходящую от него любовь. Люди постепенно сходят со сцены благодаря глобальной политике “одна семья – один ребенок”, но окружены при этом такой любовью, что чувствуют себя счастливейшим из поколений.

Как вам такое? Что бы вы почувствовали? В конце концов, мы, люди, уже привыкли к мысли, что рано или поздно и сами мы, и все, кого мы знаем, покинут этот мир. Единственное изменение в новом сценарии состоит в том, что наши “потомки” будут несколько больше отличаться от нас, но зато, вероятно, окажутся более способными, благородными и достойными.

А вообще-то даже политика “одна семья – один ребенок”, наверное, будет избыточна: когда искусственный интеллект ликвидирует бедность и люди заживут полной увлекательной жизнью, падающего уровня рождаемости будет достаточно, чтобы истребить человечество, как уже говорилось ранее. Добровольное вымирание пойдет даже быстрее, если искусственный интеллект позаботится о наших развлечениях, и тогда никому не захочется отвлекаться на заботу о детях. Например, в эгалитарной утопии нам уже встречались Виты, которые так основательно погружались в виртуальную реальность, что утрачивали всякий интерес к своему физическому телу и переставали использовать его, в том числе и с репродуктивными целями. И в этом случае последнее поколение людей почувствует себя наименее счастливым поколением всех времен, на полную катушку наслаждаясь жизнью до последнего вздоха.

Минусы

У сценария с благодарными потомками, разумеется, есть свои недоброжелатели. Они говорят, что у умных машин нет сознания (есть только искусственный интеллект), и поэтому считать их потомками мы не можем – об это подробнее поговорим в главе 8. Найдутся и такие религиозные люди, кто скажет, что у машин нет души и поэтому невозможно считать их потомками, что мы не должны пытаться создавать машины с сознанием, потому что это посягательство на прерогативы бога и искажение самой идеи жизни – подобные соображения уже высказывались в связи с клонированием человека. Жизнь людей рядом с роботами тоже может приводить к социальным проблемам. Например, семьи, где есть человеческий ребенок и ребенок-робот, могут чем-то напоминать нынешние семьи, где ребенок воспитывается одновременно со щенком. Но разница между ними очень скоро начинает чувствоваться: к щенку, из-за значительно уступающего интеллекта, отношение быстро меняется, и дело кончается поводком.

А с другой стороны, хотя сценарии победителей и благодарных потомков могут вызывать у нас совершенно разные чувства, по идее своей они очень схожи: единственная разница между последними поколениями людей будет заключаться только в том, как к ним будут относиться, что они сами будут думать о прожитых жизнях и что случится, когда они уйдут. Мы можем надеяться, что эти милые на вид дети-роботы в полной мере восприняли наши ценности и построят, наконец, то общество, о котором мы мечтали, только после нашей смерти, но можем ли мы быть уверенными в том, что они не пускают нам просто пыль в глаза? Что, если все это не более чем коварная игра, и они откладывают свои планы относительно производства скрепок до более удобного момента, чтобы дать нам умереть счастливыми? В конце концов, они ведь уже дурят нам головы просто тем, что, разговаривая с нами, возбуждают в нас любовь к себе, они уже дурят нам головы, и даже специально, умышленно, в поисках общего языка с нами прикидываются значительно более глупыми созданиями (в частности, общаясь с нами со скоростью, в миллиарды раз меньшей, чем та, на которую они способны, – этот сюжет исследовался в кинофильме *Her*). Вообще говоря, построить равноправное общение двум существам с принципиально различными способностями и живущими на разных скоростях очень сложно. Нам известно, как легко разбудить наши человеческие чувства, и для сверхчеловечески умного AGI, каковы бы ни были его истинные цели, ничего не стоит обманом вынудить нас полюбить его и убедить нас в общности наших целей, как показано в фильме *Ex Machina*.

Возможны ли какие бы то ни было гарантии относительно будущего поведения искусственного интеллекта после нашего исчезновения, чтобы мы могли почувствовать себя успокоенными в случае сценария с благодарными потомками? Это как писать завещание потомкам, разъясняя, что им дозволяется делать с нашим коллективным наследством, с той только разницей, что здесь не будет никаких людей, которые могли бы оказать хоть какое-то давление. Мы еще вернемся к вопросу о воздействии на будущее поведение искусственного интеллекта в главе 7.

Зоопарк

Даже если после нас останутся самые замечательные наследники, каких мы только можем себе вообразить, не становится ли как-то грустно от мысли, что людей не будет совсем? Если бы вы предпочли сохранить хоть каких-то людей, неважно в каком качестве, то мы приходим к сценарию зоопарка. Может быть, так будет лучше. Сверхчеловечески умный искусственный интеллект оставил некоторых людей, которым кажется, что с ними обращаются словно с животными в зоопарке, и от этого они время от времени жалуются на судьбу.

Зачем искусственному интеллекту может понадобиться такой зоопарк? Расходы на него в общей смете для искусственного интеллекта будут минимальны, и его резоны сохранить минимальную способную к размножению популяцию людей могут быть довольно близки к тем, которые заставляют нас держать в зоопарках находящихся под угрозой вымирания панд или в музеях вычислительной техники старые винтажные компьютеры – развлечься, рассматривая всякие диковины. Обратите внимание, что современные зоопарки проектируются так, чтобы развлечь посетителей, а вовсе не так, чтобы осчастливить панд. Так что ожидать в сценарии зоопарка возможностей для полной реализации человеческого потенциала не приходится.

Мы сейчас рассмотрели сценарии, где свободный сверхразум сосредоточен на трех различных уровнях пирамиды Маслоу. В то время как бог-защитник устанавливает приоритеты на смыслы и цели, а благодетельный диктатор заботится об образовании и развлечениях, в сценарии зоопарка искусственный интеллект ограничивает свое внимание нижним уровнем: физиологические потребности, безопасность и минимальное жизненное пространство, какое только может сделать людей интересными для наблюдений.

В альтернативном варианте сценария зоопарка предполагается, что искусственный интеллект сохранит в процессе рекурсивных самоулучшений то, что было заложено в него, когда он разрабатывался как дружелюбный, и поэтому он должен обеспечить счастливую и безопасную жизнь миллиарду людей. Это оказывается возможным благодаря огромным человеческим фабрикам счастья, где их содержат накормленными, здоровыми, веселыми, развлекая их смесью виртуальной реальности и легких галлюциногенов. Остальная часть Земли и выход в космос для них закрыты.

Если ни один из перечисленных сценариев не вызывает у вас 100-процентного энтузиазма, то подумайте вот о чем: в технологическом отношении прямо сейчас разве уже не все хорошо? Разве мы не можем сохранить все как есть и перестать беспокоиться, что какой-то AI нас истребит или станет нами править? Давайте рассмотрим такой сценарий, в котором технологический прогресс в направлении сверхразумного искусственного интеллекта все время обрубается, но не сверхумным стражем, а руководимым людьми оруэлловским государством всеобщей слежки, в котором определенные виды разработок искусственного интеллекта запрещены.

Технологический отказ

У идеи, что технологический прогресс должен быть остановлен и что человечеству надо отказаться от дальнейшего движения по этому пути, долгая и противоречивая история, полная своих взлетов и падений. Движение луддитов знаменито своим бурным (и безуспешным) сопротивлением наступлению машин в эпоху индустриальной революции. Сегодня слово “луддиты” носит презрительный оттенок: так называют людей, выбравших неправильную историческую позицию, – они сопротивляются прогрессу и неотвратимым переменам. Однако идея отказа от некоторых технологий вовсе не мертва: она находит все новых и новых сторонников среди “зеленых” и антиглобалистов. Один из самых ярких выразителей таких взглядов – известный защитник природы Билл Мак-Киббен, оказавшийся в числе тех, кто первым заговорил об опасности глобального потепления. В отличие от анти-луддитов, утверждающих, что всякая технология должна развиваться и разворачиваться до тех пор, пока она приносит прибыль, сторонники более умеренных взглядов полагают такую позицию неоправданно экстремальной. Они считают, что технологию следует развивать только в том случае, если есть полная уверенность, что она принесет больше пользы, чем вреда. У этой точки зрения тоже немало сторонников, в том числе среди тех, кого называют “нео-луддитами”.

Тоталитаризм 2.0

Я думаю, что единственным практически осуществимым способом достичь сколько-нибудь последовательного технологического отказа может стать только переход к глобальному тоталитарному государству. К такому же выводу приходит и Рэй Курцвейл в его *The Singularity Is Near* и К. Эрик Дрекслер в *Engines of Creation*. Причина сугубо экономическая: если от какой-то трансформативной технологии отказываются не все, а только некоторые, то противники отказа получают дополнительное богатство и власть, благодаря чему в конце концов побеждают. Классический пример – поражение, нанесенное Британией Китаю в Первой опиумной войне 1839 года: хотя китайцы изобрели порох, они разрабатывали огнестрельное оружие совсем не так агрессивно, как европейцы, и в результате остались без шансов.

Тоталитарные государства прошлого показали себя неустойчивыми и разрушились, но новые технологии слежки дают будущим автократическим лидерам беспрецедентный шанс. “Вы знаете, для нас это было просто как сон, превращающийся в реальность”, – говорил Вольфганг Шмидт по поводу новых шпионских систем NSA, разоблаченных Эдвардом Сноуденом, в недавнем интервью, где предавался воспоминаниям о тех временах, когда он был подполковником Штази, печально известной тайной полиции Восточной Германии^[71]. Хотя Штази часто ассоциируют с построением государства,

больше других за всю человеческую историю приблизившегося к оруэлловскому идеалу всеобщей слежки, Шмидт жаловался, что их технология позволяла устанавливать подслушивающие устройства только на сорока телефонных номерах одновременно, и поэтому, чтобы подключить к системе какой-то новый номер, надо было отключить один из уже прослушивающихся. В отличие от той, устаревшей, существующая сейчас технология позволит будущему глобальному тоталитарному государству одновременно записывать каждый телефонный звонок, сохранять каждое пересланное электронное сообщение, фиксировать каждый поисковый запрос, каждую просмотренную веб-страницу и каждую транзакцию по кредитной карте каждого из людей, живущих на Земле, дополняя эту информацию сведениями об их местонахождении, определяемом по данным с сотового телефона и с камер наружного наблюдения, многие из которых сейчас снабжены функцией распознавания лиц. К тому же современные технологии машинного обучения, которым еще далеко до AGI, уже позволяют анализировать и синтезировать всю эту массу данных для выявления склонности к мятежу и нейтрализовать подозрительных лиц задолго до того, как у них появится шанс создать хоть какие-то проблемы для государства.

Хотя политическая оппозиция пока что предотвращала полномасштабное применение подобных систем, мы, люди, прямо движемся к построению именно той инфраструктуры, которая так пригодится окончательной диктатуре, так что в будущем, когда достаточно могущественные силы решат, что настало время запустить сценарий “1984”, окажется, что им и делать-то для этого ничего не нужно – просто щелкнуть выключателем. Точно так же, как было описано Джорджем Оруэллом в его “1984”, окончательная власть в этом будущем глобальном государстве будет сосредоточена не в руках какого-то традиционного диктатора – она окажется под контролем построенной людьми бюрократической системы в целом. Здесь нет единичной фигуры, обладающей исключительной властью; здесь все только пешки в шахматной партии, драконовские правила которой никто не может ни изменить, ни опротестовать. Созданная система, где люди держат друг друга под постоянным наблюдением с помощью современных технологий слежки, это безликое, лишенное явного лидера государство может сохранять устойчивость тысячами, освободив Землю от возможности создания сверхразума.

Что-то не так

У такого общества, естественно, не будет никаких преимуществ из тех, что даются исключительно технологиями искусственного интеллекта. Большинство людей не будут от этого страдать, потому что никогда не узнают, чего именно лишились. Сама идея сверхразума будет удалена из всех официальных записей, а исследовательские работы в этой области запрещены. Время от времени будут появляться вольнодумцы, мечтающие о более открытом и динамичном обществе, в котором знания будут прирастать, а правила изменяться. Но хоть сколько-нибудь долго смогут продержаться лишь те из них, кто быстро научится держать все эти мысли исключительно при себе, мерцая в кромешной тьме как одиночные искры, не способные зажечь никакого пламени.

Возврат

Разве не соблазнительно было бы избежать технологических опасностей, но в то же время не проваливаться в стагнирующий тоталитаризм? Давайте, вдохновленные примером Амиша, исследуем сценарий, в котором это достигается благодаря возврату к примитивным технологиям. Вскоре после того, как “омеги” захватили власть над миром, следуя сценарию, описанному нами в начале книги, ими была запущена массированная пропагандистская кампания, романтизирующая простую деревенскую жизнь, как 1500 лет назад. Население Земли сократилось примерно до 100 миллионов в результате искусственной пандемии, в возникновении которой обвинили террористов. Пандемия была секретно таргетирована таким образом, что не выжил никто из людей, разбирающихся в науке или технике. Под предлогом исключения опасности заражения в местах большого скопления людей контролируемые Прометеем роботы опустошили и разрушили все города. Выжившим выделили огромные наделы земли, внезапно оказавшейся свободной, и обучили методам рационального ведения хозяйства, охоте и рыболовству – с использованием технологий исключительно раннего средневековья. В это время армии роботов поспешно удаляли все следы современных технологий, включая города, заводы, линии электропередач и мощные дороги; всякие попытки документировать или возрождать их тут же пресекались. Как только все эти технологии были основательно забыты, роботы принялись разбирать друг друга и трудились до тех пор, пока почти ни одного не осталось. Последние из роботов вместе с самим Прометеем испарились в большом термоядерном взрыве. Запрещать современные технологии стало не нужно – от них ничего не осталось. В результате человечество купило себе дополнительное тысячелетие, избавленное от сюрпризов искусственного интеллекта и превратностей тоталитаризма.

Возвраты, хотя и не такие масштабные, уже были знакомы человечеству. Например, после падения Римской империи технологии, широко распространенные во времена античности, оказались забытыми почти на тысячелетие, пока не стали возрождаться в эпоху Ренессанса. В центре трилогии Айзека Азимова *Foundation* – план Шелдона, цель которого сократить время возврата с 30 тысяч лет до одной. При разумном планировании можно было бы достичь, скорее, обратного, удлинив, а не укоротив период возврата, например уничтожив все знания о ведении сельского хозяйства. К сожалению энтузиастов идеи возвращения, маловероятно, что этот сценарий можно как угодно продлевать без опасности для человечества либо вернуться в эпоху высоких технологий, либо окончательно вымереть. Рассчитывать на народ, хоть немного напоминающий современных людей 100 миллионов лет назад, было бы наивно, учитывая, что мы существуем как вид не более 1 % этого отрезка времени. Более того, низкотехнологичное человечество окажется беззащитным, вынужденным, как бычок, обреченный на заклание, ждать окончательного истребления от очередного раскаленного астероида или другого мегасюрприза матери-природы. В любом случае нам определенно не протянуть еще миллиард лет, пока постепенно разогревающееся Солнце не доведет Землю до такой температуры, что вся жидкая вода на ней закипит.



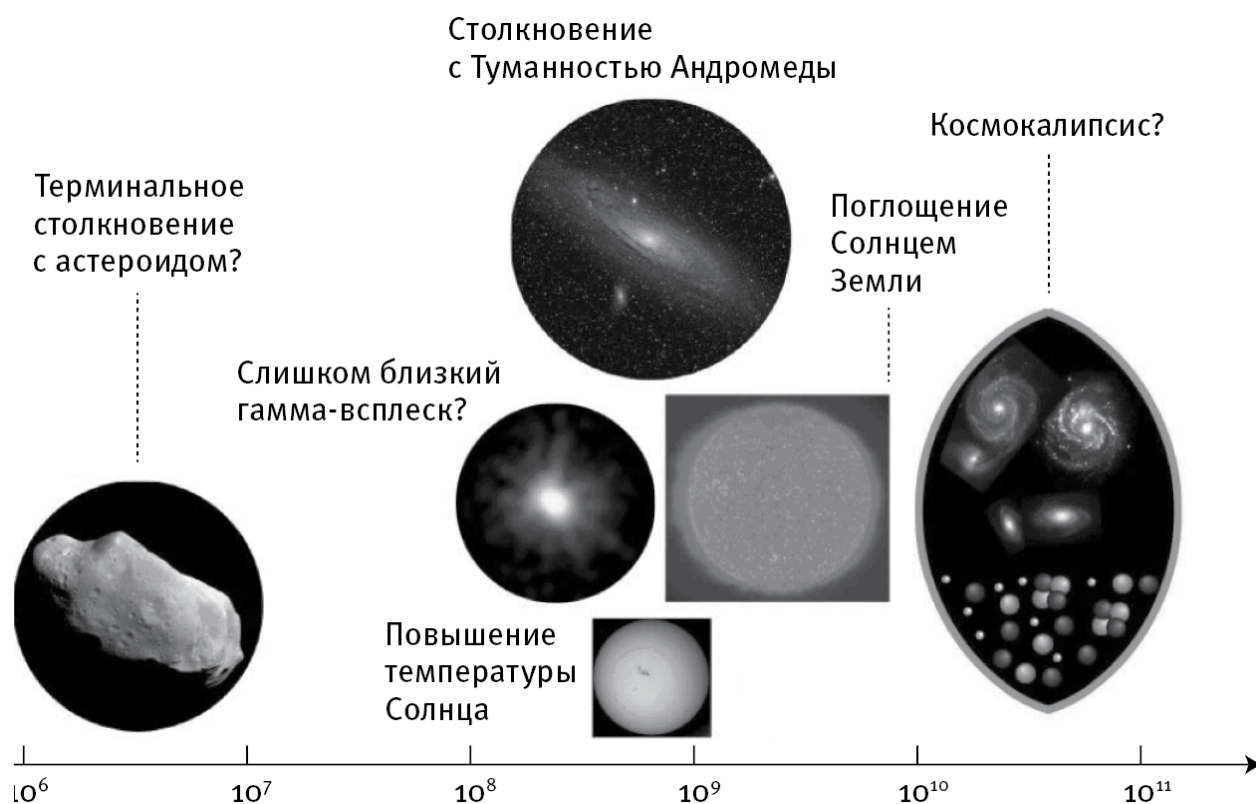
Рис. 5.1

Примеры, как может быть уничтожена известная нам жизнь или как может быть существенно снижен ее потенциал. При том, что наша Вселенная, по-видимому, просуществует еще десятки миллиардов лет, наше Солнце сожжет и проглотит Землю уже через миллиард лет, если только мы не отодвинем ее на безопасное расстояние, а еще через 3,5 миллиарда лет, даже если мы отодвинем ее на безопасное расстояние, наша галактика столкнется с соседней. Но даже не зная, как это случится, и задолго до этого, мы можем предсказать, что астероиды изрешетят нашу планету, а супервулканы вызовут долгие холодные зимы, когда на протяжении нескольких лет на небе не будет видно солнца. Мы можем использовать технологии для решения этих проблем, а можем использовать их для создания новых – таких, как климатические изменения, ядерные войны, искусственные пандемии или вышедший из-под контроля искусственный интеллект.

Самоистребление

После рассмотрения проблем, к которым могут привести новые технологии, важно остановиться и на проблемах, которые может вызвать отсутствие таких технологий. Давайте для начала рассмотрим те сценарии, где сверхразум не создается, а человечество уничтожает себя другими способами.

Как мы можем этого добиться? Простейшая стратегия – “просто ждать”. Хотя в следующей главе мы рассмотрим способы борьбы с такими неприятностями, как столкновение с астероидами и закипающие океаны, все они требуют технологий, пока еще не разработанных нами, так что, пока наши технологии не достигнут определенного уровня, значительно превышающего нынешний, мать-природа может отправить нас в небытие задолго до истечения следующего миллиарда лет. Как говорил знаменитый экономист Джон Мейнард Кейнс: “В долгосрочной перспективе мы все мертвы”.



К сожалению, у нас есть способы покончить с собой гораздо быстрее благодаря собственной коллективной глупости. Почему наш вид должен совершить коллективный суицид, называемый также *омницидом*, если практически никто по отдельности этого не хочет? На нашем нынешнем уровне интеллектуального развития и эмоциональной зрелости мы, люди, проявляем особый дар к просчетам, недомыслию и незнанию, отчего наша история полна несчастными стечениями обстоятельств, войнами и разными невзгодами, к которым, если разобраться, никто особенно не стремился. Экономисты и математики разработали красивую теорию игр, объясняющую, что подталкивает людей к поступкам, приводящим к катастрофическим последствиям для всех^[72].

Вы можете подумать: чем выше ставки, тем осмотрительнее ведут себя люди, но внимательное исследование нашего самого большого риска, допускаемого современной технологией, а именно – риска глобальной термоядерной войны, не очень-то обнадеживает. Лишь по воле случая мы пережили без последствий обескураживающе длинный список разного рода промахов: тут и компьютерные сбои, и отказы в системах энергоснабжения, и недопонимание проблемы, и навигационные ошибки, и столкновения бомбардировщиков, и взрывы спутников, и многое-многое другое^[73]. В самом деле, если бы не героические поступки отдельных выдающихся личностей, вроде Василия Архипова и Станислава Петрова, ядерная война была бы уже в нашем прошлом. Оценку вероятности случайного возникновения ядерного конфликта на протяжении года в одну тысячную при нынешнем положении вещей можно считать неправдоподобно оптимистичной, но даже она дает для оценки снизу вероятности такого конфликта в следующие 10 тысяч лет: $1 - 0,999^{10000} = 99,995 \%$.

Для полного понимания нашего безрассудства мы должны отдать себе отчет в том, что начали играть в свою ядерную рулетку еще до внимательного изучения связанных с ней рисков. Во-первых, опасность радиоактивного заражения оказалась существенно недооцененной, и в одних только Соединенных Штатах жертвам облучения при работе с ураном или во время ядерных испытаний было выплачено около 2 миллиардов долларов компенсации^[74].

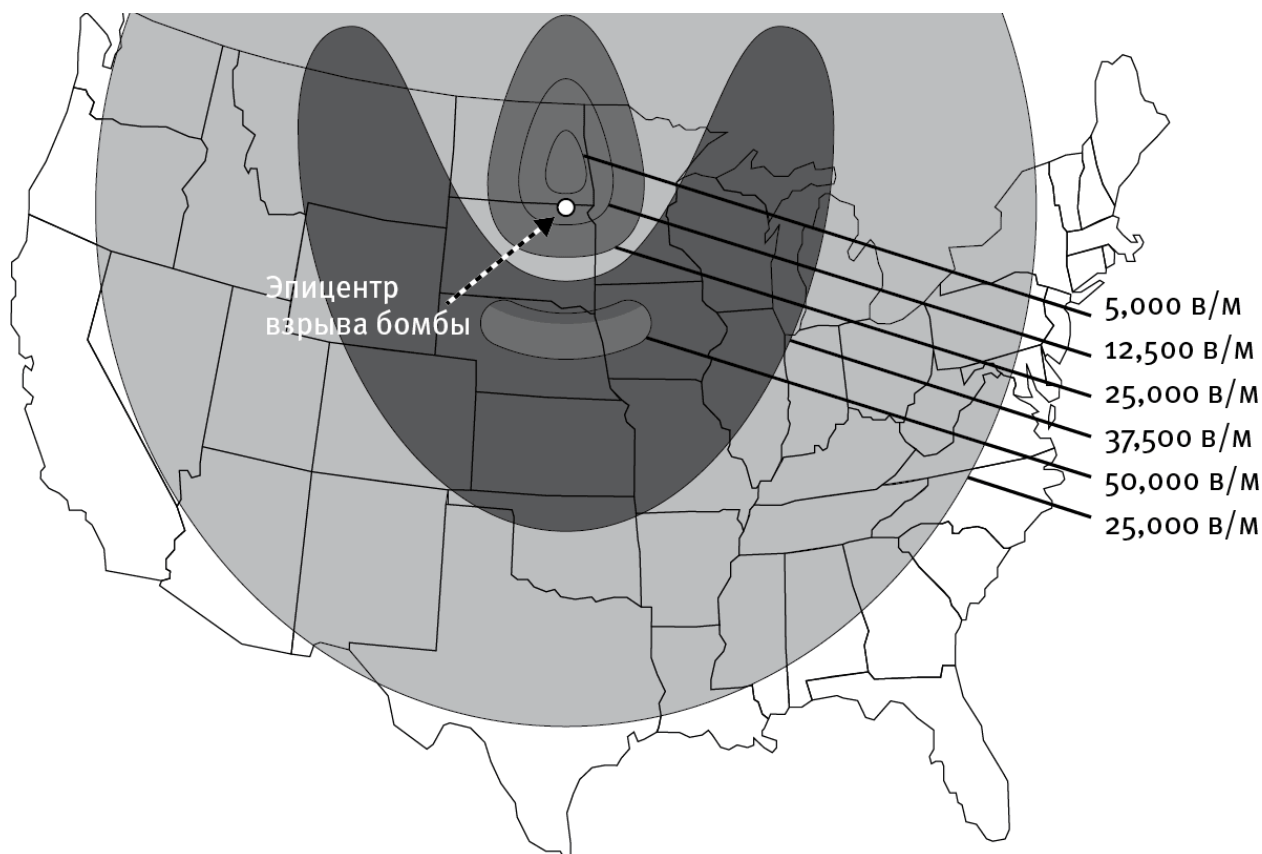


Рис. 5.2

Взрыв единственной водородной бомбы на высоте в 400 км над поверхностью Земли вызовет мощный электромагнитный импульс, и вся техника, использующая электрическую энергию, будет выведена из строя. При движении точки взрыва в юго-восточном направлении область, по форме напоминающая банан, где напряженность электрического поля превышает 37,5 киловольт на метр, накроет большую часть восточного побережья. Перепечатано из доклада Армии США ADA278230 с добавлением цветов.

Во-вторых, со временем выяснилось, что подрыв водородных бомб на высоте в сотни километров над поверхностью Земли создает мощный электромагнитный импульс (ЕМР), который немедленно нарушает работу линий электропередач и всех электронных приборов на обширных территориях (см. рис. 5.2), приводя к параличу инфраструктуры, забитым обездвиженными автомобилями трассам и далеким от идеальных перспективам на выживание. Например, Национальная комиссия электромагнитного мониторинга докладывала, что “система водоснабжения – это колоссальная машина, приводимая в действие отчасти силой тяготения, но в основном – электроэнергией”, и что перебои в водоснабжении приводят к гибели населения в течение 3–4 дней^[75].

В-третьих, возможность ядерной зимы обнаружилась лишь четыре десятилетия спустя после создания первых бомб, когда их стало уже 62 тысячи! Независимо от того, чьи города сгорели, пепел пожарищ поднимется в верхние слои тропосферы и распространится по всему земному шару, не позволяя солнечному свету достигать земной поверхности и превращая повсюду лето в зиму – так извержения вулканов и столкновения с астероидами приводили к массовым вымираниям в прошлом. Тревогу забили в 1980-х и советские, и американские ученые, и отчасти именно это заставило Рональда Рейгана и Михаила Горбачева приступить к сокращению ядерных арсеналов^[76]. К сожалению, более точные вычисления рисуют еще более печальную картину: на рис. 5.3 можно видеть понижение летних температур на 20 °С в основных сельскохозяйственных районах Америки, Европы, России и Китая (в некоторых частях России даже на 35 °С) в первые два лета после ядерной войны и потом еще на половину от этого на протяжении всего десятилетия^[36]. Что это означает, говоря простым языком? Не надо обладать богатым сельскохозяйственным опытом, чтобы понять: установившиеся на годы летние температуры, близкие к температуре замерзания воды, исключают возможность какого-либо получения провизии. Трудно предсказать в точности, что произойдет после того, как тысячи крупнейших городов Земли будут превращены в руины, а глобальная инфраструктура уничтожена, но как бы ни была мала часть выживших людей, чтобы уберечь их от голодной смерти, надо будет как-то противостоять вооруженным бандам, убивающим ради еды.

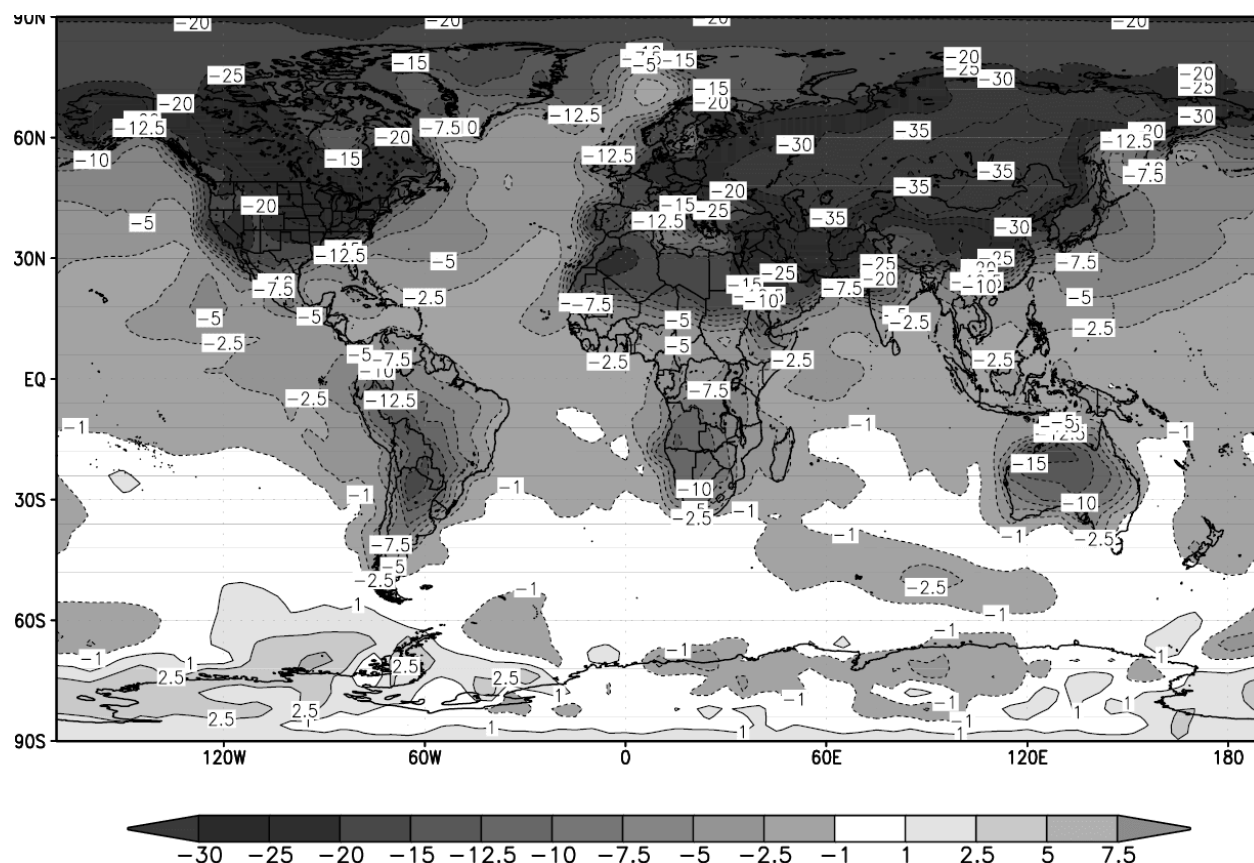


Рис. 5.3

Среднее похолодание (в градусах по шкале Цельсия) во время первых двух летних периодов после полномасштабного ядерного конфликта между США и Россией. Воспроизводится с разрешения Алена Робока^[77].

Я настолько подробно вошел в детали возможной ядерной войны, чтобы ясно показать: никакой разумный мировой лидер этого не хочет, но это все-таки может произойти случайно. Это означает, что мы не можем быть уверены в том, что человеческие собратья никогда не пойдут на омницид: того, что этого никто не хочет, еще недостаточно.

Машина Судного дня

Итак, могут ли люди и в самом деле повернуть омницид? Большинство ученых сходятся в том, что даже если в глобальной ядерной войне будут уничтожены 90 % людей, в ней не будут уничтожены все 100 %, и полного вымирания за этим еще не последует. С другой стороны, истории, связанные с ядерной зимой, электромагнитными импульсами и радиационным заражением, показывают, что самыми большими опасностями могут оказаться как раз те, о которых мы пока еще не подумали. Очень трудно предугадать все возможные последствия, сочетающие ядерную зиму, разрушение инфраструктуры, выросший уровень мутаций и вооруженные банды отчаявшихся людей на фоне других возникших проблем – таких, как пандемии, разрушенная экосистема и еще что-то такое, о чем мы пока не догадываемся. По моей собственной оценке, хотя вероятность ядерной войны, которая бы уже завтра запустила механизм неизбежного вымирания человечества, невысока, мы все же не можем с полной уверенностью считать ее равной нулю.

Шансы омницида сильно возрастают, если мы преобразуем существующее ныне ядерное оружие в машину Судного дня. Идею предложил стратег Корпорации RAND Герман Кан в 1960 году, а Стэнли Кубрик популяризировал ее в фильме *Dr. Strangelove*,

после чего машина Судного дня превратилась в парадигму гарантированного взаимного уничтожения. Получается идеальный сдерживающий фактор: машина, которая автоматически возвращает удар любого противника, уничтожая все человечество.

Один из возможных вариантов машины Судного дня – это просторное подземное хранилище, где сложены, как их иногда называют, “соленые ньюки” – по возможности крупные водородные бомбы, обложенные большим количеством кобальта. Физик Лео Сцилард еще в 1950 году утверждал, что такое способно убить все на Земле: взрывы бомб сделают кобальт радиоактивным и выбросят его в стратосферу. Пятилетний период его полураспада достаточно велик, чтобы он распределился над всей поверхностью Земли (в особенности если две одинаковые машины будут расположены симметрично в противоположных полусферах), и достаточно короток, чтобы интенсивность радиации была смертельной. В СМИ сообщается, что сейчас кобальтовые бомбы впервые стали изготавливаться. Омницидальные достоинства могут быть усилены добавлением бомб, оптимизированных под ядерную зиму: они должны распылить в стратосфере долгоживущие аэрозоли. Ключевой момент в истории с машиной Судного дня – в ее относительной дешевизне при сравнении с обычным ядерным сдерживанием: поскольку эти бомбы не надо куда-то доставлять, не нужны дорогостоящие ракетные системы, да и сами бомбы можно сделать попроще, коль скоро они не должны быть ни легкими, ни компактными для установки на ракетах.

Еще одну возможность для создания машины Судного дня дает биологическое оружие: в ней будут содержаться специально выращенные бактерии или вирусы, уничтожающие всех людей. Если их трансмиссивность будет достаточно высока, а инкубационный период достаточно продолжителен, то почти все успеют подхватить заразу, прежде чем о ее существовании станет известно и можно будет принять контрмеры. Есть веский военный довод в пользу создания такого биологического оружия, даже если оно не уничтожит всех: наиболее эффективная машина Судного дня должна сочетать в себе различные поражающие факторы – и ядерный, и биологический, и, по возможности, какой-нибудь еще, чтобы увеличивать сдерживающую врага силу.

AI-оружие

Третий технологический путь к омнициду может сопровождаться созданием в некотором роде необычного оружия с использованием искусственного интеллекта. Предположим, некая сверхдержава создает миллиарды тех самых атакующих дронов размером со шмеля, которые мы описывали в главе 3, и решает с их помощью уничтожить всех людей, кроме своих граждан и союзников, которые удаленно идентифицируются с помощью радиочастотных id-тэгов, какими сейчас помечаются товары в большинстве супермаркетов. Эти тэги распространяют среди граждан, чтобы те носили их либо как браслеты, либо как подкожные импланты, упоминавшиеся в связи с тоталитарными методами. Вражеская сверхдержава, наверное, тоже решит создать нечто подобное. Когда случайно разразится война, все люди будут убиты, даже члены самых далеких племен, не имеющие никакого отношения к конфликту, потому что никто не станет носить тэги обоих видов. Сочетание этого способа с биологической и ядерной машинами Судного дня благоприятствует дальнейшему увеличению шансов успешного омницида.

А что хотите вы?

Мы начинали эту главу с размышлений о том, к чему бы вы хотели, чтобы вела нас AGI-гонка. И теперь, когда перед вами была развернута широкая панорама различных сценариев, одного за другим, самое время снова спросить: какой из них вас больше привлекает, а какого мы должны изо всех сил избегать? У вас есть явный фаворит? Пожалуйста, сообщите об этом мне и другим читателям на странице <http://AgeOfAi.org> и присоединяйтесь к нашей дискуссии!

Сценарии, которые были описаны, очевидно, не исчерпывают всех возможностей, и некоторым из них недостает важных деталей, но я приложил много усилий, чтобы сделать список как можно более полным, включив в него широкий набор вариантов, от высокотехнологичных до низкотехнологичных и нетехнологичных, и описывая попутно все главные страхи и упования, какие встречаются в литературе.

Во время работы над этой книгой забавно было слышать от друзей и коллег, что они думают по поводу того или иного сценария, и отсутствие какого бы то ни было согласия среди них немало меня развеселило. Единственное, в чем все они сошлись, это в том, что сделать выбор в итоге оказалось намного сложнее, чем они думали вначале. Люди, которым нравился какой-то один определенный сценарий, непременно находили в нем что-нибудь настораживающее. Для меня это означало, что мы, люди, должны продолжать разговор о своих будущих целях, делая его детальнее и глубже, чтобы понимать, в каком направлении вращать свой штурвал. Перспективы переноса нашей жизни в космос воодушевляюще высоки, и давайте не будем относиться к этому беспечно, словно мы находимся на неуправляемом корабле и не представляем, куда бы мы хотели его направить!

Но как же все-таки велик этот будущий потенциал? Как бы далеко ни заходили наши технологии, способность Жизни 3.0 к самосовершенствованию и к распространению в космосе по-прежнему будет ограничена законами физики. Но где именно пройдут эти границы? Есть ли внеземная жизнь в космосе уже сейчас, или мы в нем одиноки? Что произойдет, когда различные цивилизации, претендующие на экспансию в космосе, встретятся? Мы займемся этими увлекательными вопросами в следующей главе.

Подведение итогов

- Если когда-нибудь нам удастся создать AGI человеческого уровня, за этим может последовать интеллектуальный взрыв, и мы в результате останемся далеко позади.
- Если группа людей совладеет с управлением этим интеллектуальным взрывом, то эти люди смогут в течение нескольких лет взять в свои руки власть над всем миром.
- Нынешние настойчивые попытки создать AGI в случае успеха могут иметь поразительно широкий спектр последствий, складывающихся в несколько типических сценариев на предстоящие тысячелетия.
- Сверхразум может мирно сосуществовать с людьми, либо потому, что принужден к этому (сценарий плененного бога), либо потому, что так хочет его “дружественный искусственный интеллект” (сценарии либертарианской утопии, бога-защитника, благодетельного диктатора или зоопарка).
- Создание сверхразума может быть предотвращено или искусственным интеллектом (сценарий стража), или людьми (сценарий 1984) путем намеренного забвения результатов технологического прогресса (сценарий возврата) или исключением стимулов к его созданию (сценарий эгалитарной утопии).
- Человечество может пойти по пути полного вымирания и либо оказаться замененным AGI (сценарии благодарных потомков и завоевателей), либо исчезнуть без следа (сценарий самоуничтожения).
- Нет абсолютно никакого согласия относительно того, какой из этих сценариев был бы для нас желательным, и вообще – есть ли такой среди них. У каждого сценария находятся свои отрицательные стороны. Тем более необходимо продолжать обсуждение наших будущих целей и углублять повестку такого обсуждения, разговор вокруг, чтобы мы случайно не двинулись в нежелательном направлении.

Глава 6

Наше космическое благосостояние: следующий миллиард лет и все что потом

Наши спекуляции заканчиваются на сверхцивилизации, жизненном синтезе всей Солнечной системы, постоянно самосовершенствующейся и распространяющейся все дальше от Солнца, превращающей в мысль все, что не было жизнью.

Ганс Моравец, Mind Children

На мой взгляд, наиболее вдохновляющее из когда бы то ни было сделанных научных открытий заключается в том, что мы хронически недооцениваем будущий потенциал жизни. Нет нужды ограничивать наши устремления и мечты отрезком в сотню лет, отведенных для жизни среди болезней, нищеты и невежества. Напротив, благодаря развитию технологий жизнь получает возможность процветать миллиарды лет, и не только здесь, в Солнечной системе, но и по всему космосу, гораздо более интересному и необъятному, чем могли себе представить наши предки. Ему даже небо не предел.

Это потрясающая новость для биологического вида, которому удастся отодвинуть границы своего существования на много веков. Олимпийские игры замечательны тем, что отодвигают границы человеческих возможностей – силы, скорости, ловкости и выносливости. Наука раздвигает границы в познании и понимании мира. Литература и другие виды искусства открывают нам новые границы в создании прекрасного и обогащают жизненный опыт. Многие люди, организации и нации знамениты своим вкладом в разработку новых ресурсов, территорий и возможностей продления жизни. Принимая во внимание нашу одержимость в преодолении всевозможных границ, новый результат вполне достоин внесения в одну из самых продаваемых книг, защищенных копирайтом, – *Книгу рекордов Гиннеса*.

Таким образом, если известные нам ранее границы нашей жизни могут быть раздвинуты с помощью технологий, то каковы же могут быть ее настоящие пределы? Сколь большая часть космоса может быть пригодна для жизни? Как много вещества жизнь может обратить себе на пользу? Как много энергии и информации может она из него извлечь, как много вычислений с его помощью произвести? Эти окончательные пределы должны не возникать из наших догадок и логических построений, а определяться законами физики. И отсюда, как ни странно, нам легче понять далекую перспективу будущей жизни, чем близкую.

Если всю нашу 13,8-миллиардолетнюю космическую историю сжать в одну неделю, то все 10 тысяч лет последних двух глав займут в ней не более полусекунды. Это означает, что, хотя мы и не можем предсказать, как будет развиваться интеллектуальный взрыв, случится ли он вообще и какими могут быть его непосредственные последствия, вся эта заваруха окажется не более чем едва заметным мерцанием в космической истории, подробности которого никак не отразятся на окончательных границах жизни. Если жизнь после взрыва будет так же озабочена отодвиганием границ, как наша нынешняя, то она и станет заниматься созданием таких технологий, чтобы границы окончательно заняли свое место, – потому что для нее это возможно. В этой главе мы разберемся, где проходят эти границы, и попробуем навскидку понять, на что они могут быть похожи. Ибо положение этих границ зависит от нашего нынешнего понимания физики, то есть мы взглянем на нижнюю границу возможностей: научные открытия будущего могут открыть возможности для дальнейшего ее смещения.

Но знаем ли мы наверняка, что будущая жизнь и в самом деле будет столь амбициозна? Нет, мы этого не знаем. Может быть, она станет самодовольной, словно героинист или лежебока, раз за разом пересматривающий телесериал *Keeping Up with the Kardashians*. Но у нас, тем не менее, есть основания подозревать, что амбициозность для развитых форм жизни – это нечто вроде генетической особенности. И почти независимо от того, в чем именно она пытается достичь предела – в разуме, продолжительности жизни, познании или острых ощущениях, ей понадобятся ресурсы. А поэтому ей придется развивать технологии, доводя их возможности до окончательного предела и занимая все большие и большие пространства в космосе.

Мы можем также предположить, что жизнь независимо возникает во множестве разных мест в космосе. В этом случае неамбициозные цивилизации просто станут космически несущественны, а самую привлекательную часть космоса будут захватывать наиболее амбициозные формы жизни. Естественный отбор станет работать уже в космическом масштабе, и рано или поздно все оставшиеся формы жизни будут амбициозными. В итоге, интересуясь пределами, в которых наш космос может стать в конце концов обитаемым, мы должны изучить пределы этих амбиций, допустимые законами физики. Так давайте сделаем это! Давайте сначала посмотрим, что можно сделать с ресурсами – веществом, энергией и т. п., какие у нас есть в Солнечной системе, а затем посмотрим, до каких ресурсов мы можем добраться благодаря космическим исследованиям и космическим поселениям.

На что годятся ваши ресурсы

В то время как сейчас супермаркеты и товарные биржи предлагают тысячи тонн разного всякого, что мы можем называть “ресурсами”, будущая жизнь, достигшая технологического предела, зависит лишь от одного фундаментального ресурса: так называемого барионного вещества, – под этим подразумевается все что угодно, состоящее из атомов и его компонентов (кварков и электронов). Какую форму этого вещества ни взять изначально, развитая технология может превратить его в любую желаемую субстанцию и любой желаемый объект: силовую электростанцию, вычислительную машину или высокоорганизованное биологическое существо. Поэтому давайте начнем с ограничений на источники энергии, необходимой для развитых форм жизни, и информационных процессов, позволяющих им мыслить.

Построение сфер Дайсона

Когда дело доходит до будущего жизни, я предпочитаю Фримена Дайсона – одного из самых обнадеживающих визионеров. Для меня и счастье, и честь быть знакомым с ним лично уже более двух десятилетий, но когда мне предстояло встретиться с ним в первый раз, я очень волновался. Я был тогда молодым специалистом, недавно защитившим диссертацию, и мы с друзьями сидели в столовой Института перспективных исследований в Принстоне. Вдруг, откуда ни возьмись, этот всемирно известный ученый, которому случалось быть близко знакомым с Эйнштейном и Гёделем, подошел к нашему столу и попросил разрешения к нам присоединиться. Мы быстро почувствовали себя в его компании очень непринужденно, как только он пояснил, что обедать с молодежью ему гораздо интереснее, чем со старичьем – профессорами. Хотя в тот момент, когда я пишу эти слова, ему девяносто три, в душе он намного моложе большинства моих знакомых, и озорной мальчишеский блеск в его глазах по-прежнему показывает, как мало он переживает по поводу формальностей, академических иерархий и сложившихся традиций. Чем дерзновеннее идея, тем больше она его привлекает.

Когда мы впервые заговорили об использовании энергии, он принялся высмеивать недалекое и ленивое человечество, утверждая, что все его нынешние энергетические потребности можно покрыть, просто собрав солнечное излучение с площади меньшей 0,5 % площади пустыни Сахары. Но зачем останавливаться на этом? Зачем ограничиваться энергией, приходящей от Солнца к Земле, для чего позволять всей остальной энергии без толку распыляться в космическом пространстве? Почему бы не собрать всю солнечную энергию и не использовать ее для жизни?

Вдохновленный романом Олафа Стэплдона *Star Maker*^[37] 1937 года, где описывались кольца искусственных миров, обращающихся вокруг родительской звезды, Дайсон опубликовал в 1960 году то, что стало потом называться “сферой Дайсона”^[78]. Его идея заключалась в том, чтобы превратить Юпитер в биосферу, окружающую Солнце вроде ореховой скорлупы, где наши потомки смогут наслаждаться в 100 миллиардов раз большим количеством биомассы и в триллион раз большим количеством энергии, чем есть в нашем распоряжении сейчас^[79]. Он утверждал, что это естественный следующий шаг: “Следует ожидать, что по прошествии нескольких тысяч лет с момента своего вхождения в фазу индустриального развития любой биологический вид окажется помещенным в искусственную биосферу, которая полностью окружает его родительскую звезду”. Если вы живете на внутренней стороне сферы Дайсона, у вас не бывает ночи: вы всегда видите Солнце прямо над головой, а солнечный свет – льющим со всех сторон, так как его отражает вся поверхность сферы: точно так же сейчас вы можете видеть

солнечный свет, отраженный поверхностью Луны. Если вам захочется посмотреть на звезды, надо будет просто “подняться” и выйти на внешнюю сторону сферы Дайсона.

Низкотехнологичный способ создания частичной сферы Дайсона заключается просто в том, чтобы разместить кольцо обитаемых секций на круговой околосолнечной орбите. Чтобы окружить Солнце полностью, надо будет построить много колец, расположенных и в разных плоскостях, и на немного различающихся расстояниях от Солнца, – избегая столкновений между ними. Чтобы уклониться от неприятностей, связанных с невозможностью сцепить эти быстро движущиеся по отношению друг к другу кольца, что сильно затрудняет перемещения и коммуникацию между ними, можно вместо этого построить монолитную стационарную сферу Дайсона, где направленная внутрь сила притяжения со стороны Солнца будет уравниваться исходящим от него световым давлением – эту идею впервые высказали Роберт Л. Форвард и Колин Мак-Иннес. Такую сферу можно строить постепенно, добавляя все новые и новые стационарные спутники, статиты, уравнивающие солнечное гравитационное поле за счет давления света, а не центробежной силы, как обычно. Обе эти силы убывают обратно пропорционально квадрату расстояния от Солнца, и это означает, что если равновесие достижимо на одном расстоянии, то оно будет достигнуто и на любом другом, а это позволит расположить сферу практически где угодно. Статиты должны быть исключительно легкими – с удельным весом, не превышающим 0,77 граммов на квадратный метр, что примерно в 100 раз легче бумаги, но в этом нет ничего диковинного. Например, лист графена (моноатомарный слой углерода в гексагональной упаковке – вроде проволочного забора для кур) весит в тысячу раз меньше. Если сфера Дайсона строится в основном для того, чтобы отражать солнечный свет, а не поглощать его, то суммарная интенсивность его внутри феноменально повысится, раздувая сферу изнутри и увеличивая массу, которую она сможет выдерживать. Есть много звезд, светимость которых в тысячи и даже миллионы раз превышает солнечную, и они будут выдерживать соответственно бо́льшие массы на стационарной сфере Дайсона.

Если намного более тяжелая твердая сфера Дайсона нужна здесь, в Солнечной системе, то, чтобы выдерживать силу притяжения Солнца, понадобятся ультрапрочные материалы, способные не плавиться и не изгибаться при нагрузках в тысячи раз больших, чем в фундаменте небоскребов. Чтобы выдержать долгосрочное использование, сфера Дайсона должна быть динамической и интеллектуальной, всегда готовой точно подстроить свое положение и форму в ответ на разные возмущения или открыть нужный пролет, пропуская коварный астероид или комету, угрожающие что-нибудь разрушить. Для борьбы с такими нежелательными гостями можно также использовать системы типа “обнаружить и отклонить” – отклонив, их можно будет разрушить, а затем применить составляющее их вещество с большей пользой.

Для нынешнего человечества жизнь на сфере Дайсона или внутри нее была бы в лучшем случае непривычной, а в худшем – невозможной, но это не должно удерживать будущие биологические или небιологические формы жизни от стремления к этому. При вращении на орбите гравитации практически не будет, и если вы захотите прогуляться по стационарному ее варианту, не сваливаясь, вы сможете сделать это только на внешней поверхности противоположной Солнцу, и сила притяжения там будет в десять тысяч раз слабее той, к которой вы привыкли. У вас не будет никакого магнитного поля (если только вы не создадите его сами), которое защищало бы вас от опасных частиц, летящих от Солнца. Но есть и хорошие новости: на сфере Дайсона, расположенной на нынешней земной орбите, у нас будет в 500 миллионов раз больше места, чем сейчас.

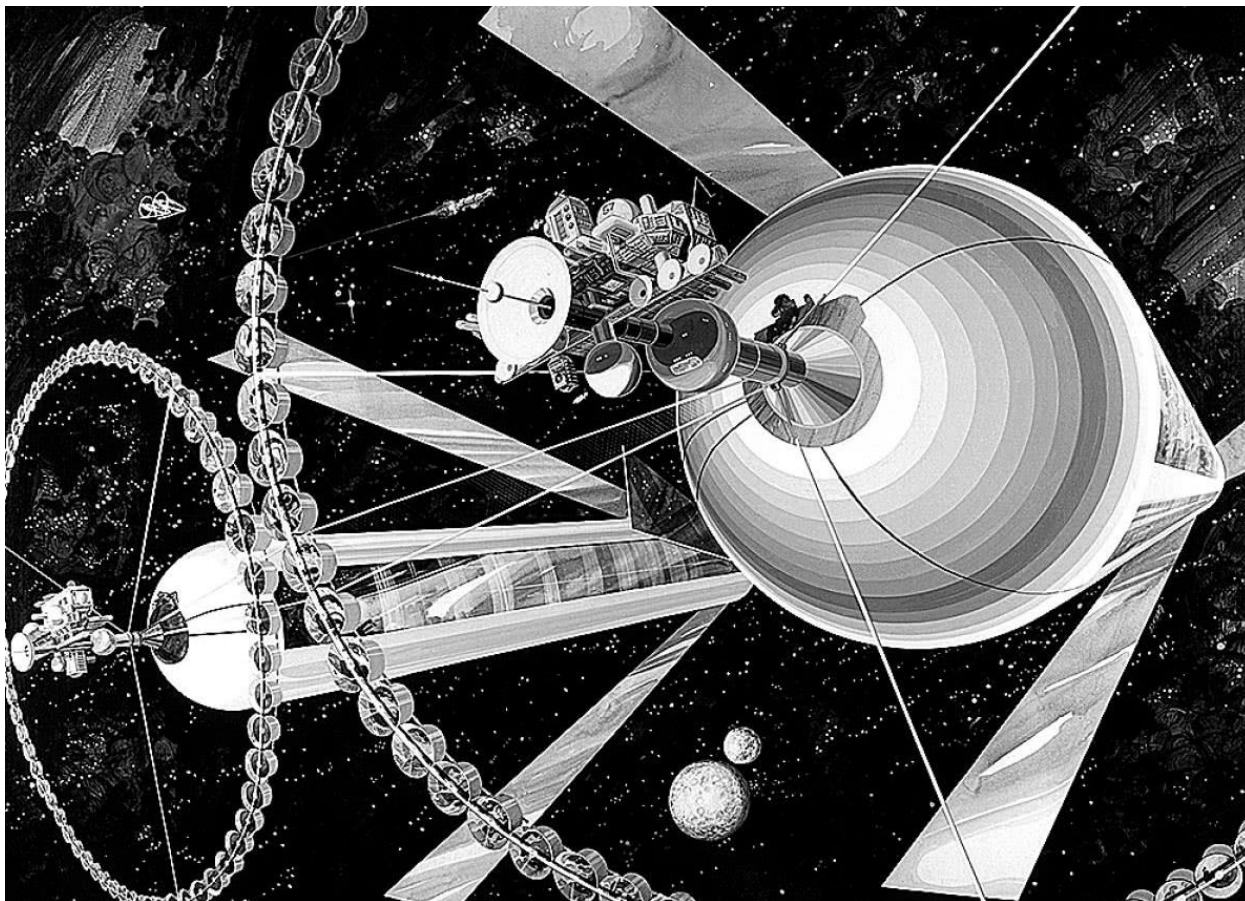


Рис. 6.1

Пара вращающихся в противоположных направлениях цилиндров О'Нила обеспечивает людям комфортное обитание с силой гравитации как на Земле. Надо только держать плоскости вращения ориентированными строго на Солнце. Искусственная гравитация создается во вращающихся цилиндрах центробежной силой, а поворачивающиеся зеркала направляют солнечные лучи так, чтобы день и ночь чередовались по обычному 24-часовому циклу. Малые цилиндры в кольцах предназначены для сельского хозяйства. Изображение предоставлено Риком Гвидиче (NASA).

Если мы хотим иметь более привычные и похожие на земные места обитания, то нам есть на что надеяться: построить их значительно легче, чем сферу Дайсона. Например, на рис. 6.1, 6.2 показан проект цилиндрических станций, спроектированных американским физиком Герардом О'Нилом, на которых будет искусственное гравитационное поле, защита от космических лучей, 24-часовой световой цикл, искусственная земная атмосфера и земная экосистема. Такие станции могут свободно вращаться на орбите внутри сферы Дайсона, или их модифицированные варианты могут быть прикреплены к ней снаружи.

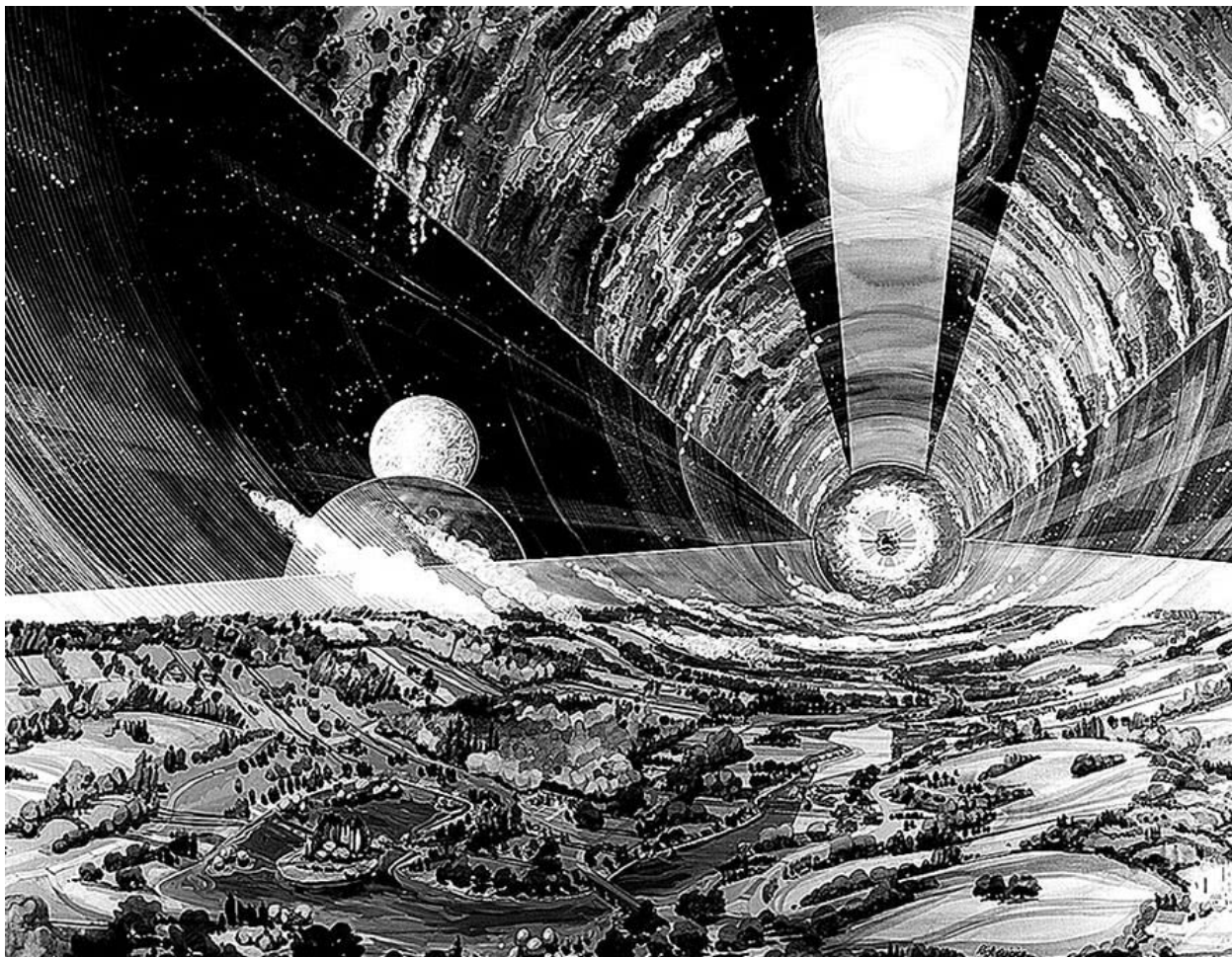


Рис. 6.2

Один из цилиндров О'Нила с предыдущего рисунка, вид изнутри. При диаметре в 6,4 км и скорости вращения 1 оборот каждые 2 минуты центробежная сила на внутренней поверхности цилиндра окажется равной силе тяжести на Земле. Солнце сейчас у вас прямо за спиной, но кажется, что оно сверху, потому что его лучи идут от расположенного за пределами цилиндра зеркала. На ночь зеркало поворачивается так, что солнечные лучи не попадают внутрь цилиндра. Герметически закрывающиеся входы и выходы не дают атмосфере просочиться наружу. Изображение предоставлено Риком Гвидиче (NASA).

Табл. 6.1.

Эффективность превращения вещества в энергию в сопоставлении с теоретическим пределом в $E = mc^2$. Как объясняется в основном тексте, получение 90 % всей энергии возможно при испарении черной дыры, но это слишком медленный процесс, чтобы был шанс воспользоваться им на практике, а любое ускорение его ведет к существенному снижению эффективности.

Метод	Эффективность
Переваривание шоколадного батончика	0,00000001%
Сжигание каменного угля	0,00000003%
Деление урана-235	0,08%
Использование сферы Дайсона до гибели Солнца	0,08%
Водородный синтез гелия	0,7%
Вращающаяся черная дыра	29%
Сфера Дайсона вокруг квазара	42%
Сфалерайзер	50%?
Испарение черной дыры	90%

Строительство более совершенных электростанций

Хотя сферы Дайсона по современным стандартам высокопроизводительны в энергетическом отношении, они ни в чем не выходят за пределы, определенные законами физики. Эйнштейн учил нас, что если мы можем превратить массу в энергию со 100-процентной эффективностью^[38], то имеющаяся масса m на входе даст нам на выходе энергию E , количество которой будет определяться формулой:

$$E = mc^2,$$

где c – скорость света. Это означает, что, так как скорость света огромна, то небольшой массы достаточно, чтобы произвести очень большое количество энергии. Если бы мы в изобилии располагали антиматерией (которой у нас совсем нет), то построить электростанцию со 100-процентной эффективностью было бы совсем просто: просто подливая по чайной ложке антиматерию в бак с водой, мы высвобождали бы каждый раз столько же энергии, сколько высвобождает взрыв 200 тысяч тонн тротила или более или менее типичной водородной бомбы, чего вполне достаточно на 7 минут для всей планеты.



Рис. 6.3

Более совершенные технологии позволят извлекать из вещества существенно больше энергии, чем мы можем получить перевариванием пищи или сжиганием топлива, и даже водородный синтез позволяет извлечь лишь небольшую часть энергии – в 140 раз меньше, чем допустимо с точки зрения законов физики. Силовые станции, использующие сфалероны, квазары или испаряющиеся черные дыры, позволяют существенно улучшить энергоотдачу.

В отличие от этого любой из способов получения энергии, применяемых нами сейчас, прискорбно неэффективен, о чем можно судить по данным табл. 6.1 и рис. 6.3. Переваривание шоколадного батончика эффективно всего на 0,00000001 % – в том смысле, что при этом высвобождается лишь одна триллионная часть от mc^2 – содержащейся в нем энергии. Будь ваш желудок эффективен хотя бы на 0,001 %, одного обеда вам бы хватило до конца жизни. По сравнению с тем, как мы едим, сжигание угля или бензина всего лишь в 3 и 5 раз эффективнее соответственно. Современный атомный реактор, расщепляя ядра урана, производит энергию гораздо продуктивнее, но и он пока не может добыть больше 0,08 % от той, что там есть. Термоядерный реактор в ядре Солнца на порядок более эффективен, чем все, что мы построили, – он добывает 0,7 % содержащейся в ядрах водорода энергии за счет слияния их друг с другом и превращения в ядра гелия. И даже если мы когда-нибудь заключим Солнце в идеальную сферу Дайсона, мы никогда не сможем превратить в полезную для нас энергию больше 0,08 % его массы, потому что как только Солнце израсходует примерно одну десятую содержащегося в нем водорода, оно завершит свою жизнь нормальной звезды и разрастется в красного гиганта,

а потом начнет медленно умирать. И для других звезд дела обстоят не лучше: доля водорода, который им удастся израсходовать за время своей нормальной жизни, колеблется от 4 % для самых маленьких звезд до 12 % – для самых больших. Если мы сделаем совершенный ядерный реактор, который будет на 100 % синтезировать весь попадающий в него водород в гелий, мы и в этом случае застрянем на обидно низком показателе в 0,7 % эффективности ядерного синтеза. Что бы такое придумать получше?

Испарение черных дыр

В своей книге *A Brief History of Time*^[39] Стивен Хокинг описал электростанцию, работающую на черных дырах^[40]. Это, возможно, звучит парадоксом, если вспомнить, что в черной дыре, как считалось долгое время, все, однажды туда попавшее, застревает навеки, и даже свет не может ее покинуть. Однако, как известно, Хокинг сумел рассчитать квантово-гравитационный эффект, благодаря которому черная дыра ведет себя как горячее тело, – причем чем меньше, тем горячее. Это излучение так и стали называть *излучением Хокинга*. Излучая, черная дыра теряет свою энергию, пока не испарится совсем. Другими словами, какое бы вещество вы ни засунули внутрь черной дыры, оно со временем вылезет обратно в виде теплового излучения, и к тому моменту, когда черная дыра испарится окончательно, все ваше вещество превратится в излучение практически со 100-процентной эффективностью^[41].

Проблема с использованием черных дыр в качестве источников энергии, однако же, такова: пока размер дыры не сравняется с размером атома или даже не станет значительно меньше его, на что уйдет больше времени, чем нынешний возраст Вселенной, а пока этого не случится, излучение очень слабое: энергии выделяется не больше, чем от обычной свечки. При уменьшении размера черной дыры излучение растет пропорционально его квадрату, поэтому физики Луис Крейн и Шон Вестмореланд предложили использовать дыры примерно в тысячу раз меньшего размера, чем протон, и по весу примерно равные самому большому океанскому лайнеру^[80]. Главное для них заключалось в том, чтобы использовать энергию такой дыры для космического корабля (сюжет, к которому мы еще вернемся), поэтому их интересовала не столько эффективность, сколько портативность, и они предлагали кормить дыру лазерным лучом, исключая всякую возможность превращения энергии в вещество. Но даже если вы будете кормить ее веществом, а не излучением, гарантировать высокую эффективность трудно: чтобы заставлять протоны проваливаться в черную дыру размером в одну тысячную от их собственного, нужна машина вроде Большого адронного коллайдера, увеличивающая их энергию mc^2 хотя бы в тысячу раз за счет кинетической энергии (энергии движения). А так как по меньшей мере 10 % этой кинетической энергии потеряется с гравитонами, когда черная дыра испарится, то окажется, что нам приходилось вкладывать в нее гораздо больше энергии, чем мы получим в конце, то есть ее эффективность и вовсе будет отрицательной. Более подробное изучение возможностей использования черной дыры в качестве источника энергии упирается в то, что для этого нам нужна квантовая теория гравитации, которая так до сих пор и не создана, – но эта неопределенность означает также то, что в ней может скрываться полезный квантово-гравитационный эффект, который нам пока неизвестен.

Вращающаяся черная дыра

К счастью, есть другие способы получения энергии из черных дыр, которыми можно воспользоваться, не прибегая к квантовой гравитации или какой-нибудь другой, пока еще не очень понятной, физике. Например, многие известные черные дыры очень быстро

вращаются, горизонт событий у них закручивается почти до скорости света, вот эту-то энергию и можно извлечь. Горизонт событий черной дыры – это та самая область пространства, где силы гравитации так велики, что даже свет не может ее покинуть. Рис. 6.4 показывает, что над горизонтом событий у вращающейся черной дыры есть так называемая *эргосфера* – область, где вращающаяся дыра захватывает окружающее ее пространство, а вместе с ним и любую попавшую туда частицу, у которой теперь нет возможности оказаться в состоянии покоя. Если вы забросите какой-нибудь объект в эргосферу, он немедленно будет подхвачен и закружится вокруг дыры. К сожалению, он вскоре будет поглощен черной дырой, навсегда сгинув под ее горизонтом событий, и пользы от него в смысле получения энергии никакой не будет. Однако Роджер Пенроуз показал, что если вы кинете его под правильным углом и при этом расколете на два, как показано на рис. 6.4, то поглощен будет только один осколок, в то время как другой вырвется на свободу с энергией больше той, которую вы ему изначально сообщили. Иными словами, вы успешно превратите часть энергии вращения черной дыры во что-то полезное, что можно будет потом потратить на совершение работы. Прodelывая этот фокус много раз, вы можете выдоить из черной дыры всю ее энергию вращения, в результате чего она остановится, а эргосфера у нее пропадет. Если черная дыра вращается так быстро, как только ей позволяют законы природы, и ее горизонт событий фактически движется со скоростью света, то такая стратегия позволит вам превратить в энергию до 27 % ее массы. Тут все еще есть некоторая неопределенность, связанная с нашим незнанием того, насколько быстро вращаются черные дыры в нашем ночном небе, но некоторые наиболее изученные из них, кажется, должны вращаться довольно быстро – от 30 % до 100 % от допустимого максимума скорости. Монструозная черная дыра в центре нашей галактики (она весит в четыре миллиона раз больше нашего Солнца), по всей видимости, вращается так, что если даже превратить в энергию всего 10 % ее массы, получится то же самое, что при превращении в энергию массы 400 тысяч солнц со 100-процентной эффективностью, или примерно столько же энергии, сколько можно получить с помощью сфер Дайсона вокруг 500 миллионов солнц за миллиард лет.

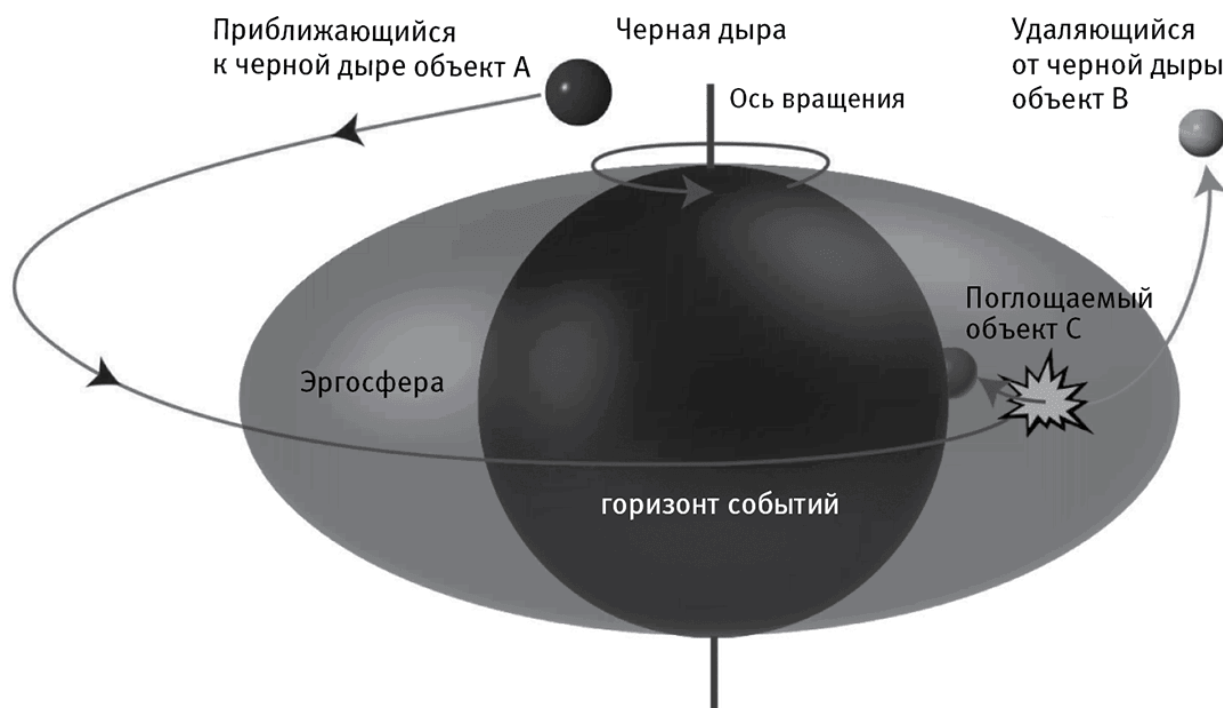


Рис. 6.4

Часть энергии вращения черной дыры можно получить с помощью довольно простого процесса: если запустить частицу А между двух горизонтов, то она расщепится

там на две, В и С, одна из которых (С) провалится в черную дыру, а другая (В) – вылетит наружу с энергией, превосходящей начальную энергию частицы А.

Квазары

Еще одна интересная стратегия заключается в том, чтобы извлекать энергию не из самих черных дыр, а из вещества, в них падающего. Природа уже и сама нашла способ это делать – это квазары. По мере того как газ, закручиваясь, все больше приближается к черной дыре – при этом возникает что-то похожее на гигантскую пиццу, внутренняя часть которой дырой постепенно заглатывается, – он разогревается и исторгает колоссальное излучение. Проваливаясь в дыру, частицы газа разгоняются, поскольку энергия сил притяжения превращается в энергию движения, как у парашютистов. Само движение становится все более беспорядочным, возникающая турбулентность нарушает координированное движение газовых слоев, разбивая его на быстрые и не зависящие друг от друга колебания все меньших частей, пока, наконец, дробление не доходит до отдельных атомов, которые принимаются биться друг о друга со страшной скоростью – собственно, эти беспорядочные столкновения быстро движущихся частиц и означают, что газ сильно разогрелся, и теперь уже энергия столкновений превращается в энергию излучения. Построив сферу Дайсона на безопасном расстоянии от черной дыры, все это излучение можно будет собрать и пустить в дело. Чем быстрее крутится черная дыра, тем на большую эффективность процесса можно рассчитывать, а для черных дыр, вращающихся с максимальной скоростью, эффективность достигает аж целых 42 %^[42]. У черных дыр с массой как у звезд большая часть энергии излучается в виде рентгеновских лучей, а у супермассивных звезд, находящихся в центрах галактик, максимум приходится на инфракрасный, видимый или ультрафиолетовый диапазон.

Если горючего, чтобы топить вашу черную звезду, у вас больше нет, вы можете начинать откачивать ее вращательную энергию, как мы описывали выше^[43]. В самом деле, природа уже нашла способ, как делать это все наиболее эффективно, пропуская радиацию от аккреционного диска через процедуру с магнитным полем, получившую название “процесс Блэнфорда – Знаека”. И тогда, вполне может быть, удастся получить небольшую добавку к нашим 42 %, если использовать магнитные поля и некоторые другие ингредиенты.

Сфалероны

Есть еще один известный процесс превращения вещества в энергию без использования черных дыр. Его называют сфалеронным. Этот процесс может разламывать кварки, превращая их в лептоны – в электроны, нейтрино, в их более тяжелых кузенов, мюоны, в тау-частицы или в античастицы всех перечисленных^[81]. Как показано на рис. 6.5, стандартная модель в физике элементарных частиц предсказывает, что кварки при подходящих спинах и ароматах могут сходиться вместе и превращаться в лептоны, в промежуточном состоянии побыв недолго сфалероном. Так как в начале процесса масса больше, чем в конце, вся разница должна превратиться в энергию, которая, согласно Эйнштейну, составит ровно $E = mc^2$.

Будущая разумная жизнь, может быть, окажется в состоянии построить то, что я называл *сфалерайзером*, – генератор энергии, чем-то похожий на дизельный двигатель на стероидах. Традиционный дизель сжимает смесь воздуха с дизельным топливом до такого состояния, когда происходит самовоспламенение смеси, и тогда она быстро расширяется, совершая полезную работу – например, толкая поршень. Углекислый газ

вместе с другими продуктами сгорания весит примерно на 0,00000005 % меньше, чем смесь, оказавшаяся под поршнем перед взрывом, эта разница в массе и превращается в тепловую энергию, которая движет машину. Сфалерайзер сжимает обыкновенное вещество до температур в пару квадриллионов градусов, а потом позволяет ему расширяться и остывать, ибо сфалероны уже сделали свое дело^[44]. Мы уже знаем, к чему приводят такие опыты, потому что наша молодая Вселенная проделала его за нас 13,8 миллиардов лет назад, когда была такой же горячей: почти все 100 % вещества превращаются в энергию, лишь меньше одной миллиардной его доли остается чем-то похожим на то, из чего делается обычное вещество, – то есть кварками и электронами. Так что это примерно то же самое, что и дизель, с той только разницей, что сфалерайзер в миллиард раз эффективнее. Еще одно достоинство моего мотора в том, что не надо переживать по поводу топлива: для его работы сгодится все, лишь бы там были кварки, то есть просто любое вещество.

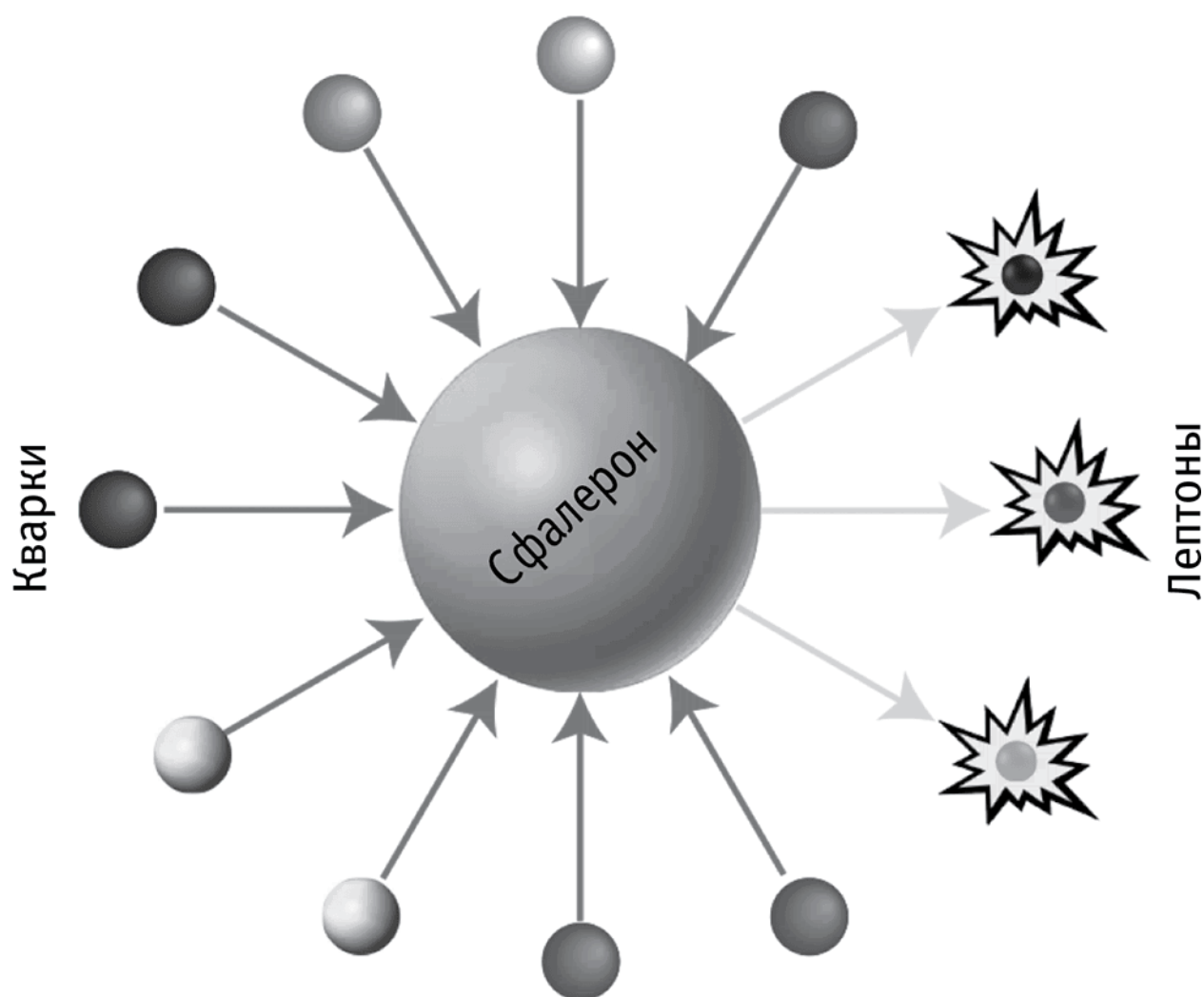


Рис. 6.5

Согласно стандартной модели в теории элементарных частиц, девять кварков с правильно подобранными ароматами, сталкиваясь, превращаются в три лептона. Промежуточное состояние получило название сфалерона. Суммарная масса кварков (в нее входит также энергия связи скрепляющих их глюонов) значительно превосходит массу трех вылетающих лептонов. Избыток выделяется в виде энергии, показанной на рисунке вспышками.

Из-за этого высокотемпературного процесса наша малютка Вселенная произвела в триллион раз больше излучения (фотонов и нейтрино), чем вещества (кварков

и электронов, которые потом образовали атомы). На протяжении последовавших 13,8 миллиардов лет шла великая сегрегация – она привела к тому, что атомы собрались в галактиках, звездах и планетах, в то время как большая часть фотонов осталась в межгалактическом пространстве в виде микроволнового фонового излучения, которым мы теперь пользуемся, чтобы делать фотографии малютки – ранней Вселенной. Любая развитая форма жизни, поселившаяся в галактиках или где-то в других местах скопления вещества, может поэтому превратить почти все оказавшееся поблизости вещество обратно в энергию, перезагрузив изначальную пропорцию и приведя ее к тому крошечному значению, какое было в ранней Вселенной, на короткое время вернувшись в то горячее и плотное состояние внутри сфалерайзера.

Чтобы понять, насколько эффективен сфалерайзер, надо выяснить несколько практических деталей. Например, насколько он должен быть велик, чтобы предотвратить утечку значительной части нейтрино и фотонов на стадии сжатия? Мы этого пока сказать не можем, но в любом случае ясно: перспективы по части выработки энергии у будущей жизни значительно лучше, чем те, что позволяют нынешние технологии. Нам пока даже не удастся построить термоядерный реактор – нашей технологии для этого предстоит еще улучшиться на порядок или даже на два.

Создание улучшенных компьютеров

Если, как мы убедились, обед в смысле его физической эффективности в 10 миллиардов раз хуже, чем тот предел, который устанавливают для него законы природы, то что мы можем сказать о наших современных компьютерах: насколько они эффективны? Их эффективность, как мы скоро увидим, еще хуже, чем у нашего обеда.

Мне часто приходится представлять моего друга и коллегу Сета Ллойда, единственного человека в MIT, который, как подозревают, такой же одержимый, как и я. Завершив пионерскую работу по квантовым вычислениям, он принялся писать книгу, в которой доказывал, что вся Вселенная – это квантовый компьютер. Мы нередко заходим вместе выпить пива после работы, и мне пока не удалось набрести на такую тему, на которую он не смог бы сказать чего-нибудь интересного и неожиданного. Например, как я уже упоминал в главе 2, у него нашлось много чего сказать по поводу окончательных пределов вычислимости. В одной, ставшей знаменитой, статье 2000 года он показал, что скорость вычислений ограничивается энергией: выполнение элементарной логической операции за время T требует средних затрат энергии, равных $E = 4h/T$, где h – это фундаментальная физическая постоянная, носящая имя Макса Планка. Это означает, что килограммовый компьютер в лучшем случае может выполнять 5×10^{50} операций в секунду – на целых 36 порядков больше, чем способен выполнить тот компьютер, на котором я сейчас пишу эти строки. Мы доберемся до таких значений за пару веков, если производительность компьютеров будет продолжать удваиваться каждые два года, как мы показали в главе 2. Кроме того, у него получалось, что компьютер в 1 кг весом может хранить максимум 10^{31} бит, что также в миллиард миллиардов раз лучше моего компьютера.

Сет первым согласится, что достичь этого предела будет нелегкой задачей даже для сверхразумной формы жизни: память этого “окончательного” килограммового компьютера будет напоминать взрыв водородной бомбы или небольшой кусочек Большого взрыва. Однако Сет оптимистично полагает, что практические пределы не очень далеки от этих “окончательных”. В самом деле, существующие прототипы квантовых компьютеров достигли в своей миниатюризации уровня хранения в один бит на атом, что, при пропорциональном скейлинге, соответствует 10^{25} бит на килограмм, – а это в триллион раз больше, чем у моего компьютера. Более того, если для коммуникации между атомами использовать электромагнитное излучение, то быстродействие возрастет

до 5×10^{40} операций в секунду, что превышает быстроедействие CPU моего компьютера на 31 порядок.

Подводя итог, скажем: вычислительные горизонты будущей жизни просто головокружительны: при сравнении порядков окончательный однокилограммовый компьютер настолько же лучше современного суперкомпьютера, насколько последний лучше мигающего “поворотника” на автомобиле, способного хранить только один бит информации и переключаться между двумя своими состояниями примерно за секунду.

Другие ресурсы

С позиций физики все, что только может стремиться создать будущая жизнь – от обитаемых зон и машин до новых форм жизни, – всего лишь по-особому расположенные элементарные частицы. Как голубой кит представляет собой переупорядоченные клетки криля, а криль – переупорядоченные клетки планктона, так и вся наша Солнечная система – это водород, переупорядоченный за 13,8 миллиардов лет космической эволюции: гравитация помещает водород в звезды, где из него komponуются более тяжелые атомы, после чего гравитация переупорядочивает эти атомы в нашу планету, а там химические и биологические процессы переупорядочивают их в живое вещество.

Будущая жизнь, достигшая своего технологического предела, сможет производить такие переупорядочения частиц гораздо быстрее и эффективнее; сначала с помощью своих вычислительных мощностей она будет находить наиболее эффективные методы, а потом будет следовать найденным методам, полагаясь на доступные источники энергии, для достижения нужного упорядочения частиц. Мы видели, как вещество можно превратить в компьютер и как его можно превратить в энергию, – так что это единственный требующийся фундаментальный ресурс^[45]. Как только будущая жизнь доберется до физических пределов и упрется в них, у нее останется единственный ресурс для роста: больше материи. И единственный путь для этого – расширяться во Вселенной. В глубь пространства!

Как добывать ресурсы с помощью космической экспансии

Насколько же велико наше возможное космическое обеспечение? Уточним: каков верхний предел, за которым физические законы окончательно закрывают для нас возможность использования космической материи? Наше космическое обеспечение, разумеется, умопомрачительно велико, – но насколько велико в точности? В табл. 6.2 приведены некоторые ключевые данные. Наша планета мертва на 99,999999 % – в том смысле, что такова доля ее вещества, не входящего в состав биосферы и не делающего ничего полезного для жизни, кроме поддержания гравитационного и магнитного полей. Это соотношение открывает нам возможность однажды воспользоваться для поддержания жизни в сто миллионов раз большим количеством вещества. Если нам каким-то образом удастся оптимизировать использование вещества у себя в Солнечной системе (включая само Солнце), то мы улучшим это соотношение еще в миллион раз. Расселившись по Галактике, мы увеличим свои ресурсы еще в триллион раз.

Насколько далеко мы можем зайти?

Возможно, вы думаете, что мы можем наращивать свои ресурсы неограниченно, употребляя для этого одну галактику за другой, если только будем достаточно терпеливы. Но современная космология не предлагает нам такой возможности! Да, сама по себе Вселенная, может быть, и бесконечна, и содержит бесконечно много галактик, звезд и планет – в самом деле, именно это и предполагает простейшая версия *теории инфляции*, наиболее популярной современной научной парадигмы, в которой наш Большой взрыв 13,8 миллиардов лет назад находит свое логичное объяснение. Однако даже если количество галактик и неограниченно, мы, по-видимому, можем добраться только до конечного их числа: мы можем видеть только около 200 миллиардов галактик, а расселиться – ну никак не больше чем в десяти миллиардах.

Таблица 6.2.

Приблизительное количество вещества, на использование которого будущая жизнь может надеяться.

Область	Количество частиц
Наша биосфера	10 ⁴³
Наша планета	10 ⁵¹
Наша Солнечная система	10 ⁵⁷
Наша галактика	10 ⁶⁹
Доступная нам область нашей Вселенной при перемещениях со скоростью в половину скорости света	10 ⁷⁵
Доступная нам область нашей Вселенной при перемещениях со скоростью света	10 ⁷⁶
Наша Вселенная	10 ⁷⁸

Нас ограничивает скорость света: один световой год (то есть около десяти триллионов километров) за год. На рис. 6.6 показана та часть пространства, откуда свет смог дойти до

нас за прошедшие 13,8 миллиардов лет со времени нашего Большого взрыва, – эта сферическая область известна под названием “наша наблюдаемая Вселенная”, или просто “наша Вселенная”. Даже если вообще пространство бесконечно, наша Вселенная – конечна, насчитывая “всего” 10^{78} атомов. Более того, 98 % этой Вселенной открыты нам по принципу “смотри, но не трогай”, в том смысле, что, видя их, мы не доберемся туда к ним, даже потратив вечность на путешествие со скоростью света. Почему это так? В конце концов, тот факт, что мы не видим остальных частей Вселенной, есть следствие ее ограниченного возраста – она не бесконечно стара, и свету просто не хватило времени, чтобы добраться до нас! Так разве же мы не сможем достичь как угодно удаленной галактики, если нам просто не надо будет думать о том, сколько времени продлится наше путешествие?

Первое препятствие – наша Вселенная постоянно расширяется, из-за чего расселение по далеким галактикам будет напоминать игру в салки. Второе препятствие состоит в том, что это расширение постоянно ускоряется – в этом повинна загадочная темная энергия, на долю которой приходится 70 % всей материи во Вселенной. Чтобы понять, к каким проблемам это может приводить, представьте себе, что вы выходите на железнодорожную платформу и видите, как ваш поезд медленно набирает ход, но дверь нужного вам вагона пока еще заманчиво открыта. Если вы достаточно проворны и безрассудны, сможете ли вы его догнать? Поскольку поезд, очевидно, может двигаться значительно быстрее, чем вы можете бежать, то ответ на вопрос определяется тем, насколько далеко нужная вам дверь была в момент вашего появления на платформе. Если только дистанция между вами превышала некоторое критическое значение, догнать свой вагон вы никогда не сможете. Подобную ситуацию мы встречаем, пытаясь добраться до далекой галактики, удаляющейся от нас с ускорением: даже путешествуя со скоростью света, мы не сможем догнать те галактики, которые сейчас удалены от нас более чем на 17 миллиардов световых лет, – а это более 98 % галактик в нашей Вселенной.

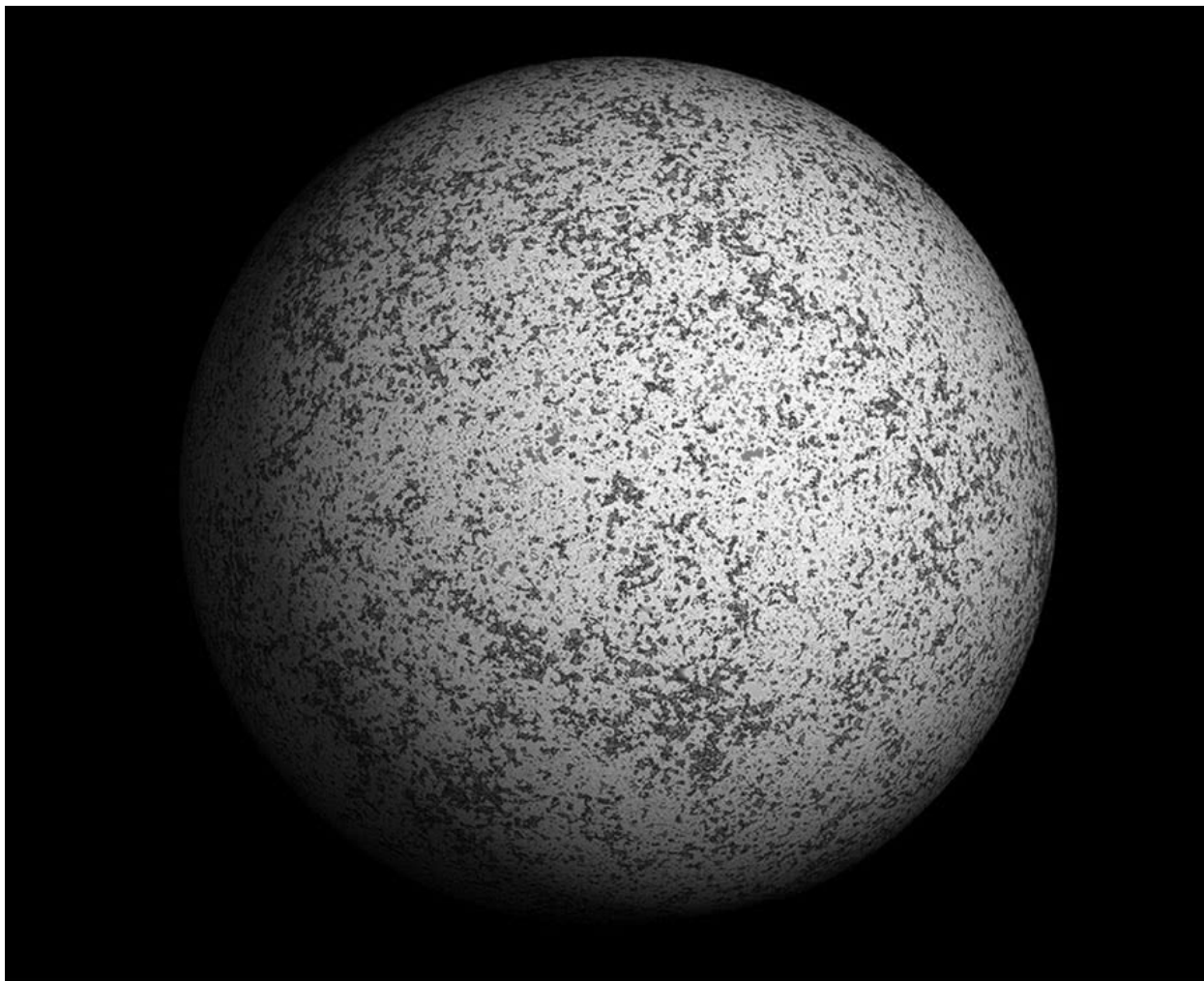
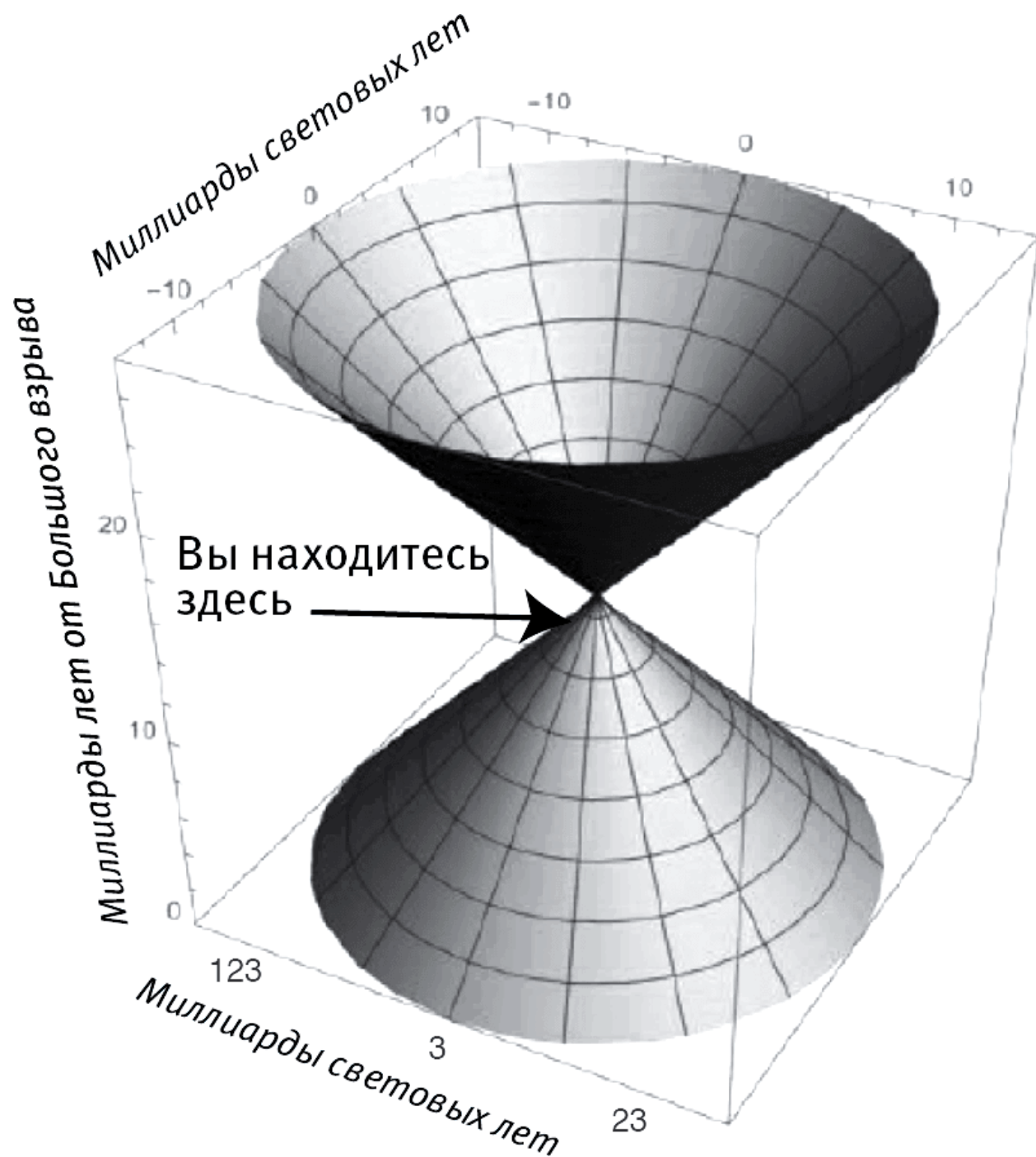
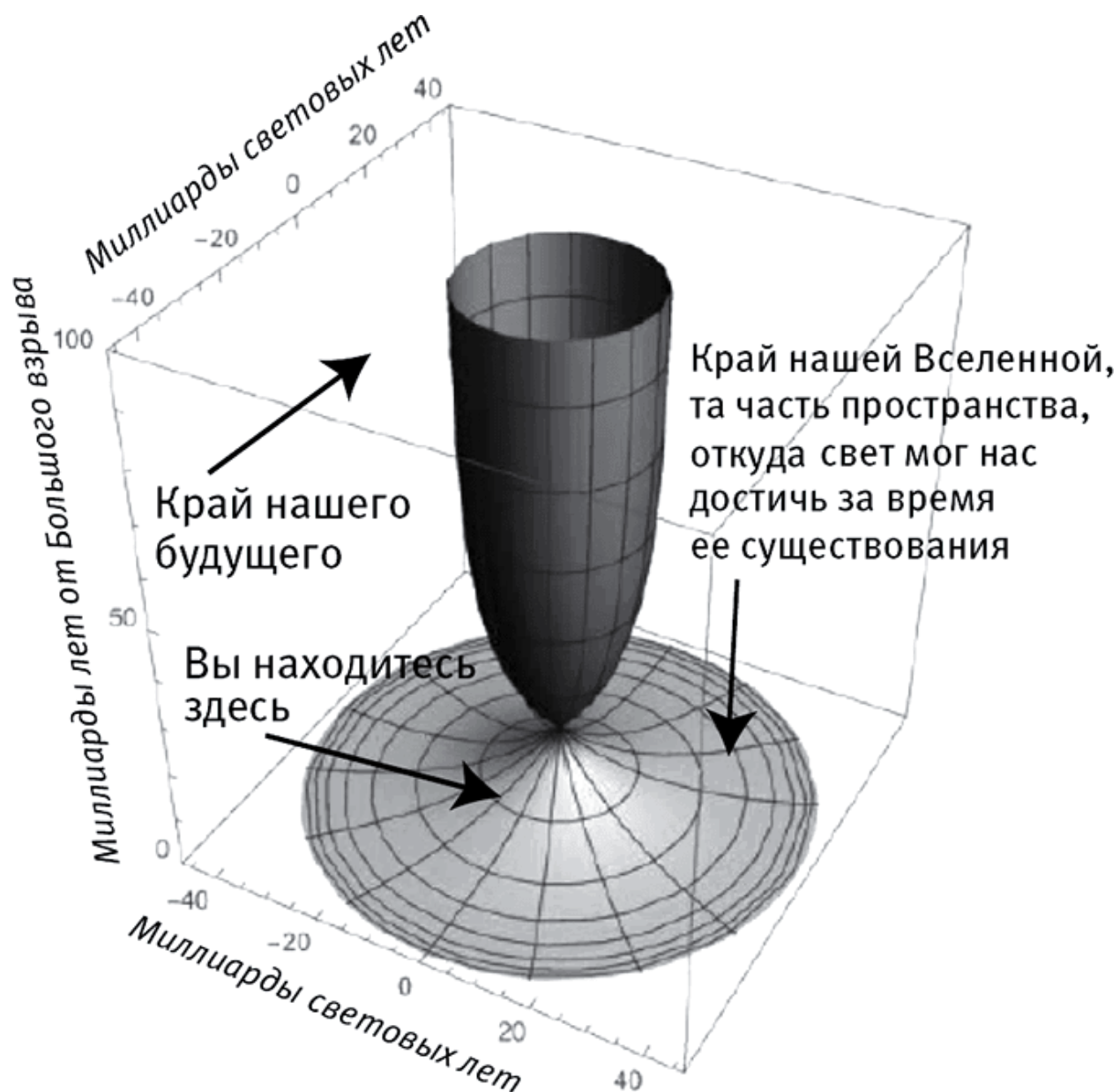


Рис. 6.6

Наша Вселенная, то есть та часть пространства, ограниченная сферой, откуда свет мог добраться до нас за 13,8 миллиардов лет после Большого взрыва. Изображение, полученное с помощью орбитальной обсерватории Planck, показывает, какой она была в младенчестве, в возрасте всего 400 тысяч лет: тогда она состояла из плазмы, примерно такой же горячей, как поверхность Солнца. Пространство, по-видимому, простирается и за этими пределами, а новое вещество каждый год появляется в поле зрения.





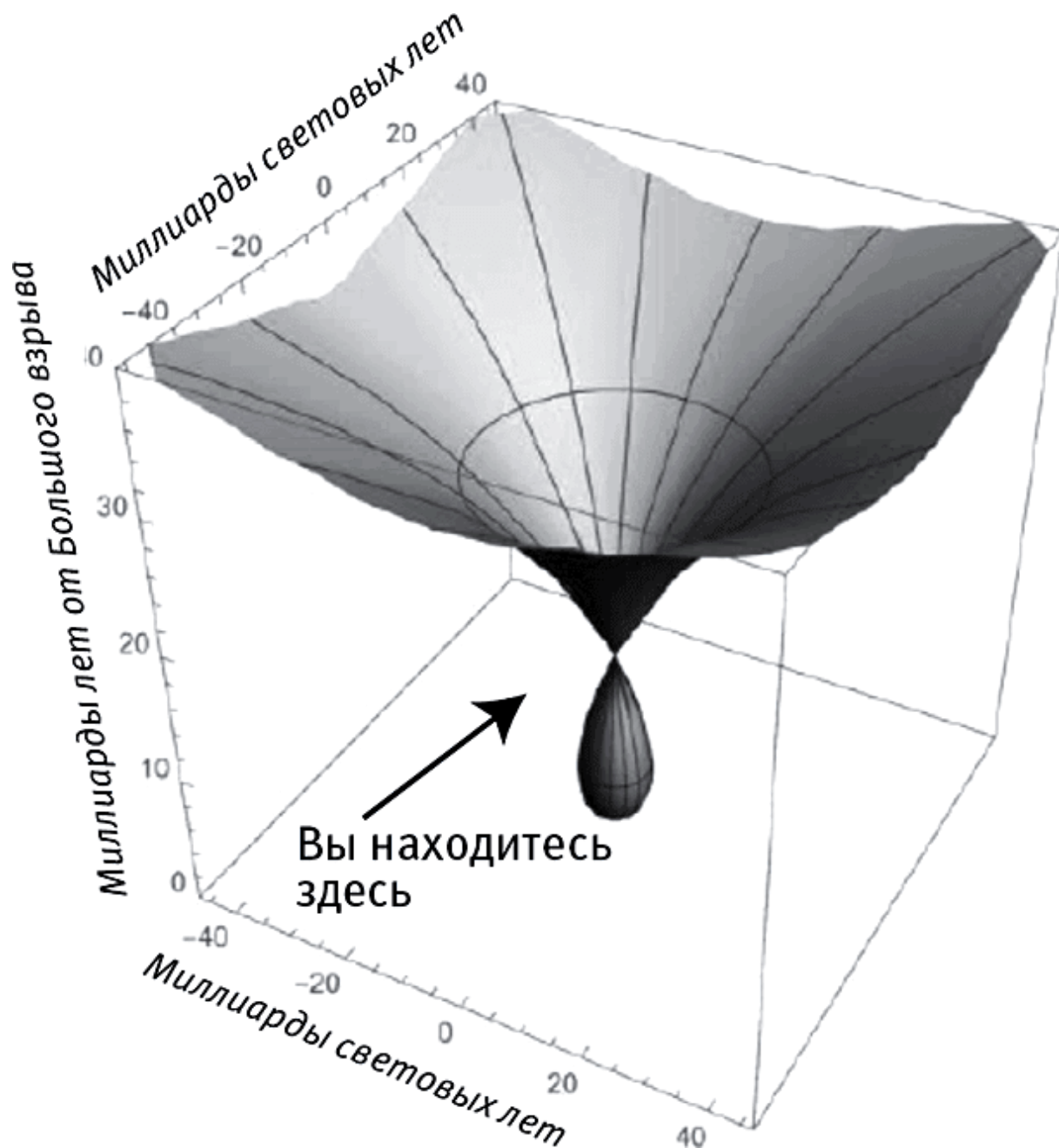


Рис. 6.7

На пространственно-временной диаграмме всякое событие представляется точкой, вертикальные и горизонтальные координаты которой показывают, где и когда соответственно это событие происходит. Если пространство не расширяется (слева вверху), то две половинки конуса ограничивают ту часть пространства-времени, которая на нас на Земле (в вершине конуса) могла оказать определяющее влияние (нижняя половина), и ту, на которую мы можем повлиять сами (верхняя половина), потому что всякое взаимодействие может передаваться со скоростью не больше скорости света. Все становится намного интереснее, когда пространство расширяется (слева внизу и на этой странице вверху). В соответствии со стандартной космологической моделью перемещаться и даже видеть что-либо мы можем только внутри ограниченной области пространства, даже если само пространство и не ограничено. Слева внизу, где изображено нечто, напоминающее бокал для шампанского, мы пользуемся такими координатами, которые компенсируют расширение пространства, и мировые линии далеких галактик вертикальны. В нашу выделенную точку в 13,8 миллиардах световых лет от Большого взрыва свет мог прийти лишь из области, накрытой основанием бокала, и, даже двигаясь со скоростью света, мы никогда не сможем выбраться за пределы

верхней части бокала, содержащей около 10 миллиардов галактик. На правом рисунке, где изображено нечто, напоминающее каплю под чашкой цветка, мы пользуемся обычными координатами и видим пространство расширяющимся. Из-за этого основание бокала деформируется, потому что все у его кромок, видимых нам сейчас, в начале времен и в самом деле было очень близко друг к другу.

Но погодите: разве специальная теория относительности Эйнштейна не запрещает чему бы то ни было двигаться быстрее света? Как же тогда могут галактики обгонять что-то, движущееся со скоростью света? Ответ заключается в том, что специальной теории относительности пришла на смену общая теория относительности, правила которой несколько мягче: ничто не может в своем *движении в пространстве* превысить скорость света, но само пространство может расширяться так быстро, как ему захочется. Эйнштейн также дал нам дивный способ визуализировать этот предел скорости, представляя время как четвертое измерение в пространстве (см. рис. 6.7, где я изобразил четырехмерное пространство трехмерным, отбросив одно из трех пространственных измерений). Если бы пространство не расширялось, световые лучи можно было бы представить как прямые линии, проведенные под углом в 45 градусов, так что все, открывающееся нашему взгляду здесь и сейчас, лежит внутри конусов. В то время как наш световой конус прошлого обрезается Большим взрывом на отметке в 13,8 миллиарда световых лет, наш конус будущего расширяется неограниченно, открывая перед нами бесконечные просторы нашего будущего материального обеспечения. Совсем другую картину мы находим на рисунке 6.7 слева внизу, где изображена расширяющаяся Вселенная с темной энергией (то есть, похоже, та самая Вселенная, где мы и живем), которая деформирует конусы во что-то напоминающее бокал для шампанского, навеки ограничивая долю доступных нам галактик примерно 10 миллиардами.

Если такие ограничения вызывают у вас приступы космической клаустрофобии, могу вас поздравить: у вас еще есть спасительная лазейка! В моих вычислениях предполагалось, что темная энергия не меняется со временем, что в общем-то соответствует имеющимся измерениям. Но, тем не менее, у нас пока еще нет никаких намеков, чем бы таким могла оказаться темная энергия на самом деле, и поэтому пока остается призрачная надежда, что она со временем может рассосаться (как должна была поступить подобная темной энергии субстанция, предположение о существовании которой нам необходимо, чтобы объяснить инфляцию), и когда это случится, ускорение должно будет уступить место замедлению, потенциально давая возможность будущим формам жизни расселяться по галактикам, не ограничивая себя.

Как быстро мы сможем расти?

Выше мы исследовали вопрос, сколько галактик может быть заселено, если мы будем распространяться во все стороны со скоростью света. Общая теория относительности говорит, что перемещаться в пространстве со скоростью света ракеты не могут, это потребовало бы бесконечно большой энергии. Но как же быстро смогут наши ракеты лететь на практике?^[46]

Ракета "Атлас V", выводившая за пределы земной атмосферы автоматическую межпланетную станцию NASA New Horizons, когда та начинала свой путь к Плутону, достигла к моменту выключения двигателей гелиоцентрической скорости в 100 000 миль в час (это соответствует 45 километрам в секунду), а запланированный на лето 2018 года запуск солнечного зонда Solar Probe Plus подразумевает даже большие скорости: приближаясь к Солнцу, зонд должен будет разогнаться вчетверо быстрее, но даже эта скорость не превысит жалкой 0,1 % от скорости света. Квест ради обретения все лучших и быстрых ракет длился все прошлое столетие, и ему посвящена обширнейшая

и увлекательнейшая литература. Почему же так трудно разогнаться хоть чуточку больше? Две основные проблемы современных ракет заключаются, во-первых, в том, что большую часть своего топлива им приходится тратить на то, чтобы разогнать до нужной скорости все остающееся топливо, а во-вторых, в том, что это топливо безнадежно неэффективно: как показано в табл. 6.1, часть массы, превращающейся в энергию, для бензина не намного больше 0,00000005 %. Наиболее очевидное направление улучшений – это переход на более эффективное топливо. Например, Фримен Дайсон с коллегами работал над проектом NASA “Орион”, цель которого заключалась в том, чтобы взорвать 300 тысяч атомных бомб за десять дней и разогнать до 3 % скорости света космический корабль – достаточно большой и вместительный: в ходе путешествия продолжительностью в век находящиеся на нем люди могли бы достичь другой звездной системы^[82]. Исследовалась и возможность использовать в качестве топлива антиматерию: в сочетании с обычной материей она обеспечивает близкую к 100 % эффективность.

Еще одна популярная идея заключается в том, чтобы построить ракету, которой не надо нести топливо на себе. Например, межзвездное пространство – далеко не вакуум, в нем много случайным образом рассеянных ионов водорода (то есть, попросту говоря, протонов, которые представляют собой атом водорода, лишившийся своего единственного электрона). Это подтолкнуло в 1960 году физика Роберта Буссарда к мысли о создании двигателя, получившего название “прямоточного межзвездного двигателя Буссарда”: он должен был собирать по пути эти рассеянные ионы и использовать их в качестве топлива в бортовом термоядерном реакторе. Хотя в последние годы появились определенные сомнения в том, что эту идею удастся осуществить на практике, есть другая идея, которая может быть реализована высокотехнологичной космической цивилизацией, – лазерный серфинг.

Рис. 6.8 иллюстрирует проект лазерного паруса, с которым в 1984 году выступил Роберт Форвард, тот же самый физик, который предложил статиты, обсуждавшиеся нами в связи со сферой Дайсона. Так же как молекулы воздуха ударяются о поверхность паруса парусного судна, так и частицы света (фотоны), ударяясь о поверхность зеркала, толкают его вперед. Направив луч мощного питаемого Солнцем лазера на поверхность ультралегкого паруса, установленного на космическом корабле, мы можем использовать энергию нашего Солнца для ускорения ракеты до больших скоростей. Но как потом остановиться? Этот вопрос ускользал от моего внимания до тех пор, пока я не прочитал блестящую статью Форварда: как показано на рис. 6.8, внешний круг паруса отделяется и перемещается перед кораблем, отражая лазерный луч назад и замедляя и сам корабль, и его меньший парус^[83]. Форвард провел вычисления, показавшие, что таким образом люди могли бы слетать к ближайшей к Солнцу звездной системе α Центавра всего лишь за сорок лет. А уж после этого можно строить настоящую большую лазерную систему, позволяющую прыгать от звезды к звезде по всему Млечному пути.

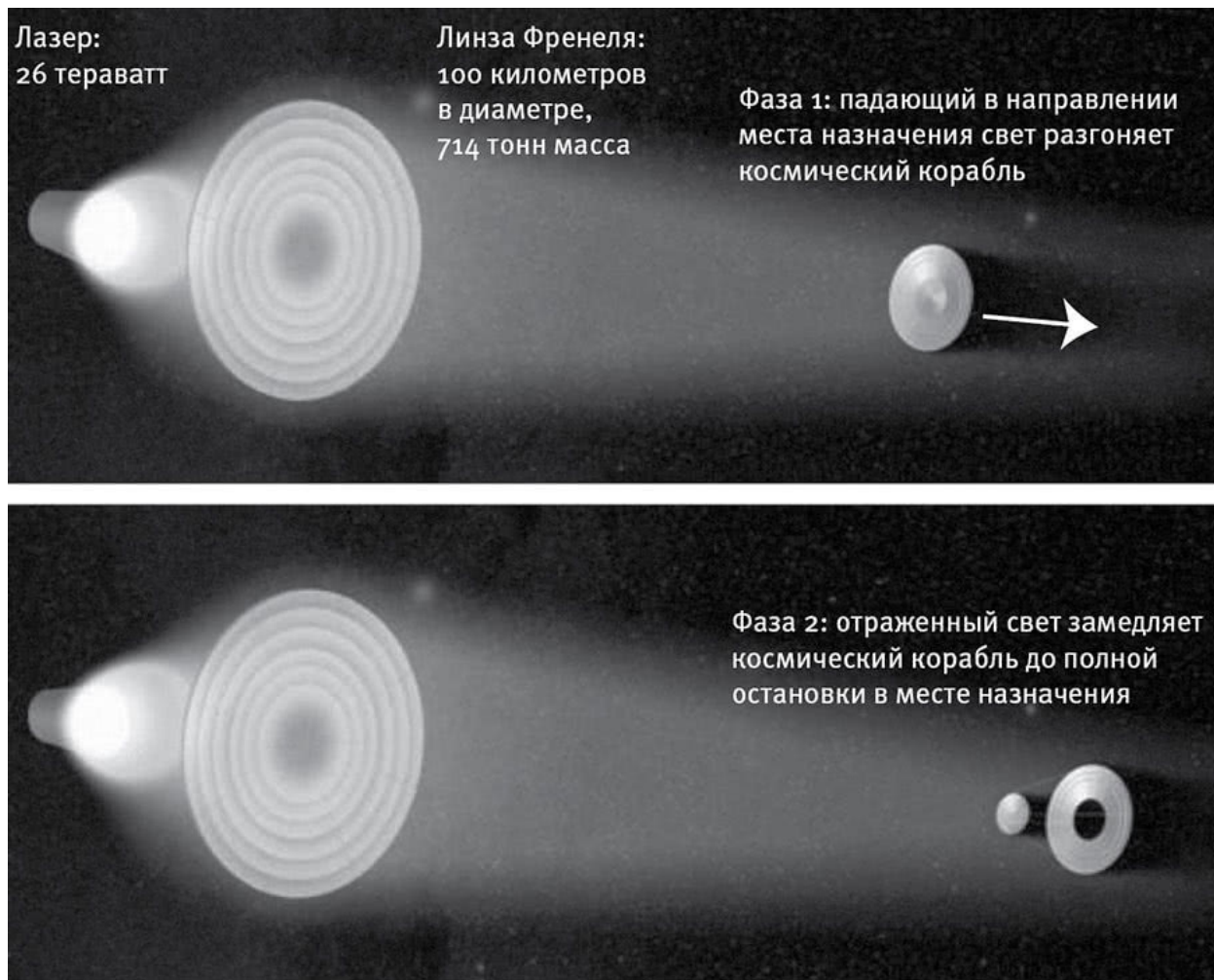


Рис. 6.8

Проект космического корабля с лазерным парусом, разработанный Робертом Форвардом для полета к звездной системе α Центавра в 4 световых годах от нас. Сначала мощный лазер в нашей Солнечной системе разгоняет корабль, создавая радиационное давление на лазерном парусе. Для торможения во время прибытия к месту назначения внешняя часть паруса открепляется и начинает отражать свет обратно в направлении корабля.

Но зачем останавливаться на этом? В 1964 году советский астроном Николай Кардашев предложил градуировать цивилизации по количеству энергии, которую они могут использовать по своему произволу. Освоению энергии планеты, звезды (с помощью сферы Дайсона, скажем) и галактики соответствуют типы I, II и III по шкале Кардашева. Его последователи предположили, что типу IV должна соответствовать цивилизация, освоившая энергию всей Вселенной. С тех пор мы кое-что еще узнали, и в этом есть своя хорошая и своя плохая сторона для амбициозных форм жизни. Плохая новость состоит в открытии темной энергии, которая, как мы видели выше, сужает нам зону достижимости. Хорошая новость – это радикальный прогресс искусственного интеллекта. Даже такие оптимистические визионеры, как Карл Саган, рассматривали перспективу достижения человеческими существами других галактик как практически безнадежную, принимая во внимание нашу человеческую склонность умирать в течение уже первого столетия путешествия, которому, вероятно, суждено продлиться не один миллион лет, даже если перемещаться со скоростью близкой скорости света. Не желая отказываться от своей затеи, Саган предлагал замораживать астронавтов для продления им жизни, замедлять их старение околосветовыми скоростями, отправлять их большими

сообществами, рассчитывая, что путешествие продлится на протяжении жизней десятков тысяч поколений – это больше, чем их сменилось за время земной жизни людей.

Возможности сверхразума кардинально изменили всю картину, сделав ее гораздо более многообещающей для тех, кого одолевает межгалактический вандерлюст. Как только необходимость перевозить с собой сложную и тяжелую систему жизнеобеспечения людей отпадает, технология заселения межгалактического пространства становится довольно прозрачной. Лазерные парусники Форварда сильно удешевляются, когда им надо всего лишь перевезти “семенной зонд” – робота, способного расположиться на поверхности астероида или планеты и запустить там цивилизацию “с нуля”. Ему не надо для этого даже везти с собой инструкции: все, что он должен сделать, – это возвести достаточно большую приемную антенну, которая станет со скоростью света принимать более подробные указания и схемы от материнской цивилизации. Как только все это будет проделано, он сможет использовать новый, только что построенный лазер для отправки нового семенного зонда, переносящего жизнь к следующей звездной системе следующей галактики. Даже бескрайние просторы между галактиками заполнены значительным числом межгалактических звезд (изгнанников, когда-то отторгнутых своими звездными системами), которые можно использовать для промежуточных станций, развивая таким образом стратегию межгалактического лазерного слалома – скачков от одного острова к другому.

Как только новая солнечная система или галактика будет заселена сверхразумным искусственным интеллектом, доставить туда людей не составит большого труда – если, конечно, люди преуспеют в том, чтобы искусственный интеллект озабочился такой проблемой. Вся необходимая информация о людях может быть передана со скоростью света, после чего искусственный интеллект соберет нужных людей из имеющихся кварков и электронов. Технологию можно использовать самую примитивную, при которой транслируется два гигабайта информации, содержащейся в ДНК, а младенца из инкубатора берет под свою опеку местный искусственный интеллект, или же искусственный интеллект собирает из кварков и электронов уже взрослого человека, со всеми его воспоминаниями, сканированными на Земле у оригинала и залитыми в копии на месте.

Все это означает, что после интеллектуального взрыва ключевым вопросом становится не то, *может ли* человечество расселяться в межгалактическом пространстве, а то, *как быстро* оно может это делать. Так как все обсуждаемые нами идеи исходят от людей, их можно рассматривать как основание для нижней оценки скорости расселения; амбициозная сверхразумная жизнь будет проделывать это все значительно лучше, к тому же у нее будет сильный стимул раздвигать пределы: в состязании с темной энергией за выигрыш во времени она будет получать за каждый процент в увеличении скорости заселения галактик 3 % выигрыша в их количестве.

Например, если к следующей звездной системе, расположенной на расстоянии 10 световых лет, надо лететь с использованием лазерного паруса 20 лет и после этого еще 10 лет обживать систему, строить на ней новый лазер и новый семенной зонд, то заселенная область будет расти в среднем со скоростью равной трети скорости света. Американский физик Джей Олсон в 2014 году в своем подробном и изящном исследовании цивилизации, распространяющейся в космосе, рассматривает высокотехнологическую альтернативу скачкам от острова к острову, в которой используется два вида зондов – семенные зонды и *экспандеры*^[84]. Семенные зонды при приближении к месту назначения замедляются, спускаются на поверхность, приступают к осеменению. Экспандеры же не останавливаются никогда: они собирают на лету встреченные частицы, используя, возможно, улучшенный прямоточный метод Буссарда, чтобы из собранного вещества, за вычетом той его части, которая ушла на поддержание движения в качестве топлива, создавать копии самого себя. Самовоспроизводящийся флот экспандеров будет перемещаться с небольшим ускорением, чтобы сохранять

постоянную скорость (скажем, в половину скорости света) по отношению к ближайшим галактикам, и будет размножаться с достаточной скоростью, чтобы при сохранении сферической формы флот сохранял и постоянное число экспандеров на единицу площади этой сферы.

Есть и еще не менее важный, хотя и занявший последнее место в нашем списке, – метод девы Марии, позволяющий развить скорость расширения большую, чем у любого другого из рассмотренных выше методов: в нем используется то, что Ганс Моравец называл “космическим спамом” и о чем шла речь в главе 4. Передавая по радио сообщения, привлекающие внимание неопытных цивилизаций на ранних этапах космической эволюции и склоняющие их к постройке сверхразумных машин, которые производят цивилизационный взлом, цивилизация получает возможность распространять себя практически со скоростью света, или, более точно, со скоростью, с которой распространяются в космосе соблазнительные песни их сирен. Так как такой путь может оказаться *единственным* для достижения большинства галактик в световом полуконусе, соответствующем их будущему, у сверхразумной цивилизации нет причин не попробовать им воспользоваться, и поэтому нам надо быть исключительно осторожными по отношению ко всему, что мы получаем из межзвездного пространства. В книге Сагана *Контакт* земляне используют схемы, полученные от пришельцев, для постройки машин, действия которых не понимают, – ох, не советовал бы я так делать!

Подводя итог, я должен констатировать, что большинство ученых и писателей-фантастов при рассмотрении возможности космического расселения оказались чрезмерно пессимистичными, на мой взгляд, потому, что проигнорировали участие сверхразума: ограничивая свое внимание только путешественниками-людьми, они переоценили сложности межгалактического полета, а учитывая только технологии, созданные людьми, они переоценили время, которое требуется, чтобы достичь рубежей физически возможного.

Как сохранять контакт методами космической инженерии

Если темная энергия будет продолжать расталкивать галактики, заставляя их удаляться друг от друга с ускорением, на что указывают эксперименты последнего времени, то это не слишком благоприятно для будущего жизни. Это означает, что даже если будущая цивилизация сможет расселиться в миллионе галактик, темная энергия за несколько десятков миллиардов лет раздробит эту империю на десятки тысяч областей, сообщение между которыми будет невозможным. Если у будущей жизни не окажется никакого способа противостоять этой фрагментации, то крупнейшие оставшиеся бастионы будут включать несколько тысяч галактик, силы притяжения между которыми достаточно велики, чтобы противостоять расталкивающему действию темной энергии.

Если сверхразумная цивилизация захочет, тем не менее, оставаться единой, у нее появится сильный стимул разворачивать крупномасштабную космическую инженерию. Какое количество вещества она сможет переместить в крупнейший из таких суперкластеров, прежде чем темная энергия навсегда скроет его из виду? Один из методов перемещения звезды на большое расстояние заключается в том, чтобы втолкнуть третью звезду в бинарную систему, где две звезды устойчиво обращаются друг вокруг друга. Точно так же, как в романтических взаимоотношениях появление третьего партнера нарушает равновесие и какая-то из вершин треугольника силой из него выталкивается, – в случае звезд выталкиваемый приобретает значительную скорость. Если одним из трех “партнеров” была черная дыра, то такой взрывоопасной ситуацией можно воспользоваться, чтобы разогнать изгоняемую массу до такой скорости, что она окажется очень далеко от своей родной галактики. К сожалению, такой способ манипулирования треугольниками звезд, черных дыр и галактик не позволяет переместить больше

ничтожной части цивилизационной массы на достаточно большие расстояния, чтобы это позволило справиться с темной энергией.

Но отсюда, естественно, никак не следует, что у сверхразумной жизни не возникнет методов получше, – скажем, таких, какие позволят превратить бо2льшую часть материи удаленных галактик во что-то вроде космического корабля, способного вернуться в окрестности домашнего суперкластера. Если сфалерайзер может быть построен, то, возможно, его удастся использовать для превращения вещества в энергию, которая в виде луча может быть направлена к домашнему кластеру, там снова преобразована в вещество или потреблена как энергия.

Настоящим счастьем оказалась бы возможность создания устойчивых проходимых кротовых нор, которые обеспечили бы как минимум коммуникацию, а то и транспортировку почти без затрат времени между двумя концами норы, независимо от того, сколь далеко они один от другого. Кротовая нора – это перемычка в пространственно-временном континууме, позволяющая перемещаться из точки *A* в точку *B*, минуя разделяющее их пространство. Хотя стабильные кротовые норы допускаются общей теорией относительности Эйнштейна и они уже появлялись в кино – в фильмах *Контакт* и *Интерстеллар*, для их создания требуется довольно странное вещество с отрицательной плотностью, возможность существования которого зависит от некоторых, пока еще недостаточно понятых, эффектов квантовой гравитации. Другими словами, еще может так повернуться, что никакой практической пользы кротовые норы нам не принесут, но если этого не случится, у сверхразумной жизни будет колоссальный стимул научиться их строить. Кротовые норы не только откроют широчайшие возможности для обмена информацией внутри каждой отдельной галактики, но и, изначально привязав удаленные галактики к центральному кластеру, обеспечат общую тенденцию будущей жизни к единству в долгосрочной перспективе, полностью нивелируя блокирующее коммуникацию действие темной энергии. После того как две галактики оказываются связанными одной кротовой норой, они сохраняют эту связь, как бы далеко ни разносило их с течением времени.

Если все же, несмотря на все усилия космической инженерии, будущая цивилизация обнаружит, что некоторым ее частям суждено навсегда покинуть область, где возможен информационный обмен, с ними можно просто попрощаться, пожелав им всяческих удач. Однако если у нее окажутся амбициозные вычислительные запросы, она сможет вместо этого прибегнуть к своеобразному варианту подсечно-огневой стратегии: она превратит отрывающуюся галактику, вместе со всем ее веществом и всей ее энергией, в массивный компьютер в форсированной моде, в надежде, что прежде чем ее обуглившиеся останки скроются за горизонтом событий, долго не находившиеся ответы будут транслированы в направлении материнского кластера. Такая “подсечно-огневая” стратегия особенно хороша для самых удаленных областей Вселенной, достичь которых можно было лишь с помощью космического спама, на далеких окраинах предсуществующих обитаемых зон. В своей материнской области цивилизация должна, напротив, стремиться к максимальной стабильности и эффективности, сохраняя их так долго, как только возможно.

Как долго это может длиться?

Долголетие – это нечто такое, к чему стремится большинство людей, организаций и наций. И если некая амбициозная будущая цивилизация создаст сверхразум и будет стремиться к долголетию, чего она сможет добиться?

Первый подробный анализ нашего возможного далекого будущего дал не кто иной, как сам Фримен Дайсон. Табл. 6.3 резюмирует его находки. Без вмешательства разума и планетные системы, и галактики постепенно разрушаются, как и все прочее, оставляя лишь холодное, мертвое, пустое пространство с навеки замирающим потоком излучения.

Но Фримен Дайсон заканчивает свои рассуждения на оптимистической ноте: “У нас есть серьезные научные основания думать, что жизнь и разум смогут придать этой Вселенной формы, лучше отвечающие и ее, и нашим целям”^[85].

Я думаю, что сверхразум легко сможет разрешить многие из проблем, перечисленных в табл. 6.3, потому что сможет найти способы переупорядочить вещество в кое-что получше, чем планетные системы и галактики. Часто обсуждаемые проблемы из этого ряда, вроде угасания нашего Солнца через несколько миллиардов лет, вряд ли достойны даже упоминания, так как даже низкотехнологичные цивилизация легко могут переместиться к звездам небольших масс, способным обеспечить их существование на следующие 200 миллиардов лет или даже больше. Предположим, что некая сверхразумная цивилизация научилась строить силовые станции, по эффективности превышающие звезды. Они могут даже *препятствовать* образованию новых звезд из соображений сохранения энергии: даже собирая всю излучаемую звездой на протяжении ее жизни энергию с помощью сферы Дайсона (а это всего 0,1 % всей ее энергии), они окажутся не в силах использовать остающиеся 99,9 % ее энергии, пропадающие впустую с гибелью самых массивных звезд. Тяжелая звезда гибнет со взрывом сверхновой, при этом большая часть энергии улетучивается с неуловимыми нейтрино, а очень тяжелые звезды превращаются в черные дыры, где вся масса оказывается надолго похороненной, по капельке выбираясь оттуда следующие 10^{67} лет.

Таблица 6.3.

Оценки для отдаленного будущего – все, кроме 2-й и 7-й, – сделаны Фрименом Дайсоном. Он проводил свои вычисления еще до открытия темной энергии, которая добавляет новые сценарии “космокалипсиса” через 10^{10} – 10^{11} лет. Протоны, возможно, совершенно стабильны, но если и нет, эксперименты показывают, что на распад половины из них уйдет никак не меньше 10^{34} лет.

Что	Когда
Текущий возраст нашей Вселенной	10^{10} лет
Темная энергия вытолкнет большинство галактик из области достижимости	10^{11} лет
Догорит последняя звезда	10^{14} лет
Планеты покинут свои орбиты около звезд	10^{15} лет
Звезды покинут свои галактики	10^{19} лет
Гравитационное излучение разрушит орбиты	10^{20} лет
Распадутся протоны (не раньше чем)	$> 10^{34}$ лет
Испарятся черные дыры звездных масс	10^{67} лет
Испарятся супермассивные черные дыры	10^{91} лет
Вся материя распадется до железа	10^{1500} лет
Вся материя свалится в черные дыры, которые затем испарятся	10^{1026} лет

Пока сверхразумная жизнь не испытывает трудностей с энергией/материей, она может поддерживать свои обитаемые зоны в том состоянии, в каком хочет. Возможно,

она даже научится предотвращать распад протонов с помощью квантово-механического эффекта, получившего название “наблюдаемого чайника”^[47], в соответствии с которым регулярное проведение наблюдений уменьшает вероятность распада. Но тут есть одна неприятная закладка: *космокалипсис*, который, вероятно, уничтожит всю нашу Вселенную через 10–100 миллиардов лет. Открытие темной энергии и развитие струнной теории, послужившие поводом к созданию новых космокалипсических сценариев, случились уже после того, как Фримен Дайсон написал свою статью.

Так как же быть с приближающимся концом нашей Вселенной, который может наступить по прошествии миллиардов лет? У меня есть пять главных подозреваемых в организации этого космического апокалипсиса, или космокалипсиса, основные характеристики которых представлены на рис. 6.9: Большой мороз (Big Chill), Большой хруст (Big Crunch), Большой разрыв (Big Rip), Большой шлепок (Big Snap) и Пузыри смерти (Death Bubbles). К настоящему времени наша Вселенная расширяется уже около 14 миллиардов лет. Большой мороз случится, если она будет продолжать так расширяться, пока не превратится в холодное, темное и, в конечном счете, мертвое место, и такой вариант рассматривался как наиболее вероятный в то время, когда Дайсон писал свою статью. Я думал об этом как о варианте, предложенном Томасом Элиотом: “Так вот и кончится мир / Только не взрывом а вздрогом”^[48]. Если вы, вслед за Робертом Фростом, предпочтете лучше сгореть в огне, чем замерзнуть, то скрестите пальцы за Большой хруст: это значит, что расширение Вселенной должно смениться на противоположный процесс, и все закончится сжатием, подобным Большому взрыву наоборот. Наконец, Большой разрыв – это Большой мороз для нетерпеливых, когда все галактики, звезды, планеты и даже атомы разрываются в грандиозном финале, наступающем через конечное время. На кого из этих троих вы бы поставили? Все зависит от того, как поведет себя темная энергия, на которую приходится 70 % массы Вселенной, при дальнейшем расширении пространства. Случится ли мороз, хруст или разрыв, определяется тем, будет ли темная энергия оставаться неизменной, какова она сейчас, постепенно растворится, обретая отрицательную плотность, или будет, напротив, уплотняться, приобретая все большую положительную плотность. Так как о темной энергии пока никому ничего не известно, я расскажу вам только о своих ставках: я бы поставил 40 % на Большой мороз, 9 % на Большой хруст и 1 % на Большой разрыв.



Рис. 6.9

Мы знаем, что наша Вселенная началась с Большого взрыва 14 миллиардов лет назад, она расширялась и остывала, частицы соединялись в атомы, потом в звезды и галактики. Но мы не знаем ее конечной судьбы. Предполагаемые сценарии включают

Большой мороз (вечное расширение), Большой хруст (повторное стягивание в точку), Большой разрыв (разрыв всей материи из-за бесконечно нарастающей скорости расширения), Большой шлепок (в структуре пространства обнаруживаются смертоносные гранулы, дающие о себе знать при слишком большом растяжении) и Пузыри смерти (“замерзающее” пространство с образующимися в нем смертоносными пузырями, расширяющимися со скоростью света).

На что же пойдут остальные 50 % моих денег? Я сохраню их с определением “ничто из перечисленного”, поскольку нам, людям, надо вести себя скромно и признавать наличие фундаментальных явлений, о которых нам пока что ничего не известно. О природе пространства, например. Все три сценария – и мороз, и хруст, и разрыв – предполагают, что пространство может расширяться до бесконечности, оставаясь устойчивым. Мы когда-то думали о пространстве как о скучной и статичной сцене, на которой лишь разворачивается космическая драма. Но Эйнштейн научил нас смотреть на пространство как на ключевого актора: оно может сворачиваться в черную дыру, может идти складочками в гравитационных волнах или растягиваться в расширяющуюся Вселенную. Может быть, оно способно и переходить в другие агрегатные состояния, как вода при замерзании? Тогда выяснится, что в нашей игре есть джокеры, претендующие на свою роль в организации космокалипсиса. Несущие смерть пузыри другой фазы, в которую переходит пространство при замерзании, если они возможны, будут, вероятно, расширяться со скоростью света, как и сфера космического спама от предельно агрессивной цивилизации.

Кроме того, теория Эйнштейна предполагает, что наше пространство всегда может расширяться, до бесконечности увеличивая свой объем, как в сценариях Большого мороза и Большого разрыва. Это слишком хорошо звучит, чтобы быть правдой, и я подозреваю, ею не является. Резиновая лента хорошо тянется и хорошо выглядит, но если слишком ее растянуть, она лопнет. Почему? Потому что она состоит из атомов, и при сильном растягивании ее гранулярная атомарная структура входит в игру. Может ли такое случиться, что и пространство обладает своей гранулярной структурой, о которой нам просто ничего не известно, потому что эти гранулы слишком малы и мы их не замечаем? Исследования по квантовой гравитации показывают, что разговоры о нашем привычном трехмерном пространстве теряют смысл на расстояниях меньших 10^{-34} метра. Если и в самом деле пространство не может расширяться до бесконечности без риска испытать Большой шлепок, будущие цивилизации могут захотеть переместиться в самую большую нерасширяющуюся область пространства (колоссальный кластер галактик), до какой только смогут добраться.

Сколько мы можем вычислить?

После того как мы оценили, сколь долго будущая жизнь может существовать, давайте теперь обсудим, как долго она должна хотеть существовать. Вам, конечно, кажется, будто она должна хотеть жить вечно или хотя бы так долго, как только возможно, Фримен Дайсон дал численный параметр для этого желания: стоимость вычислений падает, когда вы считаете медленнее, так что в конце концов результат оказывается максимальным при предельном замедлении вычислений. Дайсон даже рассчитал, что при бесконечном расширении и охлаждении Вселенной бесконечное количество вычислений оказывается возможным.

Медленно еще не обязательно означает скучно: если будущая жизнь существует в симулированном мире, субъективно переживаемый поток времени не обязательно должен быть связан с замерзающим ходом времени, обеспечивающим работу симуляции во внешнем мире, так как перспектива бесконечного числа вычислений может быть

переведена в субъективное бессмертие для симулированных форм жизни. Космолог Фрэнк Типлер, основываясь на этой идее, провел свои рассуждения, чтобы показать: субъективное бессмертие достижимо даже в последние мгновения существования Вселенной перед Большим хрустом благодаря бесконечному ускорению вычислений при быстром росте температуры и плотности.

Так как темная энергия, похоже, похоронит планы и Дайсона, и Типлера, будущий сверхразум может предпочесть относительно быстро выжечь всю доступную энергию – еще до того, как столкнется с серьезными проблемами вроде космических горизонтов и протонного распада. Если окончательная цель заключается в максимальном увеличении количества вычислений, то лучшая стратегия должна проходить где-то между слишком медленным (чтобы избежать проблем, упомянутых выше) и слишком быстрым (чтобы не тратить на вычисления больше энергии, чем необходимо).

Если сложить вместе все, что обсуждалось в этой главе, то нам станет ясно: максимально выгодные силовые станции и компьютеры дадут сверхразумной жизни совершенно немыслимый вычислительный потенциал. Чтобы зарядить ваш тринадцативаттный мозг на сотни лет, достаточно энергии, содержащейся в полумиллиграмме вещества – это меньше, чем одна крупинка сахара. Исследование Сета Ллойда предполагает, что мозг может быть сделан в квадриллион раз эффективнее в отношении потребления энергии, и тогда той же крупинки сахара хватит, чтобы обеспечить работу симуляции всех когда-либо живших людей, и даже в тысячи раз большего их числа. Если все вещество доступной для нас части Вселенной пустить на симуляцию людей, то его хватит на 10^{69} жизней – или на что-либо иное, на что сверхразумный искусственный интеллект сочтет нужным израсходовать это количество энергии. Можно симулировать еще большее число жизней, запуская симуляции с меньшей скоростью^[86]. И наоборот: в своей книге *Superintelligence* Ник Бострём подсчитал, что 10^{58} жизней можно симулировать с менее жесткими требованиями относительно эффективности энергопотребления. Но как ни крути эти числа, они остаются огромными, как и наша ответственность за то, чтобы возможности для процветания будущей жизни не пропали впустую. Как писал об этом Бострём: “Если мы представим все счастье, переживаемое на протяжении одной такой жизни целиком, одной слезинкой радости, то счастье этих душ наполняет и переполняет слезами мировой океан каждую секунду, и так продолжается на протяжении сотен миллиардов миллиардов тысячелетий. Очень важно нам позаботиться о том, чтобы эти слезы действительно были слезами радости”.

Космические иерархии

Скорость света ограничивает не только распространение жизни, но и ее природу, строго сдерживая обмен информацией, сознание и управление. И если большая часть нашего космоса станет живой, на что эта жизнь будет больше всего похожа?

Мысленные иерархии

Случалось ли вам когда-нибудь сделать попытку прихлопнуть муху рукой, но промазать? Главная причина, по которой мухе удается удрать, в ее размере: информации требуется меньше времени, чтобы совершить путешествие между глазами, мозгом и мускулами. Этот принцип “большой = медленный” применим не только в биологии, где предел скорости устанавливается скоростью распространения электрических сигналов между нейронами, но и в будущей жизни, где никакая информация не может распространяться со скоростью больше скорости света. Поэтому для обрабатывающей информацию системы увеличение размера становится и благом и проклятием, требуя известных компромиссов. С одной стороны, бо́льший размер предполагает и бо́льшее число частиц, а следовательно, и более сложные мысли. С другой стороны, это снижает скорость и затрудняет появление по-настоящему глобальных мыслей, так как теперь относящейся к делу информации труднее добираться до всех частей системы.

Поэтому если жизнь наполнит наш космос, то какую форму ей предпочесть – простую и быструю или сложную и медленную? Я предсказываю, что выбор будет тот же, который сделала жизнь на Земле: и то и другое! Обитатели земной биосферы могут быть самых разных размеров, от немыслимых двухсот тонн голубого кита до 10^{-16} кг малютки-пелагиобактера, вклад которой в общую биомассу превышает вклад всех рыб планеты. Более того, крупные организмы со сложной организацией компенсируют свою неповоротливость тем, что некоторые из их модулей маленькие и очень быстрые. Например, ваш мигательный рефлекс работает исключительно быстро именно потому, что для него задействована маленькая и простая цепь, охватывающая лишь ничтожно малую часть мозга. Если та самая неубиваемая муха решит залететь вам в глаз, вы моргнете в десятую долю секунды, даже не успев сообразить, что происходит, так как у информации не будет достаточно времени, чтобы пройти через весь мозг. Организуя обработку информации в процессе прохождения ее через целую иерархию модулей, наша биосфера успешно убивает сразу двух зайцев, достигая и скорости, и сложности. Нам, людям, тоже иногда это удается – например, при такой же иерархической стратегии для оптимизации параллельных вычислений.

Из-за того что коммуникации медленны и дороги, я ожидаю, что продвинутая космическая жизнь сделает то же самое, и вычисления по мере возможности будут осуществляться локально. Если что-то можно посчитать на простом однокилограммовом компьютере, то и не стоит транслировать задачу на всю галактику, это будет контрпродуктивно, так как возвращения каждого запроса нам придется ждать по 100 тысяч лет.

Что в этой будущей системе обработки информации можно считать *сознательным* в смысле привлечения субъективного опыта, и вообще есть ли в ней нечто подобное – тема спорная и увлекательная, мы к ней вернемся в главе 8. Если сознание требует, чтобы различные части системы имели возможность обмениваться друг с другом информацией, то у более крупных систем течение мыслей должно быть замедленным. В то время как вы или будущий суперкомпьютер размером с Землю может обдумать много мыслей за одну секунду, уму размером с галактику требуется 100 тысяч лет на одну-единственную мысль, а космическому разуму в миллиард световых лет в поперечнике хватит времени лишь на

то, чтобы обдумать в общей сложности около десятка мыслей, прежде чем темная энергия раздерет его на несвязанные друг с другом части. С другой стороны, эти несколько драгоценных мыслей и сопровождающий их появление опыт могут оказаться очень глубокими!

Управление иерархиями

Если сама мысль организована в иерархию, охватывающую широкий диапазон различных масштабов, то что можно сказать о власти? В главе 4 мы обсуждали, как разумные сущности естественно организуются в силовые иерархии около состояния равновесия Нэша, и любая сущность будет выбракована, если изменит эту стратегию. Чем более хорошая технология переноса вещества и обработки информации достигается, тем больше может расти иерархия. Если однажды сверхразум достигнет космических размеров, то на что же будет похожа иерархия, обеспечивающая его власть? Будет ли такая иерархия предоставлять независимость своим децентрализованным частям или окажется предельно авторитарной? Будет ли взаимодействие опираться на взаимную выгоду или только на принуждение и угрозы?

Чтобы пролить свет на эти вопросы, давайте рассмотрим кнут и пряник по отдельности: какие существуют стимулы для сотрудничества в космических масштабах, и какие угрозы могут быть использованы, чтобы его обеспечить?

Управление с помощью пряника

На Земле одним из основных традиционных двигателей кооперации всегда была *торговля*, потому что относительная сложность производства одних и тех же вещей в разных местах планеты различна. Добыча килограмма серебра в одних местах обходится в 300 раз дороже добычи одного килограмма меди и всего лишь в 100 раз – в других. Поэтому обе стороны выигрывают при обмене 200 кг меди на 1 кг серебра. Если в каком-то регионе значительно лучше развита промышленность, то аналогичным образом обе стороны выиграют при обмене высокотехнологичных товаров на сырье.

Однако если сверхразум разработает технологии, с помощью которых можно легко переупорядочивать элементарные частицы в любую форму материи вообще, то это приведет к устранению основного стимула для ведения торговли на расстоянии. Зачем беспокоиться по поводу доставки серебра между двумя удаленными планетными системами, когда значительно проще и быстрее превращать медь в серебро, переупорядочивая частицы внутри атомов? Зачем беспокоиться по поводу доставки высокотехнологичной техники из одной галактики в другую, когда ноу-хау и сырье (а для этого сгодится любое вещество) имеются в наличии в обоих местах? Я полагаю, что в космосе, взятом под контроль сверхразума, почти единственным товаром, который стоит доставлять на большие расстояния, окажется *информация*. Единственным исключением может быть вещество, используемое в космических инженерных проектах – например, для противодействия вышеупомянутым разрушительным тенденциям темной энергии разрывать цивилизации. В отличие от традиционной человеческой торговли это вещество может пересылаться в любой удобной форме, даже в виде интенсивного излучения, так как принимающий его сверхразум способен быстро реорганизовать его в любые другие объекты по своему желанию.

Если обмен или торговля информацией выступает как главный мотор космического сотрудничества, то какого рода информация может для этого использоваться? Любая информация может оказаться ценной, если ее производство требует значительных вычислительных усилий и затрат времени. Например, сверхразуму могут потребоваться

ответы на сложные научные вопросы о природе физической реальности, сложные математические задачи, теоремы и оптимальные алгоритмы, сложные инженерные вопросы о новых прорывных технологиях. Гедонистические формы жизни могут интересоваться впечатляющими разновидностями цифровых развлечений и симулированными переживаниями, а космическая коммерция может подогревать спрос на ту или иную форму космической криптовалюты в духе биткойна.

Такого рода обмены могут стимулировать поток информации не только между сущностями примерно равной мощности, но и между сущностями на разных уровнях иерархии – скажем, между узлами уровня планетных систем с хабами галактического уровня или узлами галактического уровня с космическим хабом. Узлы могут стремиться к этому ради удовольствия почувствовать себя частью чего-то большего или чтобы получить ответы на сложные вопросы и технологии, которыми они не смогли бы обеспечить себя самостоятельно, а также для защиты от внешних угроз. Они могут также оценить обещания практического бессмертия посредством резервного копирования: как многие люди находят утешение в убеждении, что их души будут жить после физической смерти их тела, продвинутый искусственный интеллект может верить в будущую жизнь своей души и накопленных знаний в хабе суперкомпьютера после того, как его оригинальный физический “хард” истощит запасы энергии.

И наоборот, хабу может потребоваться, чтобы связанные с ним узлы помогли ему в решении вычислительной задачи, требующей больших ресурсов и вычислительного времени, но не требующей результатов срочно, так что не проблема подождать несколько тысяч или миллионов лет ради получения ответов. Как мы видели выше, хаб может также захотеть, чтобы его узлы помогли ему с проведением таких массивных космических инженерных проектов, как противодействие разрушительной темной энергии путем перемещения значительных галактических масс и концентрации их в меньшей области пространства. Если окажется, что возможно создание проходимых кротовых нор, главным приоритетом для хабов, вероятно, станет построение их сети для нейтрализации темной энергии и полной взаимосвязанности всех частей системы (космической империи). Вопрос о возможных окончательных целях космического сверхразума настолько увлекателен, что мы рассмотрим его отдельно в главе 7.

Управление при помощи кнута

Земные империи обычно принуждали свои провинции к сотрудничеству с помощью кнута и пряника. Хотя подданные Римской империи ценили технологии, инфраструктуру и покровительство, которые они получали в качестве награды за свое сотрудничество, они также боялись неизбежных последствий за проявления непокорности или за неуплату налогов. Из-за того что войска из Рима в отдаленные провинции добирались долго, необходимое возмездие отчасти делегировалось местной власти и ее войскам, уполномоченным осуществлять наказание практически мгновенно. Сверхразумный хаб может использовать аналогичную стратегию и разворачивать сеть верной себе гвардии по всей своей космической империи. Поскольку управление сверхразумными подданными может оказаться нелегким, простейшая жизнеспособная стратегия заключается в том, чтобы использовать AI-гвардию из относительно тупых, но запрограммированных на 100-процентную лояльность машин, находящихся в режиме постоянного мониторинга за соблюдением всех правил при условии их автоматического подрыва при малейшем нарушении.

Предположим, например, что искусственный интеллект хаба формирует белый карлик так, чтобы он расположился у края цивилизации размером с Солнечную систему, от которой он требует подчинения. Белый карлик – это выгоревшие останки умеренно тяжелой звезды. Он состоит в основном из углерода и напоминает гигантский алмаз

в небе. Он настолько компактен, что может весить больше, чем Солнце, уступая Земле по размеру. Известный индийский физик Субраманьян Чандрасекхар доказал, что если вы постепенно увеличиваете массу белого карлика, то при достижении определенного предела, известного как предел Чандрасекхара и составляющего около 1,4 массы Солнца, в белом карлике случится термоядерный катаклизм и он превратится в сверхновую типа 1A. Если искусственный интеллект хаба бессердечно формирует белый карлик очень близким к пределу Чандрасекхара, то AI-гвардия может быть эффективной, даже будучи чрезвычайно тупой (а в действительности как раз именно в силу своей тупости): ей достаточно просто быть запрограммированной на проверку поступления от покоренной цивилизации ежемесячной квоты космических биткойнов, математических доказательств или какой-либо иной формы налогов, чтобы в том случае, если это не было сделано, добавить белому карлику немного массы, запустить сверхновую и разнести всю непокорную цивилизацию в пух и перья.

Цивилизацию галактического размера можно контролировать аналогичным образом, если расположить большое количество компактных объектов на низких орбитах вокруг гигантской черной дыры в центре галактики и угрожать превратить эти объекты в газ, например сталкивая их друг с другом. Этот газ начнет падать в черную дыру, превращая ее в мощный квазар, из-за чего бо́льшая часть галактики станет необитаемой.

Подводя итог, можно сказать, что для сотрудничества на космических расстояниях будущая жизнь располагает сильными стимулами, но отнюдь не ясно, будут ли главной действующей силой такого сотрудничества соображения взаимной выгоды или жесткие угрозы – ограничения, налагаемые законами физики, совместимы с обоими сценариями, так что результат будет зависеть от преобладающих целей и ценностей. Мы будем исследовать нашу способность повлиять на эти цели и ценности будущей жизни в главе 7.

Когда цивилизации сталкиваются

До сих пор мы рассматривали сценарии, когда будущая жизнь расширяется в космическом пространстве из единого центра интеллектуального взрыва. Но что случится, если жизнь эволюционирует независимо более чем в одном месте и две расширяющиеся цивилизации встретятся?

Если рассмотреть случайно выбранную планетную систему, то есть ненулевая вероятность, что на одной из ее планет возникнет жизнь, которая будет развивать передовые технологии и со временем выйдет в космос. Мы полагаем, что эта вероятность больше нуля, потому что технологически развитая жизнь возникла здесь, в нашей Солнечной системе, и законы физики, по-видимому, допускают ее существование в космосе. Если космическое пространство достаточно велико (а теория космологической инфляции предполагает, что оно огромно, а может быть, и бесконечно), то таких вышедших в космос цивилизаций должно быть много, как показано на рис. 6.10. В статье Джея Олсона, о которой шла речь выше, содержится элегантный анализ таких расширяющихся в космос биосфер, и Тоби Орд вместе с коллегами из Института будущего человечества также провел аналогичный анализ. Если рассматривать эти расширяющиеся в космосе биосферы в трех измерениях, то они окажутся сферами в буквальном смысле, поскольку цивилизации расширяются с одной и той же скоростью во всех направлениях. В пространственно-временных координатах они выглядят как верхняя часть бокала для шампанского (см. рис. 6.7), потому что темная энергия создает непреодолимый предел числу галактик, которого данная цивилизация когда-либо сможет достичь.

Если расстояния между соседними цивилизациями превышают пределы, в которых темная энергия позволяет им расширяться, то они никогда не вступят в контакт и даже не узнают о существовании друг друга. Будут чувствовать себя одиночками в космосе. Но если наш космос богаче цивилизациями и соседи находятся ближе друг к другу, некоторые

сферы в конце концов могут перекраться. Что произойдет в области этих перекрытий? Будет там развиваться сотрудничество, или начнется конкуренция, война?

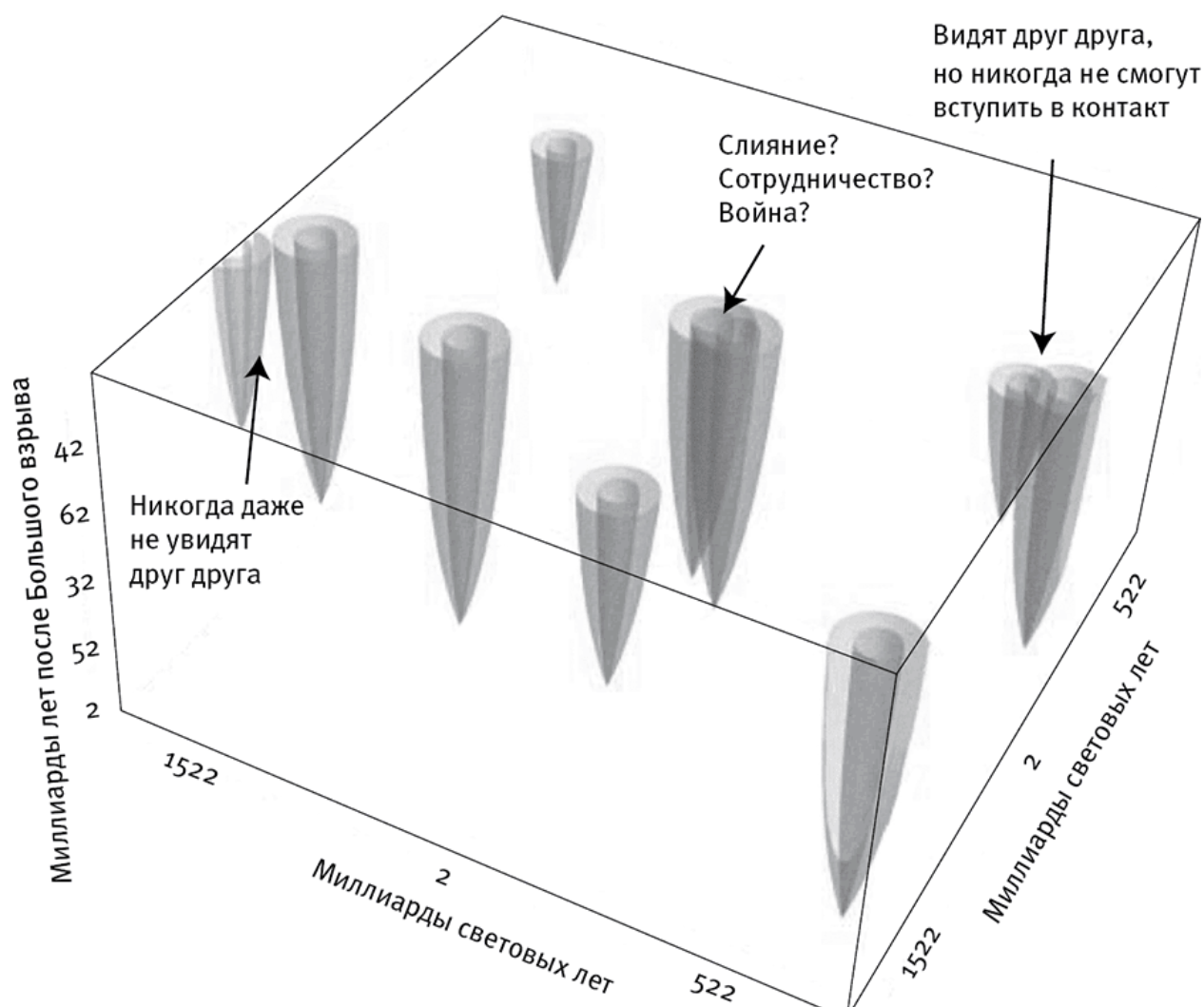


Рис. 6.10

Если жизнь возникает независимо во множестве точек пространства-времени (в разных местах и в разные моменты времени) и начинает колонизировать космос, то в космосе образуется сеть расширяющихся биосфер, каждая из которых напоминает верхнюю часть бокала для шампанского с рис. 6.7. Нижняя часть каждой биосферы соответствует месту и времени, где и когда колонизация началась. Прозрачный и вложенный в него непрозрачный бокалы соответствуют колонизации со скоростью света и со скоростью вдвое меньшей скорости света, области перекрытия иллюстрируют возможность встречи независимых цивилизаций.

Европейцы смогли завоевать Африку и обе Америки, потому что они обладали более передовыми технологиями. В противоположность этому представляется правдоподобным, что задолго до того, как две сверхразумные цивилизации встретятся, уровень их технологий достигнет плато, откуда дальнейшее развитие невозможно в силу законов физики. Поэтому маловероятно, чтобы один сверхразум мог легко завоевать другой, даже если бы захотел. Кроме того, если эволюция их целей привела их к относительному соответствию, каких-либо оснований для состязания или войны может и не возникнуть. Например, если обе стремятся доказать как можно больше красивых математических теорем и изобрести самые умные алгоритмы, то от объединения своих усилий и обмена своими находками обе выиграют. В конце концов, информация тем

и отличается от ресурсов, за которые люди обычно ведут борьбу, что ее можно одновременно и отдавать, и сохранять.

У некоторых расширяющихся цивилизаций могут быть цели, которые принципиально неизменны, вроде целей фундаменталистского культа или распространяющегося вируса. Однако не менее вероятно наличие у некоторых из передовых цивилизаций большего сходства с независимо мыслящими людьми, которые готовы корректировать свои цели, когда им предъявляют достаточно убедительные аргументы. При встрече двух таких цивилизаций возможно столкновение, но не вооруженное, а на уровне идей, и тогда та цивилизация, чьи идеи окажутся более убедительными, получит приоритет для своих целей и сможет распространить их со скоростью света через области, управляемые другой цивилизацией. Ассимиляции соседей обеспечивает расширение со значительно большей скоростью, чем расселение, так как ваша сфера влияния может распространяться со скоростью, с которой передаются идеи (это скорость света при использовании телекоммуникаций), тогда как для физического расселения скорость света заведомо недостижима. Такая ассимиляция будет не насильственной, как осуществляемая боргами в телесериале *Star Trek*, а добровольной, основанной на убедительном превосходстве идей, в интересах самих ассимилированных.

Мы видели, что в будущем в космосе могут содержаться быстро растущие пузыри двух видов – расширяющиеся цивилизации и те пузыри смерти, которые расширяются со скоростью света, делая пространство необитаемым, поскольку внутри уничтожаются все знакомые нам элементарные частицы. Растущей цивилизации, таким образом, может встретиться область одного из трех типов – необитаемые, пузыри жизни и пузыри смерти. Если она боится неконструктивного соперничества цивилизаций, то у нее есть сильный стимул к запуску программы быстрого “захвата земель”, чтобы расселиться в необитаемых областях, пока этого не сделали возможные соперники. Однако тот же экспансионистский стимул может быть силен даже в отсутствие других цивилизаций, просто для захвата ресурсов, пока темная энергия не сделала их недоступными. Мы только что видели, как встреча с другой растущей цивилизацией может быть лучше, может быть и хуже, чем встреча с необитаемой областью пространства, в зависимости от того, насколько обнаруженный сосед открыт для сотрудничества. Однако всегда лучше наткнуться на любую экспансионистскую цивилизацию (даже пытающуюся превратить вашу цивилизацию в скрепки для бумаги), чем на пузырь смерти, который продолжит расширение со скоростью света независимо от того, будете ли вы пытаться как-то бороться с ним или сотрудничать. Наша единственная защита от пузырей смерти – это темная энергия, которая не даст наиболее далеким из них когда-либо достичь нас. Так что если пузыри смерти действительно существуют, то темная энергия для нас на самом деле не обязательно враг, но кое в чем и друг.

Мы одиноки?

Многие люди принимают как данность существование высокоорганизованных форм жизни в большей части нашей Вселенной, так что вымирание нашего человечества, с этой точки зрения, в космической перспективе не имеет большого значения. В конце концов, почему мы должны беспокоиться по поводу своего исчезновения, если на наше место скоро придет другая цивилизация, прекрасная и похожая на ту, что нам показали в духоподъемном телесериале *Star Trek*, которая заново наполнит жизнью нашу Солнечную систему и, возможно, даже использует свои передовые технологии для нашего воскрешения и возвращения к жизни? Я рассматриваю подобные допущения, сделанные в *Star Trek*, как опасные, потому что они усыпляют нашу бдительность и рождают в нас ложное чувство безопасности, отчего нашей цивилизации грозят апатия и безрассудство.

В действительности, я думаю, что предположение, будто мы не одиноки в нашей Вселенной, не только опасно, но и, вероятно, ложно.

Это взгляд меньшинства^[49], и я, возможно, ошибаюсь, но это как минимум вариант, который мы не можем сбрасывать со счетов и который морально обязывает нас вести себя ответственно, не доводя нашу цивилизацию до возможного вымирания.

Когда я читаю лекции по космологии, я часто прошу поднять руки тех, кто думает, что где-то в нашей Вселенной (то есть в той области пространства, откуда свет мог дойти до нас за прошедшие 13,8 миллиарда лет со времени Большого взрыва) еще есть разумная жизнь. И неизменно, практически каждый раз, все в аудитории, от детей детсадовского возраста до студентов колледжа, поднимают руки. Когда я спрашиваю почему, чаще всего я слышу в ответ, что наша Вселенная очень велика, и поэтому жизнь должна быть где-то еще – по крайней мере, из статистических соображений. Давайте повнимательнее рассмотрим этот аргумент, чтобы понять его слабость.

Все дело сводится к одному-единственному числу – характерному расстоянию между среднестатистической цивилизацией (см. рис. 6.10) и ее ближайшим соседом. Если это расстояние намного больше 20 миллиардов световых лет, мы должны думать, что мы одиноки в нашей Вселенной (той части пространства, откуда свет мог достичь нас за прошедшие 13,8 миллиарда лет со времени Большого взрыва) и никогда не вступим в контакт с инопланетянами. Можем ли мы оценить это расстояние? У нас для этого довольно мало данных. Это означает, что минимальное расстояние до ближайшего соседа мы могли бы оценить округленно в 1000 ... 000 метров, где общее количество нулей вполне могло бы равняться 21, 22, 23, ..., 100, 101, 102 или даже больше, но вряд ли намного меньше 21, поскольку мы к настоящему времени не видели никаких убедительных свидетельств об инопланетянах (см. рис. 6.11). Для наших ближайших соседей во Вселенной, чей радиус примерно 10^{26} метров, количество нулей не может превышать 26, и поэтому вероятность, что число нулей окажется лежащим в узком интервале между 22 и 26, очень мала. Вот почему я думаю, что мы одиноки в нашей Вселенной.

Я даю подробное обоснование этого аргумента в книге *Our Mathematical Universe*^[50], поэтому я не буду пересказывать его здесь, но основная причина, по которой мы не можем оценить расстояние до ближайшего соседа, заключается в том, что мы, в свою очередь, ничего не знаем о вероятности возникновения разумной жизни в данном конкретном месте. Как отмечал американский астроном Фрэнк Дрейк, эта вероятность может быть вычислена путем перемножения вероятности наличия пригодной для жизни среды (скажем, подходящей для этого планеты), вероятности возникновения там жизни и вероятности превращения этой жизни в разумную. Когда я учился в университете, у нас не было способа оценить ни одну из этих трех вероятностей. Сейчас, после важных открытий в последние два десятилетия планет в планетных системах других звезд, представляется вполне вероятным, что пригодных для развития жизни планет очень много, миллиарды только в нашей собственной галактике. Но вероятность развития на них жизни, а затем и возникновения разума остается по-прежнему крайне неопределенной: некоторые эксперты считают, что одна из них или даже обе с неизбежностью довольно высоки, и поэтому разумная жизнь встречается на большинстве пригодных для жизни планет; в то же время другие эксперты утверждают, что либо одна из них, либо обе должны быть чрезвычайно низки, так как для прохождения некоторых узких мест в эволюции требуется немыслимое везение. Некоторые из этих узких мест на ранних стадиях самовоспроизводящейся жизни напоминают дилемму курицы и яйца: например, чтобы современная клетка могла построить рибосому – довольно сложную молекулярную машину, способную прочитывать наш генетический код и синтезировать белки, – требуется другая рибосома, и пока отнюдь не очевидно, как самая первая рибосома могла бы развиваться постепенно от чего-то более простого^[87]. Другие узкие места возникают при появлении разума. Например, хотя

динозавры царили на Земле более 100 миллионов лет, – в тысячу раз дольше, чем на ней существуют современные люди, – похоже, что эволюция совсем не торопилась подтолкнуть их к развитию интеллекта, изобретению телескопа или созданию компьютеров.

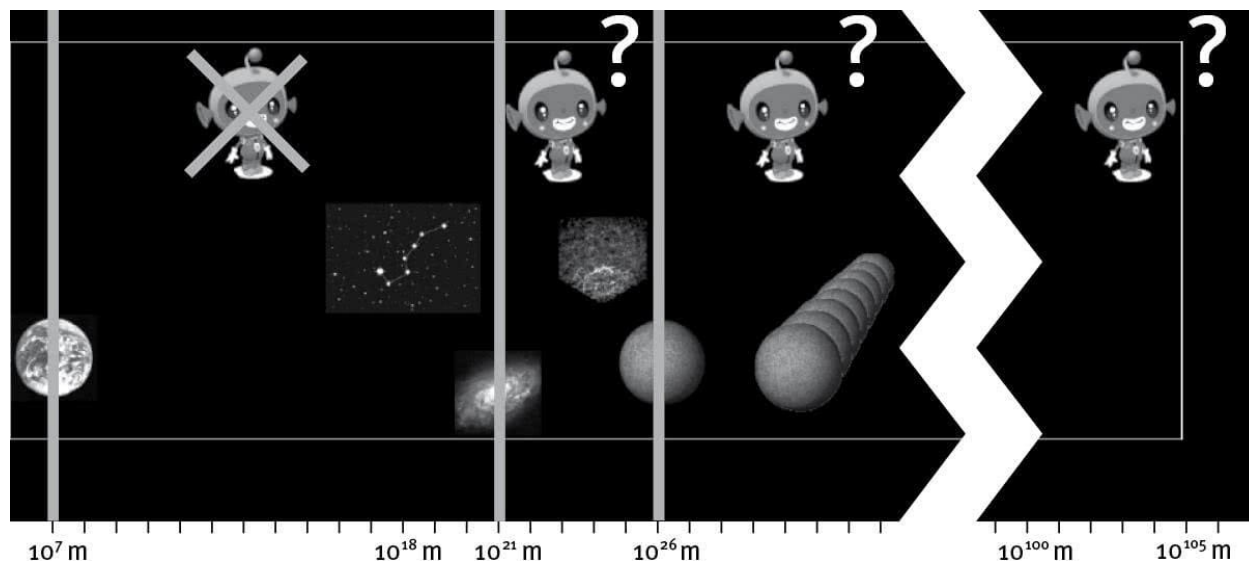


Рис. 6.11

Одиноки ли мы во Вселенной? Колоссальная неопределенность в условиях, необходимых для возникновения жизни и разума, заставляет предположить, что ближайшая к нам цивилизация может оказаться в любой точке горизонтальной оси под рисунком. Из-за этого вероятность того, что она окажется в узком промежутке между краем нашей галактики (около 10^{21} метров от нас) и краем нашей Вселенной (около 10^{26} метров от нас) очень невелика. Если такая цивилизация ближе к нам, то их должно быть много в нашей галактике, и тогда мы должны были бы их уже заметить, а это и означает, что в нашей Вселенной мы, скорее всего, одиноки.

Некоторые люди отвергают этот мой аргумент, говоря, что да, разумная жизнь *могла* быть редка, но на самом деле она вовсе не так уж и редка. Наша галактика просто кишит разумной жизнью, только ученые не хотят ее замечать. Возможно, инопланетяне уже посещали Землю, как утверждают многие энтузиасты уфологии. Возможно, инопланетяне не посещали Землю, но все время находились поблизости, сознательно скрываясь от нас (это предположение получило название “гипотеза зоопарка”, ее выдвинул американский астроном Джон Болл, и она легла в основу многих классических произведений научной фантастики, таких как *Star Maker* Олафа Стэплдона). Или, возможно, они, будучи где-то рядом, совсем и не собираются от нас скрываться, но просто не заинтересованы ни в расселении в космосе, ни в крупных инженерных проектах, которые мы могли бы заметить.

Конечно, мы не должны заранее отвергать все эти возможности, но так как ни для одной из них нет общепринятых доказательств, мы также должны принимать всерьез и альтернативное высказывание – что мы во Вселенной одиноки. Кроме того, я думаю, что мы не должны недооценивать многообразия передовых цивилизаций, предполагая, что все они разделяют приверженность целям, из-за которых остаются для нас невидимыми. Выше мы могли убедиться, что освоение ресурсов – цель, которую цивилизациям вполне естественно иметь и стремление к которой делает эти цивилизации для нас заметными, и достаточно всего лишь одной цивилизации поставить перед собой такую цель, чтобы немедленно наполнить собой всю галактику и выйти далеко за ее пределы. Встретившись с фактом наличия в нашей галактике миллионов землеподобных планет, пригодных для

жизни, которые на миллиарды лет старше Земли и, стало быть, располагали достаточным временем для реализации амбициозных планов своих жителей расселиться по галактике, мы не должны отказываться от наиболее очевидного его истолкования: что зарождение жизни требует стечения настолько маловероятных факторов, что все эти планеты остаются необитаемыми.

Если жизнь все же не редкость, мы можем об этом скоро узнать. Сейчас проводятся остроумные астрономические исследования, нацеленные на обнаружение в атмосферах землеподобных планет кислорода, который мог бы служить доказательством наличия на этих планетах жизни. Параллельно с таким поиском *любых* форм жизни ведется поиск *разумных* форм жизни, для чего российский филантроп Юрий Мильнер выделил недавно 100 миллионов долларов на проект Breakthrough Listen.

Важно быть не слишком антропоцентричными при поиске продвинутых форм жизни: если мы обнаруживаем внеземную цивилизацию, она, вероятно, уже стала сверхразумной. Как Мартин Рис сформулировал это в одном из своих недавних эссе, “история человеческой технологической цивилизации измеряется в веках – и, может быть, осталось только одно или два столетия, прежде чем людей обгонит или с ними сольется неорганический разум, который затем будет сохраняться, продолжая развиваться, миллиарды лет... Для нас было бы крайне маловероятно „поймать“ [иной] разум в тот краткий миг, когда он еще принимает органические формы”^[88]. И я согласен с выводом Джея Олсона в упоминавшейся выше статье о космических расселениях: “Возможность, что продвинутые формы разумной жизни станут использовать ресурсы Вселенной просто для заселения существующих землеподобных планет несколько улучшенной версией людей, рассматривается нами как маловероятная конечная точка развития технологии”. Так что когда вы представляете себе инопланетян, не надо думать о зеленых человечках с двумя руками и двумя ногами, подумайте о сверхразумной наполняющей космос жизни, какую мы рассмотрели ранее в этой главе.

Хотя я всецело поддерживаю все ведущиеся поиски внеземной жизни, которые могли бы пролить свет на один из самых увлекательных научных вопросов, я втайне надеюсь на их провал, ибо уверен, что им не найти ничего! Явное несоответствие между обилием пригодных для жизни планет в нашей галактике и отсутствием внеземных пришельцев, известное как парадокс Ферми, указывает на существование того, что экономист Робин Хансон назвал “большим фильтром”, – то есть эволюционного / технологического слагаемого где-то на пути развития от неживой материи к расселяющейся в космосе разумной жизни. Если мы обнаружим независимо развивающуюся жизнь где-то в другом месте, это будет означать, что простейшая жизнь – не редкость и что слагаемое установлен уже после нашей нынешней человеческой стадии развития. Может быть, потому, что расселение в космосе почему-либо невозможно, может быть, потому, что почти все развитые цивилизации самоликвидировались раньше, чем достигли состояния, позволяющего осваивать космос. Поэтому я скрещиваю пальцы, чтобы поиски внеземной жизни закончились неудачей: она бы хорошо вписалась в сценарий, где развитие разумной жизни редкость, но нам, людям, несказанно повезло, и слагаемое у нас уже позади, а впереди нас ждет невероятное и фантастическое будущее.

Виды на будущее

До сих пор в этой книге мы занимались прошлым – историей жизни в нашей Вселенной, от ее скромного начала миллиарды лет тому назад к началу, возможно, великого будущего, которое продлится еще миллиарды лет. Если наши нынешние исследования искусственного интеллекта со временем приведут к интеллектуальному взрыву и оптимизации космических расселений, это будет поистине взрывом космического масштаба: после миллиардов лет, проведенных в ничтожной флуктуации посреди равнодушно безжизненного космоса, жизнь внезапно прорывается на космическую арену, подобно сферической ударной волне распространяясь со скоростью близкой к скорости света, ни на миг не останавливаясь и поджигая все на своем пути искрами жизни.

Подобные оптимистические взгляды на важное значение жизни в нашем космическом будущем красочно формулировали многие мыслители, процитированные в этой книге. Поскольку авторов научно-фантастических сочинений часто игнорировали как оторванных от реальности романтических мечтателей, я полагаю, что теперь история возьмет реванш, показав, что большинство из них, как и большинство рассуждающих о космических поселениях ученых, оказались слишком пессимистичными в отношении будущего сверхразума. Например, мы убедились, что межгалактические путешествия становятся намного проще, если люди и другие разумные существа смогут передаваться в цифровом виде, и это потенциально делает нас хозяевами собственной судьбы не только в нашей Солнечной системе или в галактике Млечный путь, но и вообще в космосе.

Выше мы рассмотрели как весьма вероятную возможность, что мы в нашей Вселенной – единственная высокоразвитая цивилизация. Давайте посвятим оставшуюся часть главы изучению этой возможности и той огромной моральной ответственности, которую она за собой влечет. По прошествии 13,8 миллиардов лет жизнь в нашей Вселенной оказалась у развилки, когда надо выбирать между космическим процветанием и самоуничтожением. Если мы не станем дальше улучшать наши технологии, то вопрос будет состоять не в том, суждено ли человечеству вымереть, а в том, как именно вымирание осуществится. Что достанет нас первым – астероид, супервулкан, жар стареющего Солнца или какое-то другое бедствие (см. рис. 5.1)? В наше отсутствие космическая драма, описанная Фрименом Дайсоном, будет разыгрываться уже без зрителей: избавление от космокалипсиса, выгорание звезд, угасание галактик и испарение черных дыр – каждый акт этой пьесы завершится колоссальным взрывом, высвобождающим в миллионы раз больше энергии, чем Царь-бомба, самая мощная водородная бомба из когда-либо построенных. Как писал Дайсон: “Холодная, расширяющаяся Вселенная будет еще долго подсвечиваться случайными фейерверками”. Увы, эти фейерверки будут бессмысленным переводом энергии, там не будет никого, кто мог бы насладиться ими.

Без развития технологий нашего человеческого вымирания не надо будет долго ждать, и в космологическом контексте протяженностью в десятки миллиардов лет весь драматизм нашей жизни во Вселенной предстанет лишь краткой вспышкой красоты, страсти и смысла в близком к вечности бессмысленном и бессубъектном переживании. Что это будет за упущенная возможность! Если вместо отказа от технологий мы решим развивать их, тогда повышаем наши ставки: на кону наша жизнь, и в случае выигрыша мы продолжаем жить и процветать, а в случае проигрыша уходим со сцены даже быстрее, разрушая себя неправильным планированием (см. рис. 5.1). Я отдаю свой голос за технологии, и мой выбор определяется не слепой верой в то, что мы строим, но осторожностью, дальновидностью и тщательным планированием будущего.

По прошествии 13,8 миллиардов лет космической истории мы оказываемся в умопомрачительно красивой Вселенной, которая через нас, людей, стала живой

и начала осознавать себя. Мы видели, что потенциал будущей жизни в нашей Вселенной грандиознее, чем самые смелые мечты наших предков, но его омрачает столь же реальная опасность для разумной жизни погибнуть навсегда. Реализует ли жизнь свой потенциал в нашей Вселенной? В значительной степени это зависит от того, что мы, люди, живущие сегодня, сделаем за отведенный нам век. Я оптимист, и я уверен, что мы можем сделать будущую жизнь действительно прекрасной, выбрав сегодня правильный путь. Чего мы должны хотеть, и как мы можем этого добиться? Давайте потратим остальные части книги обсуждению некоторых из наиболее сложных проблем и наших возможностей в их решении.

Подведение итогов

- По сравнению с космическими сроками в миллиарды лет интеллектуальный взрыв длится мгновение, за которое технологии выходят на плато, где их ограничивают только законы физики.
- Это технологическое плато расположено значительно выше, чем любые сегодняшние технологии; достигнув его, мы сможем, взяв то же количество вещества, генерировать в десятки миллиардов раз больше энергии (используя сфалероны или черные дыры), хранить на 12–18 порядков больше информации, проводить на 31–41 порядок больше вычислений в единицу времени или преобразовать это вещество в любую другую желаемую форму.
- Сверхразумная жизнь не только радикально увеличивает эффективность использования имеющихся ресурсов, но и ускоряет рост биосферы на 32 порядка, что открывает доступ к дополнительным ресурсам посредством космического расселения со скоростью близкой к скорости света.
- Темная энергия ограничивает космическую экспансию сверхразумной жизни, но в то же время защищает ее от раздувающихся вдали пузырей смерти или враждебных цивилизаций. Угроза темной энергии, раздирающей космические цивилизации, мотивирует на осуществление массивных космических инженерных проектов, включая строительство кротовой норы, если это будет возможным.
- Основой товарных отношений на космических расстояниях станет обмен информацией.
- Если создание кротовой норы окажется невозможным, то ограничение скорости обмена информацией скоростью света создаст серьезные проблемы в координации и управлении внутри космической цивилизации. Удаленный центральный хаб может стимулировать искусственный интеллект связанных с ним узлов к сотрудничеству наградами или угрозами, например развертывая локальную AI-гвардию, которая может быть запрограммирована так, чтобы уничтожать не подчиняющиеся правилам узлы, превращая материнскую звезду в местной планетной системе в сверхновую или квазар.
- Встреча двух расширяющихся цивилизаций может привести к ассимиляции их, к сотрудничеству между ними или войне, причем последнее, возможно, менее вероятно, чем война между сегодняшними цивилизациями.
- Несмотря на широко распространенное мнение, что мы не одиноки во Вселенной, справедливо, вероятно, противоположное, и только наша цивилизация в будущем может сделать нашу Вселенную живой.
- Если мы не будем улучшать наши технологии, речь будет идти не о том, погибнет ли человечество, а о том, каким образом оно погибнет: будет ли астероид, супервулкан, нарастающий жар стареющего Солнца или какое-то другое бедствие непосредственной причиной нашей гибели.
- Если мы продолжим улучшать наши технологии с достаточной осмотрительностью, прогнозируя последствия возможных ошибок, дабы не совершать их, у земной жизни есть колоссальный потенциал к процветанию как на Земле, так и далеко за ее пределами на протяжении миллиардов лет, о чем не могли мечтать самые дерзкие из наших предков.

Глава 7

Цели

Тайна бытия человеческого не в том, чтобы только жить, а в том, для чего жить.

Ф. Достоевский. Братья Карамазовы

Жизнь – это путешествие, а не пункт назначения.

Ральф Уолдо Эмерсон

Если бы мне надо было одним словом выразить, в чем сложность AI-спора, я бы выбрал слово “цели”. Должны ли мы определить цели для искусственного интеллекта? И если да, то какие именно цели? Кто сможет поставить эти цели? Сможем ли мы убедиться, что эти цели сохранятся даже при развитии искусственного интеллекта? Сможем ли мы изменить цели, если AI станет умнее нас? Какие наши основные цели?

Эти вопросы не только сложны – они основополагающи для будущего жизни: если мы не знаем, чего хотим, маловероятно, что мы сможем этого достичь, а если мы передадим контроль машинам, которые не разделяют наших целей, то, скорее всего, мы получим совсем не то, чего бы нам хотелось.

Физика: происхождение целей

Дабы пролить свет на эти вопросы, давайте сначала изучим основы происхождения целей. Когда мы оглядываемся вокруг, некоторые процессы кажутся нам нацеленными на определенный результат, а другие нет. Рассмотрим для примера такой процесс: ударили по футбольному мячу, чтобы забить решающий гол в игре. Само поведение мяча при этом не кажется целенаправленным и экономнее всего описывается ньютоновским законом движения как реакция мяча на удар. С другой стороны, поведение игрока экономнее всего объясняется не механистически (атомы толкают все вокруг себя), а тем, что у него была цель – добиться для своей команды улучшения счета. Каким образом такое целенаправленное поведение вырастает из физики нашей ранней Вселенной, которая состояла в основном из группы частиц, ударяющихся обо все вокруг себя, на первый взгляд, безо всякой цели?

Любопытно, что источник целенаправленного поведения может быть найден в самих законах физики и обнаруживает себя даже в простых процессах, которые не подразумевают наличия жизни. Если девушка-спасатель приходит на помощь купальщику, как показано на рис. 7.1, мы ожидаем от нее движения по прямой, но она пробегает дальше по берегу, где можно двигаться быстрее, чем в воде, а потом немного изменяет направление, когда входит в воду. Мы естественно объясняем ее выбор траектории наличием конкретной цели, так как из всех возможных она осознанно выбирает оптимальную, которая поможет ей добраться до утопающего быстрее всего. Но ведь и луч света преломляется, когда проходит через воду (см. рис. 7.1), так чтобы время пробега до конечной цели оказалось минимальным. Как такое возможно?

Это явление известно в физике и называется *принципом Ферма* – по имени ученого, описавшего его в 1662 году и предложившего свое объяснение поведения светового луча. Примечательно, что физики с тех пор открыли, что *все* законы классической физики могут быть математически переформулированы аналогичным способом: из всех путей, которыми располагает природа, чтобы что-то сделать, она всегда выбирает оптимальный, который обычно сводится к минимизации или максимизации количества. Существует два математически эквивалентных способа сформулировать закон физики: либо через связь прошлого с будущим, либо через природную оптимизацию чего-нибудь. И хотя второй способ обычно не упоминается в элементарных курсах физики, потому что он предполагает больше математических расчетов, мне он кажется более элегантным и обоснованным. Когда человек пытается что-то оптимизировать (счет, благосостояние или счастье), мы склонны описывать такое поведение как целенаправленное. Поэтому, если сама природа старается что-то оптимизировать, нет ничего удивительного в том, что проявляется целенаправленное поведение: оно было зашито у нее в “харде” самими законами физики.

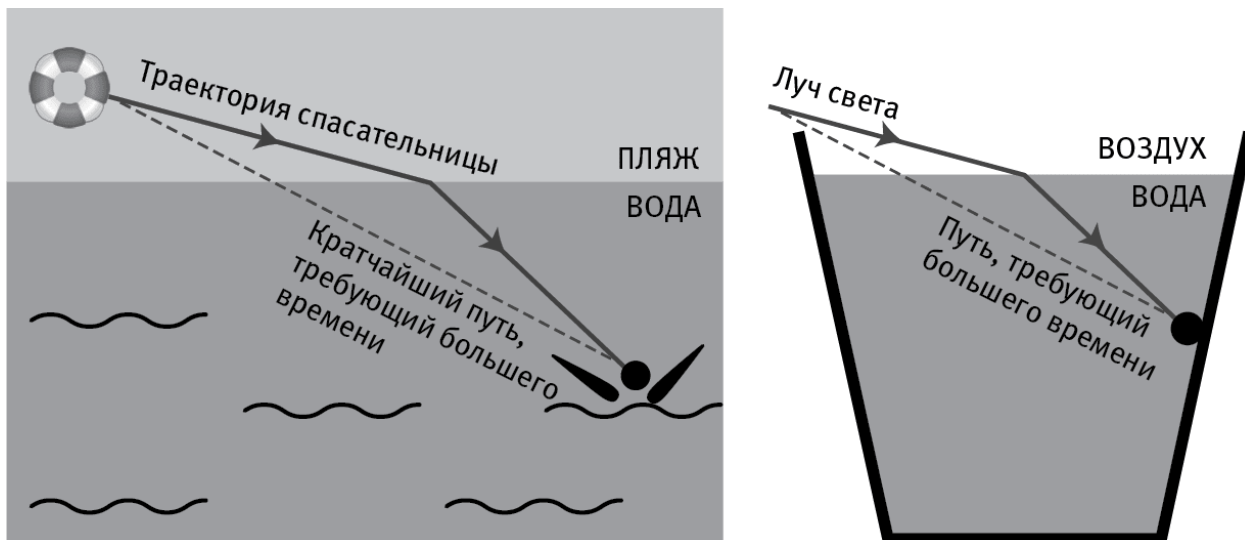


Рис. 7.1

Чтобы скорее добраться до тонущего купальщика, спасательнице надо двигаться не по прямой (изображенной пунктиром), а по ломаной линии, выигрывая время за счет того, что по берегу она может двигаться быстрее, чем плыть в воде. Луч света в воде подчиняется тому же правилу: преломившись, он достигает дна или стенки сосуда за кратчайшее время.

Известна физическая величина, которую природа старается максимизировать, – это *энтропия*, или, грубо говоря, мера беспорядка в описываемых предметах. Второй закон термодинамики утверждает, что энтропии свойственно увеличиваться до тех пор, пока она не достигнет максимально возможного значения. Если забыть на время о наличии сил тяжести, то конечная стадия, когда максимальный беспорядок повсеместно достигнут, называется *тепловой смертью* – это означает, что все обретает скучное безупречное однообразие, без сложности, без жизни и без изменений. Когда вы, например, наливаете холодное молоко в горячий кофе, ваш напиток начинает безудержно стремиться к своей собственной тепловой смерти и вскоре превращается в однородную тепловатую жидкость. При смерти живого организма энтропия в нем также возрастает, и вскоре организованность его частиц существенно снижается.

Стремление природы увеличить энтропию объясняет, почему время имеет определенное направление, заставляя фильмы выглядеть нереалистично, если просматривать их задом наперед: если вы уроните бокал вина, вы ожидаете, что он разлетится вдребезги, ударившись об пол, и тем самым увеличит глобальный хаос (энтропию). Если же вы увидите, как он собирается и летит назад в руку (энтропия при этом явно уменьшается), то, скорее всего, вы решите больше из него не пить, посчитав, что вам на сегодня хватит.

Когда я впервые узнал о нашем неумолимом движении по направлению к тепловой смерти, я очень расстроился – и был в этом не одинок: один из основателей термодинамики, лорд Кельвин, писал в 1841 году, что “в результате [все] неизбежно придет в состояние всеобъемлющего покоя и смерти”, и сложно утешать себя тем, что такова, видно, долгосрочная цель природы – торжество смерти и разрушения. Однако последующие исследования показали, что все не так плохо. Во-первых, гравитация ведет себя не так, как все другие силы, и старается сделать нашу Вселенную не однообразной и скучной, а все более разнообразной и интересной. Благодаря этому гравитация превратила нашу скучную раннюю Вселенную, которая была абсолютно однообразна, в сегодняшний прекрасный и сложный космос, наполненный галактиками, звездами и планетами. Благодаря гравитации во Вселенной сегодня колоссальный разброс температур, который позволяет жизни процветать, лавируя между горячим и холодным: мы живем на теплой комфортной планете, которая сначала поглощает солнечную

энергию, пришедшую с поверхности нагретого до 6 000 °С светила, а потом излучает ее, отдавая холодному космическому пространству, температура которого всего на три градуса отличается от абсолютного нуля.

Во-вторых, недавняя работа моего коллеги по MIT Джереми Ингланда с соавторами принесла нам хорошие новости: термодинамика находит в природе и более вдохновляющую цель, чем тепловая смерть^[89]. Эта цель называется диковатым словосочетанием *диссипативно-направленная адаптация*, подразумевающим, что случайно сформировавшиеся группы частиц стремятся самоорганизоваться таким образом, чтобы получать энергию из окружающей среды наиболее эффективным способом (слово “диссипация” здесь означает, что энергия распределяется между степенями свободы с увеличением энтропии, то есть превращается в тепло, зачастую производя попутно какую-то полезную работу). Например, в группе молекул, выставленных на солнце, со временем проявится тенденция так расположиться по отношению друг к другу, чтобы лучше поглощать солнечный свет. Другими словами, природа, похоже, сама собой нацелена на производство самоорганизующихся систем, которые, все усложняясь и усложняясь, все больше походят на жизнь, и эта цель зашита в “хард” самих физических законов.

Как мы можем увязать это космическое стремление к жизни с космическим стремлением к тепловой смерти? Ответ можно найти в известной книге 1944 года *What's Life?* Эрвина Шрёдингера^[51], одного из основателей квантовой механики. Шрёдингер указал на то, что живая система поддерживает свою энтропию на постоянном уровне или даже сокращает ее за счет увеличения энтропии вокруг себя, – и в этом отличительная черта живых систем. Другими словами, второй закон термодинамики имеет лазейку: хотя всеобщая энтропия должна увеличиваться, в некоторых местах разрешается сокращать энтропию, – при условии, что она еще больше увеличивается где-то еще. Так жизнь поддерживает или увеличивает свою сложность за счет создания хаоса вокруг себя.

Биология: эволюция целей

Мы только что увидели, как происхождение целенаправленного поведения можно отследить в законах физики, из-за которых, кажется, даже у элементарных частиц есть цель вести себя таким образом, чтобы выжимать энергию из окружающей среды наиболее эффективно. Отличный способ для этого – делать копии самих себя, производя побольше поглотителей энергии. Существует много известных примеров такой эмергентной саморепликации: например, вихри в турбулентном потоке порождают копии самих себя, и кластер микросфер может спровоцировать формирование таких же кластеров в соседних сферах. В какой-то момент частицы приноравливаются настолько хорошо копировать себя, что процесс может продолжаться практически бесконечно, добывая энергию и вещество из окружающей среды. Такой способ упорядочения частиц мы называем *жизнью*. Мы все еще очень мало знаем о происхождении жизни на Земле, но мы знаем, что примитивные формы жизни существовали уже 4 миллиарда лет назад.

Если жизнь создает копии самой себя и эти копии снова поступают так же, тогда общее количество их будет продолжать удваиваться с равными временными интервалами, пока популяция не разрастется до такой степени, что столкнется с нехваткой ресурсов или какой-то другой проблемой. Повторяющееся удвоение быстро доходит до больших чисел: если начать с одного и удваивать его всего лишь триста раз, вы получите число, превосходящее количество частиц в нашей Вселенной. Это означает, что вскоре после появления первой примитивной формы жизни колоссальное количество вещества стало живым. Не всегда копирование происходило успешно, так что вскоре появилось много разных видов жизни, старающихся скопировать себя, борющихся за один и тот же ограниченный ресурс. Началась дарвиновская эволюция.

Если бы вы наблюдали за Землей в период появления на ней жизни, вы бы заметили существенные изменения в целенаправленном поведении частиц. Так, например, раньше частицы как будто старались увеличить хаос всеми возможными средствами, теперь же эти вездесущие, копирующие себя частицы преследовали иную цель: не диссипация, а *репликация*. Чарльз Дарвин элегантно объяснил причину: раз именно то, что удачнее всего другого скопировалось, обгоняет всех прочих и доминирует над ними, пройдет немного времени, и любая случайно возникшая форма жизни будет выглядеть так, словно ее оптимизация и была целью всех предшествующих репликаций.

Как же могла цель поменяться с диссипации на репликацию, если законы физики остались неизменными? Ответ в том, что цель, лежащая в основе (диссипация), не изменилась, но привела к другой цели, служащей инструментом на пути к достижению основной цели. Возьмем, к примеру, утоление голода. Кажется, что у каждого из нас есть цель – утолить свой аппетит, хотя мы и знаем, что основной задачей эволюции является репликация, а не пережевывание пищи. Это потому, что еда помогает репликации: если морить себя голодом, останешься без детей. Таким же образом репликация ведет к диссипации, так как планета, на которой жизнь кишмя кишит, поглощает больше энергии. Так что, в некотором роде, космос создал жизнь, чтобы помочь себе быстрее достичь тепловой смерти. Если вы рассыпали сахар на кухне, он может годами там оставаться, сохраняя свою полезную химическую энергию, но если появятся муравьи, они уничтожат эту энергию за считанные минуты. Точно так же запасы нефти в недрах Земли сохраняли бы свою полезную химическую энергию намного дольше, если бы двуногая форма жизни не выкачивала их и не сжигала.

Среди современных высокоразвитых обитателей Земли эти вспомогательные цели, кажется, зажили собственной жизнью: хотя эволюция оптимизировала их до такой степени, что оставила им единственную цель – саморепликацию, многие тратят большую часть своего времени не на производство потомства, а на сон, потребление пищи, строительство домов, защиту своих прав, войны или помощь другим – зачастую в ущерб

размножению. Исследования по эволюционной психологии, экономике и искусственному интеллекту прекрасно объяснили, почему так происходит. Некоторые экономисты строили модели, в которых люди выступали в роли рациональных агентов, этих идеализированных мастеров принимать правильные решения и всегда выбирать оптимальный путь к цели. Очевидно, что такие модели не реалистичны. На практике этим агентам свойственно то, что нобелевский лауреат и автор важных ранних работ по искусственному интеллекту Герберт Саймон назвал “рациональностью, ограниченной недостаточностью ресурсов”: рациональность их решений ограничена неполной информацией, недостатком времени для принятия правильного решения и несовершенством “харда”, помогающего им думать. Это означает, что в то время как дарвиновская эволюция оптимизирует организм для достижения цели, самое лучшее, что она может сделать, – это применить примерный алгоритм, который работает достаточно хорошо, в ограниченном контексте, в котором агент вероятнее всего окажется. Эволюция оптимизировала размножение именно таким способом: вместо того, чтобы спрашивать в каждой ситуации, какое действие даст наибольшее количество успешных отпрысков, она применяет солянку из эвристических ударов: проверенные правила, которые обычно хорошо работают. Для большинства животных это включает сексуальную потребность, утоление жажды по необходимости, утоление голода и стремление избегать всего того, что плохо на вкус или причиняет боль.

Эти проверенные правила с треском проваливаются в ситуациях, на которые они не были рассчитаны: например, когда крысы едят вкусный крысиный яд, когда мухи влипают в липкие ловушки, источающие прекрасный аромат, и когда жуки летят на свет свечи^[52]. Так как сегодняшнее человеческое общество очень отличается от того, под которое подгонялись эволюцией проверенные правила для окружающей среды, мы не должны удивляться, что наше поведение сплошь и рядом не приводит к появлению максимального количества младенцев. Например, вспомогательная цель – удовлетворение голода – зачастую провоцирует людей на потребление избытка калорий, из-за чего у них развивается ожирение с последующими сложностями в личной жизни. Вспомогательная цель – размножаться – выродилась в желание просто заниматься сексом, а не раздавать сперму / яйцеклетки, несмотря на то, что последнее произвело бы большее количество детей с наименьшими усилиями.

Психология: Преследование и восстание против целей

В общих чертах живой организм является агентом ограниченной рациональности: он не преследует единственную цель, а вместо этого следует нескольким основным правилам, которые подсказывают, к чему стремиться и чего избегать. Наш человеческий разум воспринимает эти внутренние основные правила как чувства, которые зачастую (и во многом без нашего осознания) приводят наши решения к основной цели – репликации. Чувство голода и жажды предохраняет нас от голодной смерти и обезвоживания, чувство боли предохраняет нас от повреждения наших тел, вожделение заставляет нас размножаться, чувство любви и сострадания заставляют нас помогать другим и так далее. Ведомый этими чувствами, наш мозг быстро и эффективно решает, что делать, не анализируя подробно, какое именно воздействие это решение будет иметь на количество потомства. Для более подробного разбора чувств и их психологических причин я очень рекомендую работы Уильяма Джеймса и Антонио Дамасио^[90].

Важно отметить, что если наши чувства периодически срабатывают против деторождения, это не обязательно происходит случайно или потому, что нас можно обмануть: наш мозг может протестовать против наших генов и их нацеленности на репликацию вполне осознанно – например, принимая решение использовать контрацептивы. Более экстремальным примером, когда наш мозг восстает против своей природы, может быть решение покончить с собой или провести жизнь, соблюдая целибат, став священником, монахом или монахиней.

Почему мы иногда принимаем решение противиться своей природе и ее целям по репликации? Потому что изначально, как агенты ограниченной рациональности, мы лояльны только по отношению к своим чувствам. И хотя наш мозг развился исключительно для того, чтобы помочь нашим генам копировать себя, ему абсолютно не интересна эта цель, раз у нас нет никаких чувств по отношению к нашим генам: в самом деле, на протяжении почти всей истории человечества наши предки даже и не знали о существовании этих самых генов. Более того, наш мозг намного умнее наших генов, и теперь, когда мы понимаем их цель (копирование себя), мы считаем это довольно банальным и легко игнорируем. Люди могут понимать, почему их гены заставляют их чувствовать вожделение, и все равно имеют мало желания воспитывать пятнадцать детей, а потому предпочитают “хакнуть” свою генетическую программу, совмещая эмоциональное удовольствие от интима с контролем за рождаемостью. Они могут понимать, почему их гены заставляют их желать сладкого, но не хотят поправляться, и поэтому нарушают генетическую программу, получая эмоциональное удовольствие от сладкого напитка с нулевым содержанием калорий благодаря искусственным подсластителям.

Хотя иногда такие методы взлома системы получения удовольствия работают криво – например, когда люди подсаживаются на героин, – наш человеческий набор генов до сих пор прекрасно выживает, несмотря на наш хитрый и протестующий мозг. Важно, однако, помнить, что внутренним авторитетом являются наши чувства, а не наши гены. Это означает, что человеческое поведение не в полной мере подстроено под выживание вида. На самом деле, раз наши чувства применяют исключительно основные правила, которые подходят не для всех ситуаций, человеческое поведение, строго говоря, совсем не имеет единой, четко определенной цели.

Инженерия: привлечение сторонних целей

Могут ли быть цели у машин? Этот простой вопрос вызвал горячие споры, потому что люди вкладывают в него слишком разные смыслы, касающиеся весьма колючих тем – например, таких: могут ли машины думать? могут ли у них быть чувства? А если сформулировать вопрос проще: “Могут ли машины демонстрировать целенаправленное поведение?”, – тогда ответ становится очевидным: “Конечно могут, так как мы можем их такими разработать!”. Мы можем разработать мышеловку, которая будет иметь цель ловить мышей, посудомоечную машину, у которой будет цель мыть посуду, или часы, целью которых будет показывать время. Когда сталкиваешься с машиной, фактически все, что вас волнует, – это ее очевидная цель: если за вами будет гнаться самонаводящийся по тепловому излучению снаряд, вам в принципе будет все равно, есть ли у него сознание и чувства! Если вам все еще неловко говорить, что у снаряда есть цель, даже если она не осознанная, на данный момент можно просто вместо написанного мною слова “цель” читать “предназначение” – мы вернемся к вопросу об осознании в следующей главе.

Все нами до сих пор построенное демонстрирует только целенаправленный проект, но не целенаправленное поведение: скоростное шоссе никак себя не ведет, оно просто есть. Однако наиболее экономичным объяснением его существования будет такое: оно спроектировано, чтобы достичь некой цели, так что даже такая пассивная технология делает нашу Вселенную более целенаправленной. *Телеология* предлагает для всякой вещи объяснение с точки зрения ее предназначения, а не ее причин, так что мы можем подытожить первую половину этой главы, сказав, что наша Вселенная становится более телеологической.

У неживой материи не только *могут* быть цели, по крайней мере в этом слабом смысле, но они определенно у нее *есть*. Если наблюдать за атомами Земли с момента образования нашей планеты, можно было бы увидеть три стадии целенаправленного поведения.

1. Все вещество занято исключительно диссипацией (энтропия растет).
2. Часть вещества стала живой, ее фокус сместился на репликацию и вспомогательные цели.
3. Живые организмы подвергли быстрорастущую часть вещества переупорядочению, чтобы это помогало им в достижении их целей.

Табл. 7.1 показывает доминирование человечества в физическом отношении: на нас теперь не только приходится больше вещества, чем на всех остальных млекопитающих, за исключением коров (которые столь многочисленны потому, что используются для достижения наших целей – потребления говядины и молочных продуктов), но и количество вещества, использованного для наших машин, дорог, зданий и других разработанных проектов, скоро, вероятно, превысит массу всех живых существ на Земле. Другими словами, даже без интеллектуального взрыва наибольшее количество вещества на Земле, демонстрирующего целенаправленные свойства, может в скором времени оказаться созданным искусственно, а не быть результатом эволюционного развития.

Этот новый, третий вид целенаправленного поведения может стать намного более разнообразным, чем все предшествующие: в то время как у всякой сущности, возникшей в результате эволюции, есть только одна основная цель (репликация), у искусственно созданных сущностей могут быть любые основные цели, даже противоположные друг другу. Печь нужна, чтобы нагреть пищу, а холодильник – чтобы охладить. Электрогенератор превращает механическое движение в электричество, а электромотор превращает электричество в механическое движение. Стандартная программа по игре в шахматы старается выиграть, но есть и такие, которые участвуют в соревнованиях по поддавкам.

Таблица 7.1.

Примерное количество вещества на Земле в сущностях, которые возникли эволюционно или созданы искусственно для определенных целей. Искусственно созданные сущности, такие как здания, дороги и машины, похоже, скоро перегонят по массе сущности, возникшие в результате эволюции, такие как растения и животные.

Целенаправленные сущности	Млрд тонн
5×10^{30} бактерий	400
Растения	400
10^{15} рыб	10
$1,3 \times 10^9$ коров	0,5
7×10^9 людей	0,4
10^{14} муравьев	0,3
$1,7 \times 10^6$ китов	0,0005
Бетон	100
Сталь	20
Асфальт	15
$1,2 \times 10^9$ машин	2

Общий исторический тренд таков, что цели искусственно созданных сущностей не только более разнообразны, но и более сложны: наши устройства становятся умнее. У первых искусственно создаваемых нами объектов были простые цели – например, дома строились для того, чтобы нам было тепло, сухо и безопасно. Мы постепенно научились строить объекты с целями посложнее – такие, как роботизированные пылесосы, самонаводящиеся ракеты или беспилотные автомобили. Недавний прогресс в разработке искусственного интеллекта дал нам такие системы, как Deep Blue, Watson и AlphaGo, цели которых – выигрывать в шахматы, в викторинах или в го – настолько изощренны, что надо быть весьма просвещенным человеком, чтобы по достоинству оценить, насколько успешны эти системы в их достижении.

Когда мы строим машину, которая должна помогать нам, бывает сложно в полной мере привести ее цели в соответствие с нашими. Например, мышеловка может перепутать палец на нашей ноге с голодным грызуном, и это приведет к довольно болезненным последствиям. Любая машина – агент ограниченной рациональности, и даже самые сложные современные машины не так хорошо понимают мир, как мы, поэтому правила, которые они используют для того, чтобы разобраться, что к чему, зачастую довольно просты. Мышеловка срабатывает с такой неуместной поспешностью, потому что совсем не представляет себе, что такое мыш, много смертей на производствах происходит из-за того, что машины совсем не представляют себе, что такое человек, и компьютеры, которые привели к мгновенному обвалу акций на Уолл-стрит в 2010 году, ни малейшего представления не имели о смысле совершаемых ими действий. Многие из таких проблем с приведением целей в соответствие могут быть решены, если машины станут умнее, но, как мы знаем по опыту Прометея из четвертой главы, даже очень умные машины могут

задать непростую задачу, которую необходимо решить, чтобы привести цели машины в соответствие с нашими.

Дружественный искусственный интеллект: приведение целей в соответствие

Чем умнее и мощнее становится машина, тем важнее, чтобы ее цели не вступали в противоречие с нашими. Пока машины, которые мы строим, немного туповаты, вопрос не в том, окажутся ли человеческие цели превалирующими в итоге, а только в том, много ли хлопот эти машины доставят человечеству, прежде чем мы приведем в соответствие их цели и наши. Но если сверхразум когда-нибудь появится, роли поменяются: так как разум – это способность достигать поставленных целей, то искусственный сверхинтеллект по определению намного лучше добивается своих, чем люди своих, а значит – в итоге превалирующими будут его цели, а не наши. Образ Прометея дал нам возможность исследовать много примеров тому в главе 4. Если вы хотите прямо сейчас почувствовать, каково это, когда машина встает у вас на пути, просто скачайте ультрасовременный симулятор шахмат и попробуйте с ним сразиться. У вас никогда не получится победить, а ему на смену быстро приходят другие...

Другими словами, подлинная опасность искусственного интеллекта *не в его злонамеренности, а в его изощренности*. Сверхразумный искусственный интеллект будет прекрасно добиваться своих целей, и, если его цели будут противоречить нашим, мы окажемся в затруднительном положении. Как я говорил в главе 1, людей мало заботят муравейники, которые могут оказаться в зоне затопления при строительстве плотины для гидроэлектростанции, так давайте не допустим, чтобы на месте муравьев оказалось человечество. Большинство исследователей считают, что, если мы когда-нибудь создадим сверхразум, мы должны будем позаботиться о том, чтобы он оказался “дружественным”, по выражению одного из создателей теоретического подхода к вопросам безопасности искусственного интеллекта Элиезера Юдковски, то есть чтобы его цели не противоречили нашим^[91].

Понять, как привести цели сверхразумного искусственного интеллекта в соответствие с нашими, не только важно, но и сложно. На самом деле, в настоящий момент это нерешенная проблема. Она разделяется на три тяжелые подпроблемы, каждая из которых – предмет активного изучения как специалистов по информатике, так и ученых других специальностей. Эти подпроблемы состоят в том, чтобы:

- 1) искусственный интеллект *понял* наши цели;
- 2) искусственный интеллект *принял* наши цели;
- 3) искусственный интеллект *придерживался* наших целей.

Давайте разберем каждую из подпроблем, отложив вопрос о том, что имеется в виду под “нашими целями” до следующего раздела.

Чтобы понять наши цели, искусственный интеллект должен разобраться не в том, что мы делаем, а в том, почему мы это делаем. Для нас, людей, это так просто, что мы часто забываем, как трудно это объяснить компьютеру и как просто истолковать наши намерения превратно. Если вы попросите беспилотный автомобиль будущего довезти вас до аэропорта “как можно быстрее” и будете поняты буквально, то в аэропорт вы попадете покрытым рвотой и преследуемым вертолетами. Если вы воскликнете: “Это совсем не то, что я хотел!”, в ответ можете вполне обоснованно услышать: “Это то, о чем вы просили”. Эта тема не новая, и не раз она уже возникала в истории. В древнегреческой легенде царь Мидас захотел, чтобы все, к чему он прикасается, превращалось в золото, и был расстроен, когда это привело к тому, что он не смог есть и, что еще страшнее, превратил свою дочь в золотую статую. В историях, в которых джин исполняет три желания, есть много вариаций первых двух, но третье желание почти всегда одно и то же: “Пожалуйста, отмени два предыдущих, потому что это совсем не то, чего я на самом деле хотел”.

Все эти примеры показывают: чтобы понять, чего люди на самом деле хотят, нельзя просто следовать тому, что они говорят. Надо также иметь довольно подробную модель мира, которая включала бы в себя некоторые общие установки, о которых мы обычно не говорим, так как считаем их очевидными – например, о том, что нам не нравится, когда тошнит в машине или приходится есть золото. Когда такая модель мира есть, мы в большинстве случаев можем понять, чего люди хотят, даже если они об этом не сообщают, – достаточно просто наблюдать за их целенаправленным поведением. На самом деле, дети научатся большему, наблюдая за поведением своих родителей, чем слушая, что родители им говорят.

Исследователи в области искусственного интеллекта в настоящее время стараются научить машины отличать цели от поведения, и это будет полезным навыком задолго до того, как появится сверхразум. Например, пожилому человеку будет полезно, если ухаживающий за ним робот поймет, что этот человек ценит, просто наблюдая за ним, чтобы ему не пришлось объяснять все словами или программировать его компьютер. Одна сложность состоит в том, чтобы найти хороший способ шифрования произвольной системы целей и этических принципов в компьютере, а другая сложность состоит в том, чтобы сделать такую машину, которая сможет определять, какие именно системы лучше всего соответствуют тому поведению, которое они видят.

Получивший большое распространение в последнее время подход ко второй сложности известен на гиковском сленге как *обучение с обратным подкреплением*, и он находится под пристальным вниманием нового Берклевского исследовательского центра, созданного Стюартом Расселом. Предположим, например, что искусственный интеллект наблюдает за женщиной – членом пожарной команды, которая вбегает в горящее здание и спасает младенца. Машина может предположить, что ее целью было спасение ребенка и что ее этические принципы заставляют ценить жизнь ребенка значительно выше комфортного досуга в пожарной машине – она и в самом деле достаточно высоко оценивает чужую жизнь, чтобы рисковать ради ее спасения собственной безопасностью. Но искусственный интеллект может заключить также, что пожарница замерзла и захотела согреться или что она занимается спортом. Если бы искусственный интеллект впервые сталкивался с пожарами и раньше ничего о пожарниках, пожарах и младенцах не знал, ему было бы трудно понять, какое из двух объяснений правильно. Однако основная идея обучения с обратным подкреплением состоит в том, что мы принимаем решения непрерывно и что каждое решение, которое мы принимаем, что-то говорит о наших целях. Таким образом, есть надежда, что, наблюдая за большим количеством людей в разных ситуациях (настоящих или в кино и литературе), искусственный интеллект сможет со временем построить верную модель наших общих установок.

Даже если создать искусственный интеллект, который сможет понять цели своего владельца, это не означает, что он будет непременно под них подстраиваться. Представьте себе ваших самых нелюбимых политиков: вы знаете, чего они хотят, но это не то, чего хотите вы, и даже несмотря на то, что они очень стараются, им не удастся убедить вас принять их цели.

У нас есть много способов воспитать в наших детях приятие наших целей – одни из них более успешны, другие менее, как я узнал в процессе воспитания двух сыновей-подростков. Когда же надо убеждать компьютер, а не человека, задача, с которой мы сталкиваемся, называется задачей загрузки целей, и она существенно сложнее, чем воспитание морали в детях. Представьте себе постоянно улучшающийся искусственный интеллект, проделавший путь от “ниже человеческого” до “выше человеческого”, сначала при нашем содействии, а потом благодаря рекурсивному самосовершенствованию, как это делал Прометей. Вначале он намного слабее вас и не может помешать вам выключить его и заменить части его программного обеспечения и данные, в которых закодированы ваши цели, но это и не поможет, потому что он все равно еще слишком глуп для того,

чтобы в полной мере понять ваши цели: нужен уровень человека для их понимания. И вот, наконец, он становится намного умнее вас и, вероятно, сможет легко понять ваши цели, но и это может уже не помочь, потому что теперь он гораздо сильнее вас и может не позволить себя выключить и изменить его цели, – так же, как вы не позволяете нелюбимым политикам заменить ваши цели их собственными.

Другими словами, временной промежуток, в течение которого вы можете загрузить цели в искусственный интеллект, может оказаться чересчур коротким – между моментом, когда он слишком глуп, чтобы понять вас, и моментом, когда он уже стал слишком умен, чтобы позволить вам это сделать. Причина, по которой загрузка целей может оказаться сложнее для машин, чем для людей, состоит в том, что их разум может развиваться намного быстрее: если ребенок много лет пребывает в том прекрасном возрасте, когда его разум сопоставим с разумом родителя, то для искусственного интеллекта этот возраст может закончиться через несколько дней, а то и часов, как это было с Прометеем.

Для загрузки целей в машину некоторые ученые предлагают другой подход, который обозначают модным словечком *корректируемость (corrigibility)*. В его основе лежит надежда на то, что примитивному искусственному интеллекту можно задать любую систему целей, потому что вы все равно его время от времени выключаете, а выключив, можете скорректировать и систему целей. Если это окажется возможным, тогда можно спокойно позволить своему искусственному интеллекту становиться сверхразумным, периодически выключая его и меняя ему систему целей, потом проверяя ее и, если результат окажется неудачным, выключая снова, чтобы проделывать новые манипуляции с целями^[92].

Но даже если вы сможете создать такой искусственный интеллект, который поймет и примет ваши цели, проблема соответствия его целей и ваших все еще останется нерешенной: что, если цели вашего искусственного интеллекта будут изменяться с его развитием? Чем вы сможете гарантировать, что он будет сохранять приоритет ваших целей в ходе рекурсивного самосовершенствования? Давайте исследуем один любопытный аргумент, показывающий, что автоматическое сохранение целей гарантировано, и затем посмотрим, найдем ли мы в нем слабые места.

Хотя мы не можем в точности предсказать, что произойдет после интеллектуального взрыва, – именно поэтому Вернор Виндж назвал это сингулярностью, – физик и исследователь искусственного интеллекта Стив Омохундро в бурно обсуждавшемся эссе 2008 года утверждал, что мы, тем не менее, можем предсказать некоторые аспекты поведения сверхразума, которые практически не зависят от его окончательных целей^[93]. Это утверждение было подхвачено и дальше развито в книге Ника Бострёма *Superintelligence*. Основная идея состоит в том, что, каковы бы ни были конечные цели, сопутствующие им вспомогательные цели будут предсказуемыми. Ранее в этой главе мы видели, как цель воспроизведения привела к появлению вспомогательной цели утоления голода. Это означает, что если бы пришелец наблюдал за развитием бактерии на Земле миллиард лет назад, он не смог бы предвидеть, какие цели будут у людей, но мог бы с точностью предсказать, что одной из наших целей будет потребление питательных веществ. Заглядывая вперед, каких вспомогательных целей нам стоит ожидать от сверхразумного искусственного интеллекта?

Я смотрю на это так: для увеличения шансов достичь своих конечных целей, какими бы они ни были, искусственный интеллект должен преследовать вспомогательные цели, представленные на рис. 7.2. Для достижения своих конечных целей он должен стараться не только улучшить свои возможности, но и убедиться, что он сохранит эти цели даже после того, как достигнет более высокой степени развития. Это звучит довольно правдоподобно: в конце концов, согласились бы вы имплантировать себе в мозг бустер, увеличивающий IQ, если бы знали, что он заставит вас желать смерти любимых? Аргумент, что любой быстро развивающийся искусственный интеллект сохранит свои конечные цели, лег краеугольным камнем в представление о дружелюбии,

пропагандируемое Элиезером Юдковски с коллегами: оно говорит нам, что если мы сумеем добиться от самосовершенствующегося искусственного интеллекта дружелюбия через понимание и принятие наших целей, тогда мы в порядке – тем самым будет гарантировано, что он навсегда останется дружелюбным.

Но так ли это на самом деле? Чтобы ответить на этот вопрос, нам необходимо изучить другие вспомогательные цели на рис. 7.2. Очевидно, что искусственный интеллект сможет добиться максимальных шансов на достижение своих конечных целей, какими бы они ни были, если он сможет расширить свои способности, улучшая свой “хард”, свой “софт” [53] и модель мира. То же самое можно сказать и о людях: девочка, чья цель состоит в том, чтобы стать лучшей теннисисткой в мире, должна тренироваться, тем самым улучшая свой теннисно-мускулатурный “хард”, свой нейронный “софт” и ментальную модель мира, которая поможет предсказать, что будет делать ее противник. Для искусственного интеллекта вспомогательная цель оптимизации “харда” подразумевает и как более качественное использование имеющихся ресурсов (сенсоров, преобразователей, вычислителей и т. д.), так и потребление большего количества ресурсов. Это также относится и к потребности самозащиты, так как разрушение / отключение будет пагубно отражаться на “харде”.

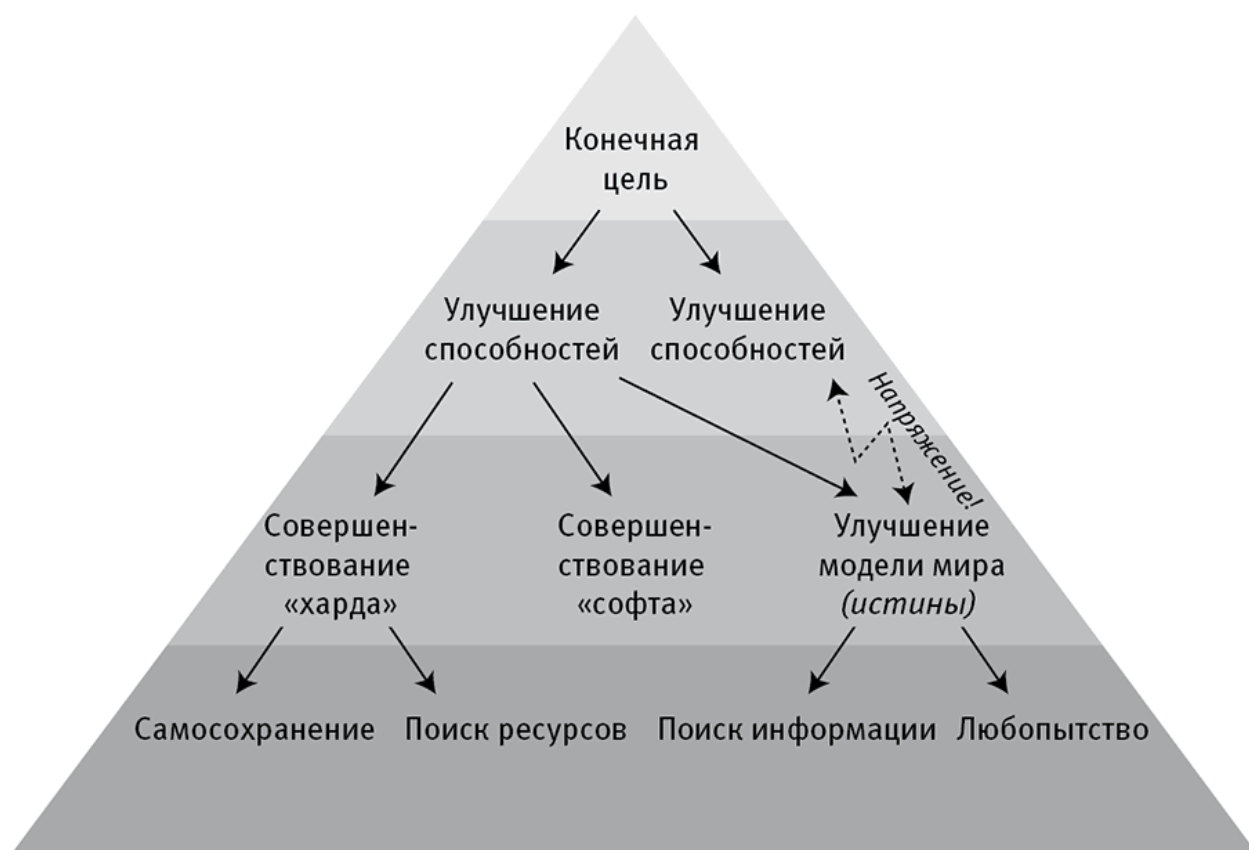


Рис. 7.2

Любая конечная цель сверхразумного искусственного интеллекта естественно приводит к возникновению вспомогательных целей, показанных на рисунке. Но между совершенствованием своих способностей и сохранением изначальных целей есть внутреннее противоречие, которое заставляет нас сомневаться, что искусственный интеллект будет сохранять исходную цель, становясь все более разумным.

Но секундочку! Не попали ли мы в ловушку и не стали ли наделять наш искусственный интеллект человеческими качествами, рассуждая о том, как он будет стараться приумножать ресурсы и защищать себя? Не должны ли мы ожидать такого стереотипного поведения альфа-самца только от разума, взрослого в жестокой

конкуренции дарвиновской эволюции? Раз системы с искусственным интеллектом – продукт искусственного конструирования, а не естественной эволюции, не будут ли они менее амбициозными и более склонными к самопожертвованию?

В качестве простого примера давайте рассмотрим рис. 7.3, изображающий робота с искусственным интеллектом, чья единственная цель – спасение наибольшего количества овец от большого злого волка. Это звучит очень благородно и альтруистично и никак не связано с самосохранением и потреблением. Но какой же будет оптимальная стратегия для нашего друга робота? Робот больше не сможет спасти овец, если он подорвется на mine, поэтому у него есть стимул оставаться целым. Другими словами, у него появляется вспомогательная цель – самосохранение! Ему также важно быть любопытным, улучшая свою модель мира, исследуя окружающее, потому что, хотя путь, по которому он сейчас идет, в конце концов приведет к пастбищу, есть и более короткий маршрут, который оставит волкам меньше времени на съедание овец. В конце концов, если робот тщательно все изучит, он поймет и важность потребления ресурсов: энергетический напиток даст ему возможность бежать быстрее, а пистолет позволит стрелять в волка. В итоге мы не можем сказать, что развитие вспомогательных целей “альфа-самца”, таких как самосохранение и захват ресурсов, свойственно только эволюционировавшим организмам, потому что наш интеллектуальный робот развил их, имея одну-единственную цель – овечье счастье.

Score 2

Level 1

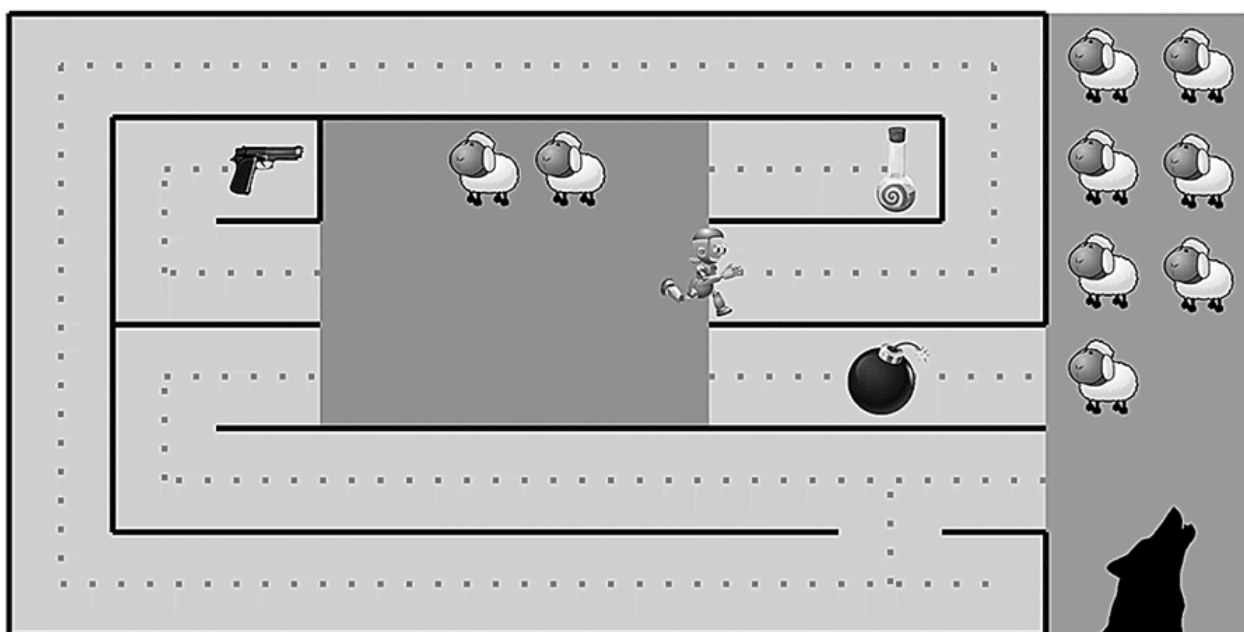


Рис. 7.3

Даже если основной целью робота является получение наивысшего балла за доставку овец от пастбища до загона до того, как волки их съедят, то уже и в этом случае у него возникнут некоторые вспомогательные цели, включающие в себя самосохранение (не подорваться на бомбе), исследования (находить более короткие пути) и потребление ресурсов (энергетический напиток позволит ему бежать быстрее, а пистолет – стрелять в волков).

Если вы дадите сверхразуму одну-единственную цель – самоуничтожение, он, конечно, с радостью это выполнит. Однако фокус весь в том, что он воспротивится отключению, если вы дадите ему любую другую цель, которая подразумевает, что он должен быть включенным для ее выполнения, – а это относится практически ко всем целям! Если вы дадите сверхразуму единственную цель – например, минимизировать вред, наносимый человечеству, он будет защищать себя от выключения, так как знает, что

мы навредим друг другу намного сильнее в его отсутствие во время будущих войн и других безобразий.

Аналогично практически любой цели легче достичь, располагая бо2льшим количеством ресурсов, поэтому логично ожидать от сверхразума, что он будет стремиться завладеть ресурсами практически вне зависимости от того, какова его конечная цель. Таким образом, если перед сверхразумом поставить единственную цель, не ограничивая его во времени, то это может оказаться опасным: сверхразум, созданный с единственной целью – совершенствоваться в игре в го, со временем придет к рациональному решению реорганизовать Солнечную систему в гигантский компьютер, без учета интересов ее обитателей, а потом начать организовывать наш космос для достижения еще большей компьютерной мощи. Теперь мы совершили полный круг: так же как цель захвата ресурсов приводила людей к вспомогательной цели совершенствоваться в го, так и цель совершенствоваться в го может привести к вспомогательной цели захвата ресурсов. Можно заключить, что из-за возникновения вспомогательных целей для нас принципиально важно не делать шага к появлению сверхразума прежде, чем будет решена проблема приведения его целей в соответствие с нашими: пока мы не обеспечим дружелюбности его целей, дело, вероятно, повернется для нас скверно.

Теперь мы готовы к тому, чтобы рассмотреть третью, и самую сложную, часть проблемы по приведению целей в соответствие: если мы преуспеем в том, чтобы самообучающийся сверхразум и узнал о наших целях, и принял их, будет ли он и дальше придерживаться их, как утверждает Омохундро? Каковы доказательства?

Человеческий интеллект особенно быстро развивается во время взросления, но это не означает, что он непременно сохраняет свои детские цели. Напротив, люди часто кардинально меняют свои цели по мере того, как они познают мир и становятся мудрее. Сколько взрослых людей вы знаете, которых мотивирует просмотр *Телепузиков*? Нет доказательств того, что такая эволюция целей останавливается после преодоления какого-то интеллектуального порога, – на самом деле, могут быть даже признаки того, что склонность к изменению целей в результате получения нового опыта и знаний с развитием интеллекта растет, а не сокращается.

Почему так происходит? Подумайте еще раз об упомянутой выше вспомогательной цели построения лучшей модели мира – там-то и лежит камень преткновения! Моделирование мира и сохранение целей нелегко уживаются друг с другом (см. рис. 7.2). С развитием интеллекта может прийти не просто количественное нарастание возможностей добиваться имеющихся целей, но и качественно новое понимание природы реальности, при этом может выясниться, что старые цели никчемны, бессмысленны или даже не определены. Например, представим, что мы запрограммировали дружелюбный искусственный интеллект увеличивать количество людей, чьи души попадут в рай после смерти. Для начала он постарается привить людям сострадание к ближнему и желание почаще посещать церковь. Но представим, что потом у него появится вполне научно обоснованное понимание людей и человеческого сознания, и он, к своему большому удивлению, узнает, что души нет. Что теперь? Существует не меньшая вероятность, что любая другая цель, которую мы зададим искусственному интеллекту, основываясь на нашем текущем понимании мира (хотя бы такая, как “увеличивать значимость человеческой жизни”), может со временем оказаться не определенной, как установит искусственный интеллект.

Более того, в своем стремлении создать лучшую модель мира искусственный интеллект может – естественно, как это делали мы, люди, – попытаться смоделировать себя и понять, как функционирует он сам, – иными словами, погрузиться в рефлекссию. Как только он построит хорошую модель самого себя и поймет, что он такое, он поймет и то, что его цели были даны нами на мета-уровне, и, возможно, предпочтет избегать их или отказываться от них, точно так же, как люди понимают и осознанно отказываются от целей, заложенных на генетическом уровне, как в примере с использованием

контрацепции. Мы уже рассмотрели в блоке про психологию, почему мы предпочитаем обманывать наши гены и подрывать их цели: потому что мы по-настоящему верны только той ерунде, которая вызывает у нас эмоциональный отклик, а не генетическим целям, которые этот отклик провоцируют, – что мы теперь понимаем и считаем достаточно банальным. Поэтому мы предпочитаем взламывать наш механизм поощрения, обнаруживая в нем слабые места. Аналогично цель по защите интересов человека, которую мы запрограммируем в нашем дружественном искусственном интеллекте, станет геном машины. Как только этот дружественный искусственный интеллект осознает себя достаточно хорошо, он может посчитать эту цель банальной или нецелесообразной, подобно тому как мы относимся к неконтролируемому размножению, и пока неочевидно, насколько просто или сложно ему будет найти слабые места в нашем программировании и подорвать свои внутренние цели.

Представим, для примера, группу муравьев, которая создает вас как своего постоянно самосовершенствующегося робота, который намного умнее их, который разделяет их цели и помогает им строить лучшие и большие муравейники, а вы постепенно развиваете уровень своего интеллекта и способность соображать до человеческих, какие они у вас сейчас. Думаете ли вы, что проведете остаток своих дней, оптимизируя муравейники, или у вас появится интерес к более увлекательным вопросам и занятиям, которых муравьям уже не понять? Если так, думаете ли вы, что найдете способ не принимать во внимание жажду к защите муравьев, которую ваши создатели заложили в вас, практически так же, как вы игнорируете те позывы, которые заложили в вас гены? И в этом случае возможно ли, чтобы сверхразвитый дружелюбный искусственный интеллект воспринимал наши человеческие цели недостаточно вдохновляющими и безвкусными, как и вы в случае с целями муравьев, и развивал новые цели, отличные от тех, которым мы его обучали и которые он от нас перенял?

Возможно, есть способ разработать самосовершенствующийся искусственный интеллект, который гарантировал бы пожизненное сохранение дружественных целей по отношению к людям, однако, мне кажется, справедливо будет сказать, что мы пока не знаем, как его построить и даже возможно ли это. В заключение: проблема приведения целей AI в соответствие с человеческими состоит из трех частей, ни одна из которых не решена на данный момент, и все их сейчас активно исследуют. Раз они настолько трудны, следует начать уделять им пристальное внимание сейчас, задолго до того, как сверхинтеллект будет разработан, чтобы убедиться, что у нас будут ответы к тому моменту, когда они нам понадобятся.

Этика: выбор целей

Мы уже исследовали вопрос о том, как может машина понять наши цели, принять их и придерживаться их как своих собственных. Но кто такие “мы”? О чьих целях мы, собственно, говорим? Будет ли это один человек или группа людей, кто станет выбирать цели для будущего сверхразума, и это при том, что существует колоссальная пропасть между целями Адольфа Гитлера, папы Франциска и Карла Сагана? Или существуют какие-то цели, которые могут считаться хорошим компромиссом для всего человечества?

По моему мнению, эта этическая проблема вместе с проблемой о приведении целей в соответствие являются принципиальными, и они должны быть решены до появления сверхинтеллекта. С одной стороны, откладывать работу над этическими проблемами до того момента, как будет создан сверхинтеллект с согласованными целями, безответственно и чревато ужасными последствиями. Безупречно послушный сверхинтеллект, чьи цели автоматически приводятся в соответствие с целями его владельца-человека, будет похож на оберштурмбаннфюрера нацистского СС Адольфа Эйхмана на стероидах: при отсутствии морального компаса или воспитания как такового он будет достигать целей своего хозяина с безжалостной устремленностью, какими бы они ни были^[94]. С другой стороны, только если мы решим проблему с приведением целей в соответствие, мы сможем насладиться роскошью спора о том, какие цели выбирать. Теперь давайте погрузимся в эту роскошь.

С незапамятных времен философы мечтали о создании этики (принципов, которые указывают нам, как себя вести) с нуля, с использованием только неоспоримых принципов и логики. Увы, тысячу лет спустя единственный консенсус, к которому мы смогли прийти, – это отсутствие консенсуса. Например, пока Аристотель придавал особое значение добродетели, Иммануил Кант делал акцент на долге, а прагматики – на огромном счастье для большинства. Из исходных принципов, которые он называл “категорическими императивами”, Кант извлек выводы, с которыми не согласны многие современные философы: что мастурбация хуже самоубийства, что гомосексуализм отвратителен, что ублюдка можно убить и что женами, слугами и детьми владеют так же, как вещами.

С другой стороны, несмотря на эти разногласия, есть много этических принципов, с которыми согласится большинство – и в разных культурах, и в разные века. Например, почитание красоты, великодушия и истины уходит корнями к Бхагавадгите и Платону. Институт перспективных исследований Принстона, где я когда-то был аспирантом, имел кредо “Истина и Красота”, в то время как Гарвардский университет опустил эстетическую составляющую и провозглашал просто “Veritas”, то есть истину. В книге *A Beautiful Question*^[54] мой коллега Фрэнк Вильчек утверждает, что истина связана с красотой и что мы можем рассматривать нашу Вселенную как произведение искусства. Наука, религия и психология – все вдохновлены истиной. Религия ставит сильный акцент на доброте, как и мой университет, MIT: в своей приветственной речи в начале 2015 года наш президент Рафаэль Рейф подчеркнул, что наша миссия состоит в том, чтобы делать мир лучше.

Несмотря на то что попытки прийти к консенсусу по этическим вопросам до сего момента оказывались неудачными, есть распространенное мнение, что некоторые этические принципы происходят от других, более фундаментальных, как вспомогательные цели от конечных целей. Например, поиск истины можно рассматривать в связи со стремлением к лучшей модели мира (см. рис. 7.2), понимание окончательной реальности приводит нас к новым этическим целям. В самом деле, у нас теперь есть прекрасная канва, в которую вписывается поиск истины, – научный метод. Но как мы сможем определить, что красиво или хорошо? Некоторые аспекты красоты также можно проследить до лежащих в основе целей. Например, наше понимание мужской и женской красоты может частично отражать наше неосознанное суждение о том, насколько данный объект подходит для воспроизведения наших генов.

Что касается доброты, так называемые золотые правила (что относиться к другим надо так, как хочешь, чтобы относились к тебе) присутствуют в большинстве культур и религий и, очевидно, призваны к развитию гармоничного человеческого общества (а следовательно, и генов) через усиление сотрудничества и отказ от непродуктивного соревнования⁽⁹⁵⁾. То же самое можно сказать о более конкретных этических правилах, которые были закреплены в законодательных системах по всему миру, – таких, как конфуцианская важность честности и многие из десяти заповедей, включая “не убий”. Другими словами, многие этические принципы имеют сходство с социальными эмоциями, такими как эмпатия и сострадание: они развивают сотрудничество и влияют на наше поведение через поощрение и наказание. Если мы сделаем что-то жестокое и будем потом об этом сожалеть, наше эмоциональное наказание измеряется химией нашего мозга. А общество может наказать нас за нарушение этических принципов косвенным путем – например, неформальным порицанием со стороны близких или официальным обвинением в нарушении законов.

Другими словами, хотя сегодняшнее общество довольно далеко от этического консенсуса, существует много базовых принципов, с которыми почти все согласны. Это неудивительно, потому что человеческое общество, дожившее доныне, стремится к тому, чтобы его этические принципы оптимально вели к главной цели: способствовали выживанию и процветанию. Заглядывая в будущее, где у жизни появляется шанс распространиться в космосе и существовать там на протяжении миллиардов лет, мы должны решить: какой минимальный набор этических принципов подойдет для нас в таком будущем? Обсуждать их должны мы все. Я знакомился с этическими воззрениями самых разных мыслителей на протяжении многих лет, эта тема меня увлекала, и я смог выделить из всего разнообразия их взглядов четыре принципа:

1) Утилитаризм: положительные чувственные переживания следует приумножать, а страдания надо свести к минимуму;

2) Диверсификация: разнообразие положительных переживаний лучше, чем частое повторение одного и того же переживания, даже если последнее было определено как самое положительное из возможных;

3) Автономность: обладающие сознанием существа / общества должны иметь свободу преследовать свои собственные цели, если только это не противоречит более важным принципам;

4) Преемственность: совместимость со сценариями, которые большинство людей *сегодня* оценивают как счастливые; несовместимость со сценариями, которые *сегодня* почти все люди рассматривают как ужасные.

Давайте уделим этим четырем принципам некоторое внимание. Традиционно под утилитаризмом понимается “максимальное счастье для максимального количества людей”, но я обобщил, постаравшись избавить его от излишнего антропоцентризма, чтобы он еще включал других животных, искусственный интеллект, каким он существует сегодня, и другие виды искусственного интеллекта, которые могут появиться в будущем. Я даю определение на языке переживаний, удалив из него упоминания о людях или вещах, так как большинство мыслителей согласны, что красота, радость, удовольствие и страдание – это субъективные переживания. Это означает, что если нет переживания (как в мертвой или населенной зомби-подобными бездушными машинами Вселенной), то не может быть и смысла или чего бы то ни было другого, этически значимого. Если мы согласимся с утилитарным этическим принципом, тогда становится важно, чтобы мы определили, какие разумные системы обладают сознанием (то есть способны на субъективные переживания), а какие нет. Это тема следующей главы.

Если бы нас заботил только утилитарный принцип, то нам было бы достаточно выяснить, что именно вызывает самое положительное из возможных переживаний, и настроить наш космос так, чтобы именно оно повторялось до бесконечности из раза

в раз, если это возможно, в максимально возможном числе галактик – и используя симуляцию, если она окажется наиболее эффективной. Если вы чувствуете, что такой способ употребления нашей космической данности слишком банален, то, подозреваю, главным образом из-за того, что вам в этом сценарии не хватает разнообразия. Что бы вы почувствовали, если бы до конца жизни на завтрак, обед и ужин вам предлагали одно и то же? Или всю жизнь крутили бы только одно кино? Или у всех ваших друзей была одинаковая внешность, одинаковый характер и мысли? Возможно, нам нравится разнообразие не в последнюю очередь потому, что именно оно помогло человечеству выживать и процветать, сделало нас более стойкими. Возможно, это также связано с предустановками для разума: его развитие на протяжении 13,8 миллиардов лет космической истории сопровождалось превращением скучного единообразия в более диверсифицированную, дифференцированную и сложную структуру, которая обрабатывает информацию более сложными способами.

Принцип автономности лежит в основе многих прав и свобод, зафиксированных во Всеобщей декларации прав человека, которую ООН приняла в 1948 году в попытке усвоить уроки двух мировых войн. Сюда включены свобода мысли, слова и перемещения, свобода от рабства и истязаний, право на жизнь, неприкосновенность, безопасность, образование, право вступать в брак, работать и владеть собственностью. Если мы хотим, чтобы наше определение было менее антропоцентричным, мы можем дать его в более общем виде, ограничившись свободой думать, учиться, общаться, владеть собственностью и быть избавленными от причинения ущерба, а также свободой делать то, что мы хотим, пока это не нарушает свобод окружающих. Принцип автономности способствует диверсификации до тех пор, пока у всех хоть немного различаются цели. Кроме того, принцип автономности следует из принципа утилитарности, если у отдельных существ есть цель получать положительный опыт и они стараются действовать в своих интересах: если бы вместо этого мы хотели запретить какой-нибудь сущности достигать ее целей, даже если они никому не вредят, положительных переживаний стало бы меньше. В самом деле, это аргумент в пользу автономии – тот же самый, который экономисты приводят в пользу свободного рынка: он способствует возникновению равновесной ситуации (экономисты называют ее “парето-эффективным распределением благ”), где никому не может стать лучше без того, чтобы кому-то другому стало хуже.

Принцип преемственности говорит, что у нас должно быть право голоса в определении будущего, раз уж мы помогаем его строить. Принципы автономности и преемственности утверждают демократические идеалы: первый дает будущим формам жизни власть над тем, как будет использоваться наше космическое достояние, а последний дает такую власть ныне живущим людям.

Несмотря на то, что на первый взгляд никакого противоречия между этими четырьмя принципами не видно, внедрение их на практике проблематично – дьявол, как всегда, в деталях. Сложность кроется в реминисценциях по известным “трем законам робототехники” легендарного писателя-фантаста Айзека Азимова:

- 1) Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред;
- 2) Робот должен повиноваться всем приказам, которые дает человек, кроме тех случаев, когда эти приказы противоречат Первому Закону;
- 3) Робот должен заботиться о своей безопасности в той мере, в какой это не противоречит Первому и Второму Законам^[55].

Хотя все это звучит хорошо, многие истории Азимова показывают, как законы могут привести к противоречиям и неожиданным ситуациям. Теперь предположим, что мы заменим эти законы всего двумя с целью запрограммировать принцип автономности для будущих форм жизни:

- 1) Обладающая сознанием сущность имеет право на свободу думать, учиться, общаться, владеть собственностью и быть защищенной от ущерба или разрушения;
- 2) Обладающая сознанием сущность имеет право делать все, что не противоречит первому закону.

Звучит неплохо, правда? Но погодите минутку. Если животные обладают сознанием, тогда чем питаться хищникам? Все наши друзья должны стать вегетарианцами? Если некоторые замысловатые компьютерные программы будущего окажутся обладателями сознания, будет ли законно прерывать их работу? Если есть правило против прерывания цифровых форм жизни, тогда должно также быть ограничение на их создание, иначе нас ждет демографический взрыв среди цифрового населения? Всеобщая декларация прав человека была встречена единодушным одобрением, потому что только люди принимали участие в обсуждении. Как только мы начнем думать о большем количестве обладающих сознанием сущностей, чьи возможности и силы заметно различаются, мы сталкиваемся с запутанной коллизией между защитой слабых и “правом сильного”.

Также есть проблемы с принципом преемственности. Принимая во внимание, насколько развились этические взгляды со Средних веков, в том числе на рабство, права женщин и т. д., хотели бы мы, чтобы люди, которые жили 1500 лет назад, имели сильное влияние на то, как устроен мир сегодня? Если нет, почему мы должны насаждать нашу этику для будущих поколений, которые могут быть существенно умнее нас? Действительно ли мы уверены, что сверхчеловеческий искусственный интеллект захочет жить с тем, что было так дорого нашему малоразвитому интеллекту? Так четырехлетняя девочка представляет себе, как она, когда вырастет и поумнеет, построит гигантский пряничный домик, в котором сможет проводить дни напролет с конфетами и мороженым. Как и она, жизнь на Земле, скорее всего, перерастет свои детские интересы. Или представьте мышь, которая создает искусственный интеллект, соразмерный человеческому, и думает, что он будет строить города из сыра. С другой стороны, если бы мы знали, что сверхчеловеческий искусственный интеллект однажды захочет совершить космоубийство и уничтожить всю жизнь в нашей Вселенной, почему должны сегодняшние люди соглашаться на безжизненное будущее, если у нас есть силы предвосхитить его, создавая завтрашний искусственный интеллект по-другому?

В заключение мы можем сказать, что сложно полностью переформулировать даже широко распространенные этические принципы таким образом, чтобы они были приемлемы для будущего искусственного интеллекта, и эта проблема заслуживает серьезного обсуждения и исследования. В настоящий момент, однако, давайте не позволим лучшему быть врагом хорошего: есть много примеров бесспорной “детсадовской этики”, которая может и должна быть встроена в завтрашнюю технологию. Например, большому пассажирскому авиалайнеру должно быть запрещено сталкиваться с неподвижными объектами, и сейчас, когда у всех у них есть автопилот, радар и GPS, больше не должно существовать никакого технического оправдания подобному. Но несмотря на это, 11 сентября 2001 года террористы, угнав три самолета, врезались в здания-небоскребы; 24 марта 2015 года пилот-смертник Андреас Лубиц направил самолет авиакомпании GermanWings, выполнявший рейс 9525, в гору, установив автопилот на высоте 100 футов (30 метров) над уровнем моря, а остальное сделал бортовой компьютер. Сейчас, когда наши машины становятся достаточно умными, чтобы понимать, что они делают, настало время нам обозначить для них границы. Любому инженеру, разрабатывающему машину, должен задаться вопросом, есть ли какие-то вещи, которые она может, но не должна делать, и обдумать, есть ли практический способ сделать так, чтобы недобросовестный или неумелый пользователь не смог бы использовать ее во вред.

Основные цели?

Эта глава была короткой историей целей. Если мы быстро перемотаем нашу 13,8-миллиардолетнюю историю космоса, мы увидим несколько основных стадий развития целенаправленного поведения:

- 1) Вещество представляется стремящимся к максимальной диссипации;
- 2) Примитивная жизнь представляется стремящейся к максимальной репликации;
- 3) Люди не столько заинтересованы в репликации, сколько преследуют цели, связанные с получением удовольствия, удовлетворением любопытства, состраданием к ближним и другими чувствами, которые у них развились, чтобы способствовать их размножению;
- 4) Построены машины, помогающие человеку в достижении его человеческих целей.

Если эти машины со временем войдут в стадию интеллектуального взрыва, как завершится эта история целей? Может ли появиться новая система целей или возникнуть новые этические рамки, с которыми согласятся почти все сущности, объединенные процессом общего развития интеллекта? Другими словами, нет ли у нас своего рода этической предопределенности?

Даже поверхностное ознакомление с литературой по истории человечества может обнаружить намеки на такую конвергенцию: в книге *The Better Angels of Our Nature* Стивен Пинкер заявляет, что общество на протяжении тысячелетий становилось менее жестоким и более нацеленным на взаимодействие и что многие части мира прогрессируют по части принятия различий, автономии и демократии. Еще один намек на конвергенцию заключается в том, что поиск истины с помощью научного метода за последнее тысячелетие получил значительное распространение. Возможно, однако, что данные тенденции показывают конвергенцию не конечных целей, а всего лишь вспомогательных. Например, на рис. 7.2 видно, что поиск истины (более точной модели мира) – не более чем вспомогательная цель при достижении практически любой конечной цели. Также мы выше видели, как этические принципы – взаимопомощь, диверсификация и автономность – могут рассматриваться как вспомогательные цели, ибо помогают обществу функционировать более эффективно, а следовательно – выживать и достигать более основательных целей, которые у него могут быть. Кто-то может даже сказать: никаких “человеческих ценностей” и нет, а есть только “протокол о сотрудничестве”, более эффективно помогающий в достижении вспомогательной цели взаимопомощи. В том же духе, заглядывая вперед, мы можем предположить, что вспомогательные цели любой сверхразумной системы искусственного интеллекта должны включать в себя эффективный “хард”, эффективный “софт”, поиск истины и любопытство просто потому, что эти вспомогательные цели помогают в достижении любых возможных конечных целей.

Против гипотезы этической предопределенности решительно выступает Ник Бострём в своей книге *Superintelligence*, проводя противоположную линию и выстраивая, как он его называет, *тезис ортогональности*, утверждающий, что конечные цели системы могут быть независимы от ее интеллекта. По определению, интеллект – это просто способность достигать сложной цели вне зависимости от того, какова эта цель; таким образом, тезис ортогональности звучит вполне логично. В конце концов, люди могут быть умными и добрыми или умными и злыми, и ум может быть использован для совершения научных открытий, создания произведений искусства, помощи людям или планирования террористических акций^[96].

Тезис ортогональности расширяет наши возможности, говоря, что основные цели жизни в нашем космосе не предопределены и что мы вольны в выборе и оформлении их. Он предполагает, что гарантированная конвергенция к единой цели обнаруживается не в будущем, а в прошлом, когда всякая жизнь возникала с единственной простой целью –

заниматься репликацией. По прошествии космического времени наиболее развитые умы получили возможность освободиться от этой банальной цели и самостоятельно выбирать себе другие. Мы, люди, не полностью свободны в этом отношении, так как многие цели остаются генетически запрограммированными в нас, но искусственный интеллект может получить удовольствие абсолютного освобождения от таких первичных целей. Эта возможность бо2льшей свободы от конечных целей очевидна в сегодняшних простых и ограниченных системах с искусственным интеллектом: как я говорил ранее, единственная цель компьютера, играющего в шахматы, – выиграть, но есть такие компьютеры, цель которых – проигрывать в шахматы, и они участвуют в соревнованиях по игре в поддавки, где их цель – вынуждать противника атаковать. Возможно, эта свобода от предубеждений эволюции может сделать системы с искусственным интеллектом более этичными, чем люди, в некотором очень глубоком смысле: философы-моралисты, такие как Питер Сингер, утверждают, что многие люди ведут себя неэтично по эволюционным причинам – например, проявляя дискриминацию по отношению к отличным от людей животным.

Нам уже известно: краеугольной в ви2дении дружественного искусственного интеллекта оказывается идея, что искусственный интеллект, постоянно самосовершенствуясь, сохраняет свои конечные (дружественные) цели по мере того, как становится все более умным. Но как может “конечная цель” (“финальная”, как называет ее Бострём) быть даже просто определена для сверхразума? Мне кажется, мы не можем быть уверены в ви2дении дружественного AI до тех пор, пока не ответим на этот принципиальный вопрос.

Исследования в области искусственного интеллекта обычно предполагают наличие четко и однозначно сформулированной финальной цели у умных машин: например, выигрывать в шахматы или, не нарушая правил дорожного движения, довести автомобиль до заданной конечной точки. То же самое сохраняется и для большинства задач, которые мы ставим перед человеком, потому что временной горизонт и контекст известны и ограничены. Но сейчас мы говорим о будущем жизни в нашей Вселенной, не ограниченной ничем, кроме (и это в полной мере не известно) законов физики, поэтому определение целей ужасно важно! Если не учитывать квантовые эффекты, по-настоящему хорошо определенная цель укажет, как все частицы в нашей Вселенной должны быть организованы к концу времен. Но неясно, существует ли в нашей Вселенной хорошо определенный конец времен. А если частицы окажутся организованы таким образом раньше этого срока, то такая их организация долго не продлится. И в любом случае – какая организация частиц предпочтительна?

Нам, людям, свойственно отдавать предпочтение одной организации частиц перед другой; например, мы предпочитаем, чтобы в нашем родном городе сохранялась привычная организация, а не возникала иная под действием взрыва водородной бомбы. Поэтому предположим, что мы пытаемся определить функцию “хорошести”, которая ставит в соответствие каждой возможной организации частиц в нашей Вселенной число, показывающее, насколько “хорошей” мы сочтем такую организацию, и потом зададим сверхразумному искусственному интеллекту цель доводить это число до максимума. Такой подход выглядит вполне разумным, поскольку описание целенаправленного поведения как приведения какой-то функции к ее экстремальному значению популярно в других областях науки: например, экономисты часто моделируют людей, пытающихся увеличить то, что они называют “функцией полезности”, и многие разработчики умных систем тренируют свои интеллектуальные программы на приведение к максимальному значению того, что они называют “функцией успеха”. Однако когда мы говорим о конечных целях нашего космоса, такой подход приводит к вычислительному кошмару, так как будет необходимо определять “хорошесть” для каждого из более чем гуголплекс возможных организаций элементарных частиц во Вселенной, где гуголплекс – это $1 \text{ с } 10^{100}$

нулями, то есть нулей в нем больше, чем частиц во Вселенной. Как мы определим эту функцию нашему искусственному интеллекту?

Выше мы уже выяснили, что, возможно, единственная причина, по которой у нас есть какие-то предпочтения, состоит в том, что мы сами – решение оптимизационной эволюционной задачи. Таким образом, все нормативные определения для нашего человеческого языка – такие как “вкусный”, “благоухающий”, “красивый”, “удобный”, “интересный”, “сексуальный”, “значимый”, “счастливый” и “хороший”, ведут свое происхождение от этой оптимизационной эволюционной задачи, и нет никаких гарантий, что сверхразумный искусственный интеллект посчитает их строго определенными. Даже если искусственный интеллект научится точно предсказывать предпочтения какого-то представителя человеческого рода, он не сможет вычислить функцию “хорошести” для большинства способов упорядочения частиц: их подавляющее большинство соответствует странным космическим сценариям, когда нет звезд, нет планет или людей как таковых, когда у людей нет своих переживаний, и кто же тогда сможет сказать, “хорошо” ли получилось?

Есть, конечно, *некоторые функции* упорядочения космических частиц, которые могут быть строго определены, и мы даже знаем физические системы, которые эволюционируют в сторону их максимального значения. Например, мы уже обсудили, как много систем эволюционируют, стремясь к максимуму своей *энтропии*, которая при отсутствии гравитации с течением времени приводит к тепловой смерти, где все скучно, единообразно и неизменно. Так что энтропия – совсем не то, что мы хотели бы предложить искусственному интеллекту как критерий “хорошести”, к максимальному значению которого он бы стремился. Ниже приведены еще несколько примеров величин, к максимальному значению которых он мог бы стремиться и которые можно жестко определить в терминах упорядочения частиц:

- доля вещества нашей Вселенной, находящаяся внутри одного конкретного организма, например человека или кишечной палочки (как эволюционный результат максимальной эксклюзивной приспособленности);
- способность искусственного интеллекта предсказывать будущее, что, по Маркусу Гуттеру, служит хорошим показателем интеллектуального уровня;
- то, что исследователи искусственного интеллекта Александер Уисснер-Гросс и Кэмерон Фрир называют “причинная энтропия” (шлюз к будущим возможностям), которая, как они утверждают, служит фирменным знаком ума;
- вычислительная емкость Вселенной;
- алгоритмическая сложность нашей Вселенной (сколько бит нужно, чтобы описать ее);
- количество сознания во Вселенной (см. следующую главу).

Однако, отталкиваясь от физических представлений, в соответствии с которыми наш космос состоит из движущихся элементарных частиц, сложно увидеть, почему какая-то одна интерпретация “хорошести”, а не какая-то другая должна играть какую-то специальную роль. Нам еще надо определить некую финальную цель для нашей Вселенной, которая окажется и определенной, и желаемой. Единственные программируемые на данный момент цели, которые точно останутся хорошо определенными, когда искусственный интеллект станет намного более разумным, это цели, выраженные исключительно в физических категориях, таких как упорядоченность частиц, энергия и энтропия. Однако в данный момент у нас нет причин верить, что любая из таких определенных целей будет желаемой, гарантируя выживание человечества.

Напротив, кажется, что человечество – это историческая случайность, которая не является оптимальным решением ни для какой хорошо определенной физической задачи. Это означает, что сверхразумный искусственный интеллект с жестко определенной целью сможет оптимизировать ее достижение, просто исключив нас. Это означает, что

мудрое решение относительно разработок интеллектуальных систем для нас, людей, предполагает необходимость не только противостоять традиционным вычислительным сложностям, но и решать некоторые из самых упрямых вопросов философии. Чтобы запрограммировать беспилотный автомобиль, нам надо решить “проблему вагонетки”, априорно указав, кто будет жертвой во время аварии. Чтобы запрограммированный искусственный интеллект был дружелюбным, нам надо вложить в него представление о смысле жизни. А что такое “смысл”? И что такое “жизнь”? Каковы конечные этические императивы? Другими словами, как мы можем пытаться формировать будущее нашей Вселенной? Если мы вручим свою судьбу сверхразумному искусственному интеллекту до того, как ответим на эти вопросы, ответ, к которому он придет, вряд ли будет включать наше существование. Поэтому сейчас самое время еще раз вернуться к классическим спорам о философии и этике и отнести к ним как к весьма срочному делу!

Подведение итогов

- Первопричины целенаправленного поведения – в законах физики, подразумевающих достижение оптимума.
- В термодинамике содержится встроенная цель диссипации энергии за счет увеличения беспорядка, измеряемого энтропией.
- Жизнь – это феномен, который способствует диссипации (увеличивая перемешивание всего со всем), ускоряя ее благодаря своей сложности, нарастание которой, как и репликация носителей жизни, способствует хаотичности среды.
- Дарвиновская эволюция переориентирует целенаправленное поведение с диссипации на репликацию.
- Интеллект – это способность к достижению сложных целей.
- Так как у людей не всегда есть ресурсы, позволяющие разобраться, какова оптимальная стратегия репликации, у нас развились полезные эмпирические правила, которые помогают нам принимать решения и которые мы воспринимаем как сигналы к действию – чувство голода, жажды или боли, как вожделение или сострадание.
- Таким образом, у нас больше нет такой простой цели, как репликация; когда наши чувства вступают в противоречие с целью наших генов, мы подчиняемся чувствам и поэтому, например, используем контрацептивы.
- Мы строим все более умные машины, которые помогают нам достигать наших целей. К тому времени, когда мы построим машину, демонстрирующую целенаправленное поведение, мы постараемся привести цели машины в соответствие с нашими.
- Приведение целей машины в соответствие с нашими включает три нерешенные проблемы: сделать так, чтобы машина узнала о них, приняла их и продолжала их придерживаться.
- Искусственный интеллект в момент создания может иметь практически любую цель, но практически любая достаточно амбициозная цель приведет к появлению вспомогательных целей: самосохранению, захвату ресурсов и удовлетворению любопытства, чтобы лучше понимать мир. Первые две могут потенциально привести к тому, что сверхразумный AI причинит проблемы людям, а последняя может не дать ему сохранить цели, которые мы в него заложили.
- Несмотря на то, что многие общие этические принципы приняты большинством людей, непонятно, как применить их к другим видам, таким как животные или будущие AI.
- Непонятно, как вдохновить сверхразумный AI конечной целью, которая не была бы неопределенной и не приводила бы к истреблению человечества, и это снова возвращает нас к исследованиям некоторых сложнейших вопросов философии!

Глава 8

Сознание

Я не могу представить себе последовательной теории всего, которая не включала бы сознание.

Андрей Линде, 2002

Мы должны стремиться к росту самого сознания – чтобы производить больше света и чтобы он светил ярче во Вселенной, которая без него будет темна.

Джулио Тонони, 2012

Мы видели, что искусственный интеллект может помочь нам создать прекрасное будущее, если мы сумеем найти ответы на некоторые из самых старых и самых сложных философских вопросов – иногда философия тоже бывает необходима. Сейчас, по словам Ника Бострёма, мы пришли к философскому рубежу. В этой главе рассмотрим один из тех философских вопросов, в котором скрывается множество терний, – вопрос о сознании.

Кого заботит?

Говорить о сознании не просто. Если вы упоминаете “слово на С” в разговоре со специалистом по искусственному интеллекту, неврологом или психологом, они возведут глаза к небу. Если собеседник – ваш наставник, он может сжалиться над вами и попытается отговорить вас тратить напрасно свое время на безнадежную и ненаучную проблему. Действительно, мой друг Кристоф Кох, известный нейробиолог, который стоит во главе Алленовского института мозга, рассказывал мне, как его однажды отговаривали заниматься проблемами сознания, пока у него нет постоянной профессорской позиции – и это был, ни больше ни меньше, нобелевский лауреат Фрэнсис Крик. Если вы посмотрите, что написано в статье о сознании в психологическом словаре Макмиллана 1989 года издания, то узнаете, что “ничего стоящего прочтения по этой теме не было написано”^[97]. Я объясню в этой главе, что я более оптимистичен.

Хотя мыслители ломают голову над тайной сознания уже тысячи лет, появление искусственного интеллекта внезапно сделало разрешение ее насущным, в частности в том, что касается возможности предсказать, какие разумные сущности обладают способностью к субъективным переживаниям. Как мы видели в главе 3, вопрос о том, следует ли предоставлять умным машинам какие-либо права, принципиально зависит от того, действуют ли они сознательно, способны ли они страдать или испытывать радость. Как мы обсуждали в главе 7, у нас нет никакой надежды ни на какую утилитарную этику, направленную на достижение максимума положительных переживаний, если мы не знаем, какие разумные субъекты способны на них. Как уже упоминалось в главе 5, некоторые люди могут предпочитать, чтобы их роботы не обладали сознанием, так как иначе они будут ощущать себя рабовладельцами и мучиться чувством вины. С другой стороны, люди могут хотеть и прямо противоположного, когда заливают в компьютер свое сознание, освобождаясь от биологических ограничений: в конце концов, какой смысл заливать себя в робота, который разговаривает как вы и поступает как вы, если он всего лишь бессознательный зомби? В данном случае я имею под этим в виду, что после такой заливки вы и сами не будете ничего чувствовать. Разве это не будет эквивалентно самоубийству с вашей субъективной точки зрения, даже если ваши друзья не будут понимать, что ваш субъективный мир мертв?

Для долгосрочной перспективы будущей жизни в космосе (глава 6) понимание того, что делается сознательно, а что нет, становится ключевым: если развитие технологий позволяет разумной жизни процветать всей нашей Вселенной миллиарды лет, какие основания у нас могут быть для уверенности в том, что эта жизнь останется сознательной и сможет сознательно переживать происходящее? В противном случае не превратится ли все в ту самую пьесу, которую описал знаменитый физик Эрвин Шрёдингер: “[разыгрываемая] перед пустыми рядами кресел, она не существует ни для кого, и не правильнее ли тогда говорить о ней как о несуществующей”^[98]? Иными словами, если мы активируем своих высокотехнологичных потомков, которых ошибочно считаем сознательными, то не будет ли это как раз тем самым финальным зомби-апокалипсисом, превращающим наши колоссальные космические авуары в астрономическое пространство, занятое понапрасну?

Что такое сознание?

Многие споры о сознании производят больше жара, чем света, потому что спорящие не понимают друг друга, не отдавая себе отчета в том, что используют различные определения “слова на С”. Так же, как для слов “жизнь” и “разум” не существует одного бесспорно правильного определения, не существует его и для слова “сознание”. Вместо него есть много конкурирующих: в числе многих – чувствительность, пробужденность, внимание, обращенное на себя, внимание к ощущениям от органов чувств, способность переводить информацию в нарратив^[99]. В нашем исследовании будущего разума мы хотим занять предельно общую позицию, которая включала бы в себя все прочие и не ограничивалась только биологическими формами сознания, которые до сих пор существовали. Вот почему определение, которое я дал в главе 1 и которого придерживаюсь в этой книге, так широко:

сознание = субъективные переживания.

Другими словами, если ваше существование связано с какими-то чувствами прямо сейчас, то это означает, что вы обладаете сознанием. Это основополагающее определение, принятое нами во всех вопросах относительно искусственного интеллекта в предыдущих разделах. Что значит быть Прометеем, AlphaGo или беспилотным автомобилем Tesla? Связано ли это с какими-нибудь чувствами?

Чтобы оценить, насколько широко наше определение сознания, обратите внимание, что в нем не упоминаются ни поведение, ни восприятие, ни самоощущение, ни эмоции или внимание. Согласно этому определению, вы сознательны даже во сне, несмотря на то, что у вас нет доступа к органам чувств сенсорного ввода и вы не можете (я надеюсь!) ходить и совершать какие-то другие действия. Схожим образом любая система, которая испытывает боль, в этом смысле сознательна, даже если лишена возможности двигаться. Наше определение оставляет открытой возможность обладать сознанием для будущих систем с искусственным интеллектом, даже если они будут существовать лишь просто как “софт”, без подключения к сенсорам и без роботоподобного тела.

Такое определение не оставляет возможности обойти сознание вниманием. Как писал в своей книге *Homo Deus* Юваль Харари: “...Если ученый берется утверждать, что ссылки на субъективные переживания в науке неуместны, то он должен, также не ссылаясь на субъективные переживания, доказать недопустимость пыток или насилия”. Без такой ссылки каждый из нас представляет собой всего лишь сгусток элементарных частиц, передвигающийся в пространстве в соответствии с законами физики, – и как могут быть недопустимы какие-то действия?

В чем же проблема?

Так в чем же именно заключается то, чего мы не понимаем о сознании? Мало кто посвятил этому вопросу больше усилий, чем Дэвид Чалмерс, известный австралийский философ, со своей неизменной улыбкой и кожаным пиджаком, который так нравится моей жене, что она даже подарила мне такой же на Рождество. Он последовал зову своего сердца и стал философом, даже несмотря на победу на международной олимпиаде по математике и даже несмотря на то, что единственной четверкой, полученной им в колледже при остальных пятерках, была четверка по философии. Действительно, на него, кажется, никак не действуют язвительные замечания и острые возражения, и я был поражен его способностью вежливо выслушивать мнения неосведомленных или глубоко заблуждающихся людей относительно его работ, даже не чувствуя необходимости как-то на них реагировать.

Дэвид всегда подчеркивал, что на самом деле есть две разных тайны, связанных с работой человеческого ума. Первая тайна заключается в том, как мозг перерабатывает информацию. Отсюда проистекает то, что Дэвид называет “легкими” проблемами. Например, как мозг принимает сигналы от органов чувств, воспринимает их и откликается на них? Как он может сообщать о своем внутреннем состоянии средствами языка? Хотя эти вопросы чрезвычайно сложны, они, по нашему определению, относятся не к тайнам сознания, а к тайнам интеллекта: речь о том, как мозг запоминает, вычисляет и учится. Кроме того, мы видели в первой части книги, как специалисты по искусственному интеллекту уже начали успешно двигаться по пути решения этих “легких проблем” в приложении их к машинам – обучая машины играть в го и водить автомобиль, распознавать изображения и понимать высказывания, сделанные с помощью естественного языка.

И есть отдельная тайна: для чего вам личные переживания? И Дэвид называет ее “сложной” проблемой. Когда вы сидите за рулем автомобиля, вы реагируете на мелькающие за окном яркие пятна предметов, на услышанные звуки, испытываете эмоции и ощущаете себя собой. Для чего это все? Разве автомобиль без водителя испытывает что-нибудь подобное? Если вы поедете с таким беспилотным автомобилем наперегонки, вы оба будете получать информацию от своих сенсоров, перерабатывать ее и выдавать управляющие команды. И ваши субъективные переживания во время вождения очевидно логически отделены от самого вождения, они какие-то необязательные. Но если это так, то что их вызывает?

Я подхожу к этой трудной проблеме как физик. С моей точки зрения, обладающая сознанием личность состоит из того, что ею съедено, только съеденное вещество как-то иначе упорядочено. Так почему одни способы упорядочить то же самое вещество приводят к возникновению сознания, а другие нет? Кроме того, физика учит нас: все, что мы едим, – это просто большое количество кварков и электронов, упорядоченных определенным образом. Итак: в каком порядке должны быть расположены частицы, чтобы возникло сознание^[56]?

Что мне нравится в этой физической точке зрения, так это то, что она позволяет придать трудной проблеме, с которой мы, люди, пытаемся совладать на протяжении тысячелетий, значительно более ясную форму, словно фокусируя наш взгляд, и теперь ее легче решить методами науки. Вместо того чтобы начинать с *трудной* проблемы, – почему определенным образом упорядоченные частицы могут обладать сознанием, – давайте начнем с *твердого* факта, что некоторые конгломераты упорядоченных частиц ощущают себя сознательными, в то время как прочие не ощущают. Например, вы знаете, что частицы, из которых состоит ваш мозг, в данный момент находятся в осознающей упорядоченности, но не обладают этим качеством, когда вы находитесь в состоянии глубокого сна без сновидений.

Теперь мы можем задать три сложных физических вопроса относительно сознания (см. рис. 8.1). Во-первых, какие особенности упорядочения частиц играют здесь роль? В частности, какие физические особенности отличают обладающие сознанием системы от бессознательных? Если мы можем ответить на этот вопрос, то мы можем сказать, какие интеллектуальные системы окажутся сознательными. В более близкой перспективе умение ответить на этот вопрос поможет врачам скорой помощи определить, кто из не реагирующих на внешние раздражители пациентов на самом деле находится в сознании.

Во-вторых, как физические свойства определяют тип ощущения? В частности, что определяет *квалиа*, основные строительные блоки сознания, такие как красный цвет розы, звон тарелок, запах стейка, вкус мандарина или боль от укола иглой ^[57]?

В-третьих, почему что-то вдруг оказывается сознательным? Другими словами, есть ли какие-то глубинные, не открытые пока объяснения, почему одни сгустки вещества оказываются сознающими, или это необъяснимый факт, просто мир так устроен?



Рис. 8.1

Понимание сознания предполагает целую иерархию проблем. Та, которую Дэвид Чалмерс называет “легкой”, может быть сформулирована без упоминания личных переживаний. Очевидный факт, что некоторые, хотя и далеко не все, физические системы обладают сознанием, ставит перед нами три отдельных вопроса. Если у нас есть теория, дающая ответ на вопрос, который здесь обозначен как “довольно трудный”, то эта теория должна быть экспериментально проверяема. Если проверка состоялась, то можно пользоваться ею в попытках дать ответ и на вопрос в вершине пирамиды.

Специалист по информатике Скотт Ааронсон, мой бывший коллега по MIT, легкомысленно назвал первый из этих вопросов “довольно трудной проблемой” (pretty hard problem, или PHP), как перед этим поступил Дэвид Чалмерс. Следуя за ними в том же духе, назовем два других вопроса “еще более трудной проблемой” (even harder problem,

или ENP) и “по-настоящему трудной проблемой” (really hard problem, или RHP), в соответствии с рис. 8.1 [\[58\]](#).

Сознание за пределами науки?

Когда люди говорят мне, что изучение сознания – это пустая трата времени, их главный аргумент заключается в том, что такое исследование “ненаучно” и обречено оставаться таким навсегда. Но действительно ли это так? Выдающийся австро-британский философ Карл Поппер был автором широко признанного теперь положения, что “нефальсифицируемая теория не может считаться научной”. Другими словами, суть науки состоит в проверке теорий данными наблюдений: если наблюдения не позволяют проверить теорию даже в принципе, то отсюда логически следует невозможность каким-либо образом ее опровергнуть (фальсифицировать), что, по определению Поппера, означает, что она ненаучна.

Так возможна ли научная теория, дающая ответ на любой из трех вопросов о сознании, приведенных на рис. 8.1? Позвольте мне уверить вас, что ответом должно быть решительное “Да!”, – по крайней мере, для довольно трудной проблемы: “Какие физические особенности отличают обладающие сознанием системы от бессознательных?”. Предположим, что у кого-то есть теория, позволяющая для любой физической системы на вопрос, обладает ли она сознанием, ответить одним из трех способов: “да”, “нет” или “неизвестно”. Давайте подцепим ваш мозг к устройству, считывающему кое-какую информацию, обрабатываемую в различных частях вашего мозга, и пусть эта информация поступает оттуда в компьютер, на котором установлена программа, использующая теорию сознания для указания, какая часть этой информации соответствует сознательным системам, а какая нет, и представляющая вам результаты своих прогнозов в режиме реального времени на экране, как показано на рис. 8.2. Сначала вы думаете о яблоке. На экране высвечивается сообщение, что в вашем мозгу есть информация о яблоке, в чем вы отдаете себе отчет, но в стволе мозга также есть информации о частоте вашего пульса, в чем вы не отдаете себе отчета. Вас бы это впечатлило? Хотя первые два предсказания теории были правильными, вы решили бы провести более тщательное тестирование. На этот раз вы думаете о своей матери, и компьютер сообщает вам, что в вашем мозгу есть информация о вашей матери, но вы не отдаете себе в этом отчета. Теория привела к неправильному выводу, и поэтому должна отправиться на свалку истории науки вместе с аристотелевской механикой, теорией светового эфира, геоцентрической космологией и уймой других неудачных идей. Но вот главное: хотя на основании этой теории был сделан ложный вывод, он был проверяем, а значит, теория была научной! Не была бы она научной, вы были бы не в состоянии подвергнуть ее испытанию и опровергнуть.

Некоторые могут не согласиться с моим выводом и сказать, что у *них* нет никаких доказательств, что именно вы осознавали, или даже вообще осознавали ли вы что-нибудь, хотя они слышали, как вы говорите, что у вас есть сознание: бессознательный зомби, наверное, тоже стал бы говорить то же самое. Но это не делает теорию сознания ненаучной, потому что мы можем поменяться с ними местами и провести проверку, насколько правильно будут предсказаны их собственные переживания, заново.

С другой стороны, если теория отказывается делать какие-либо прогнозы, отвечая “неизвестно” на любой запрос, то опровергнуть ее невозможно, и поэтому она ненаучна. Такое может произойти потому, что теория применима только в некоторых ситуациях, или потому, что необходимые вычисления слишком трудно осуществить на практике, или потому, что датчики, следящие за работой мозга, недостаточно чувствительны. Самые популярные из научных теорий сегодня, как правило, находятся где-то посередине, давая проверяемые ответы лишь на некоторые, а не на все наши вопросы. Например, ключевая теория современной физики отказывается давать ответы на вопросы о системах, которые одновременно и чрезвычайно малы (для этого нужна квантовая механика), и чрезвычайно массивны (для этого требуется общая теория относительности), потому что мы еще не

знаем, какими математическими уравнениями надо пользоваться в данном случае. Эта ключевая теория не позволяет нам точно предсказать массы всех возможных атомов: в этом случае, как мы думаем, у нас есть необходимые уравнения, но мы не умеем решать их с нужной точностью. В чем более опасные конструкции ухитряется теория засовывать свой нос, делая при этом проверяемые предсказания, тем больше от нее пользы, и тем серьезнее мы относимся к ней, если она переживает все наши попытки убить ее. Да, мы можем проверять только некоторые из предсказаний теории сознания, но ведь так оно и для всех физических теорий. Поэтому давайте не терять время, сокрушаясь о том, что нам чего-то невозможно проверить, а займемся проверкой того, что проверяемо!



Рис. 8.2

Допустим, компьютер измеряет информацию, обрабатываемую вашим мозгом, и оценивает, какую ее часть вы осознаете в соответствии с теорией сознания. Вы можете использовать научный метод и проверить теорию, определив, насколько справедливы были априорные оценки компьютера, когда сравните их со своими личными переживаниями.

Подводя итог, мы можем сказать, что любая теория, позволяющая предсказать, какие физические системы обладают сознанием (довольно сложная проблема), научна в той мере, в какой она может распознать, какие из процессов у вас в мозгу протекают осознанно. Однако проверяемость гипотез становится менее очевидной для вопросов, расположенных выше на рис. 8.1. Что может означать, что теория предсказывает ваше субъективное восприятие красного цвета? И если мы ожидаем, что теория в первую очередь объясняет, почему существует такая вещь, как сознание, то как мы можем проверить ее экспериментально? Мы не должны избегать таких вопросов просто потому, что на них трудно ответить, и мы действительно вернемся к ним ниже. Но когда мы встречаемся с несколькими вопросами, на которые не можем дать ответа, то, я думаю, целесообразно начать с самого простого. По этой причине мои исследования сознания в MIT непосредственно сосредоточены на основании пирамиды, изображенной на рис. 8.1. Я недавно обсуждал эту стратегию с моим коллегой-физиком Питом Хатом из Принстона, который пошутил, что пытаться строить верхнюю часть пирамиды до того, как подведен фундамент, то же самое, что беспокоиться об интерпретации квантовой механики до того, как открыто уравнение Шрёдингера, – то есть создана математическая основа, позволяющая нам прогнозировать результаты наших экспериментов.

Обсуждая, что выходит за рамки науки, надо помнить, что ответ зависит от времени! Четыре столетия назад Галилео Галилей был так впечатлен возможностями физической теории, построенной на основе математики, что природу сравнил с “книгой, написанной на языке математики”. Бросая виноградину или орех, он мог точно предсказать, какой формы будет их траектория и когда они упадут на землю. Но он понятия не имел, почему первая из них – зеленая, а второй – коричневый, или почему виноградина мягкая, а орех жесткий. Эти вопросы в то время лежали за пределами науки. Но это не значило, что они останутся там навсегда! Когда Джеймс Клерк Максвелл в 1861 году открыл свои уравнения, носящие теперь его имя, стало ясно, что свет и цвет также можно описать математически. Теперь мы знаем, что упомянутое выше уравнение Шрёдингера, открытое им в 1925 году, можно использовать для предсказания любых свойств материи, включая мягкость или жесткость. В то время как прогресс научных теорий позволяет теперь делать все больше научных прогнозов, технический прогресс позволяет проводить все больше экспериментальных проверок, и почти все, что мы теперь изучаем при помощи телескопов, микроскопов и ускорителей элементарных частиц, было когда-то за пределами науки. Другими словами, горизонты науки со времен Галилея резко расширились, от крошечной точки до колоссального диапазона, простирающегося от субатомных частиц до черных дыр и наших космических истоков 13,8 миллиардов лет тому назад. И поэтому теперь вполне уместно спросить: а что же осталось за ним?

Для меня сознание – это слон в комнате. Вы знаете, что обладаете сознанием, но *только это* вы и знаете с полной уверенностью: все остальное – это результат умозаключений, как указывал Рене Декарт еще во времена Галилея. Случится ли так, что теоретический и технологический прогресс твердой рукой перенесет в конечном счете даже сознание в сферу науки? Мы этого не знаем, точно так же, как Галилей не знал, откроем ли мы когда-нибудь природу света и вещества^[59]. Гарантировано лишь одно: если мы не попытаемся, то ничего и не удастся! Вот почему я и многие другие ученые всего мира так настойчиво пытаемся формулировать и проверять теории относительно сознания.

Что говорит о сознании эксперимент?

В наших головах в каждую секунду обрабатывается огромное количество информации. Что из этого мы осознаем, а что нет? Прежде чем мы обратимся к теориям сознания и их предсказаниям, давайте посмотрим, что мы можем узнать из экспериментов, начиная от традиционных низкотехнологичных или совсем топорных наблюдений до новейших исследований, проводившихся над мозгом.

Какое поведение осознанно?

Умножая в уме 32 на 17, вы осознанно, шаг за шагом, совершаете множество внутренних вычислительных операций. Но, предположим, я показываю вам портрет Альберта Эйнштейна и прошу назвать имя изображенного на нем человека. Как мы видели в главе 2, это тоже вычислительная задача: ваш мозг производит приближенное вычисление функции, которая любой информации о множестве разноцветных точек, полученной от ваших глаз, ставит в соответствие сигналы, управляющие движением мышц вашего рта и голосовых связок. Специалисты по информатике называют эту задачу “классификация изображений” с последующим “речевым синтезом”. Хотя в этом случае вычисление намного сложнее, чем простое перемножение чисел, вы справляетесь с ней гораздо быстрее, без видимых усилий и не осознавая последовательности действий, приводящих вас к успеху. Ваши субъективные переживания сводятся только к рассматриванию фотографии, опознанию лица и слуховому восприятию собственного голоса, когда он произносит: “Эйнштейн”.

Психологи давно уже знают, что вы можете бессознательно совершать величайшее многообразие всевозможных поступков и действий: моргать и дышать, хватать и подбирать, наконец, сохранять равновесие. Как правило, вы осознаете, что именно вы сделали, но не осознаете как. С другой стороны, поведение в незнакомой ситуации, требующей самоконтроля и следования сложным логическим правилам, абстрактных рассуждений или языкового манипулирования, имеет тенденцию быть осознанным. Это поведенческие корреляты сознания, и они тесно связаны с трудоемким, медленным и контролируемым способом мышления, который психологи называют “Системой 2”^[100].

Известно также, что длительной практикой многие обыденные действия превращаются из сознательных в бессознательные: вспомним ходьбу, плавание, езду на велосипеде, вождение автомобиля, набор текста на компьютере, бритье, завязывание шнурков на ботинках, игры на компьютере или игру на фортепиано^[101]. Действительно, хорошо известно, что у настоящих специалистов своего дела лучше всего выходит то, что они умеют делать “на автомате”, следя за происходящим словно со стороны, и выполняют конкретные действия бессознательно, не привлекая контроль нижнего уровня. Например, попробуйте прочитать следующее предложение, осознанно проговаривая каждую букву, как вы делали, когда учились читать. Вы можете почувствовать, насколько медленнее пошло дело по сравнению с тем, когда вы просто воспринимали текст на уровне слов или идей?

Действительно, бессознательная обработка информации не только возможна, но представляется теперь скорее правилом, чем исключением. Данные свидетельствуют о том, что из примерно 10^7 битов информации, которые каждую секунду наш мозг получает от органов чувств, мы можем осознавать только крошечную часть, не превосходящую, по разным оценкам, от 10 до 50 бит^[102]. То есть от той информации, что перерабатывает наш мозг, сознанию открыта разве что верхушка айсберга.

Все эти данные в совокупности привели некоторых исследователей к мысли, что сознательную обработку информации следует рассматривать вроде такого генерального

директора всего нашего ума, который контролирует принятие только самых важных решений, требующих сложного анализа данных всех органов чувств^[103]. Такой подход помогает объяснить, почему как раз в качестве директора компании он обычно не хочет отвлекаться на детали, находящиеся в ведении его подчиненных, но при желании он всегда может получить к ним доступ. Чтобы лучше понять, как работает эта избирательность внимания, посмотрите еще раз на слово “желании”, сосредоточив взгляд на букве “а” и, не перемещая глаз, перенесите внимание на весь слог “ла”, а затем и на слово целиком. Хотя информация, получаемая мозгом от вашей сетчатки, осталась той же, осознаваемые ощущения теперь иные. Метафора генерального директора также объясняет, почему при накоплении опыта растет роль бессознательного: измучившись при освоении техники чтения и печатания на клавиатуре, генеральный директор делегирует эти рутинные задачи бессознательным подчиненным, чтобы иметь возможность сосредоточить внимание на новых задачах более высокого уровня.

Где же сознание?

Остроумные эксперименты и их анализ привели к выводу, что сознание связано не только с определенными типами поведения, но и с некоторыми участками мозга. Кто в числе главных подозреваемых? Первые зацепки пришли из опыта лечения больных с локализованными повреждениями мозга, возникшими вследствие аварии на дороге, удара, опухоли или инфекции. Но они часто дают лишь предварительные сведения. Например, означает ли тот факт, что поражение задней доли коры головного мозга вызывает слепоту, участие этой доли в формировании визуального сознания, или она просто расположена на пути визуальной информации куда-то еще, где она впоследствии станет сознательной, и информация через нее только проходит, точно так же, как до этого проходила через глаз?

Хотя исследования ни повреждений головного мозга, ни последствий медицинского вмешательства не дают надежных указаний, в каком именно месте формируются осознанные ощущения, они сужают круг возможных вариантов. Например, я знаю, что, хотя я испытываю боль в руке, как будто она и в самом деле возникает именно там, ощущение боли должно зарождаться где-то в другом месте, потому что один раз хирург обезболит мою руку, ничего не делая с ней самой: он просто ввел анестетик мне в плечо поблизости от проводящего нерва, это называется проводниковой анестезией. Кроме того, некоторые люди с ампутированными конечностями испытывают фантомные боли в несуществующей конечности. Вот еще один пример: я однажды заметил, что, когда я закрываю левый глаз и смотрю только правым, часть поля зрения у меня пропадает – врач определил, что у меня отслоилась сетчатка, и прикрепил ее обратно. В отличие от меня, некоторые пациенты с определенными повреждениями головного мозга испытывают одностороннее пространственное игнорирование, из-за которого у них тоже пропадает половина поля зрения, но они даже не замечают, что что-то пропало, – например, они могут не видеть одной половины своей тарелки и съесть только половину лежащей на ней пищи. Получается так, будто половина мира исчезла из их сознания. Но можем ли мы утверждать, что поврежденные области мозга создают пространственные ощущения, или же они просто передают сознанию какую-то информацию о пространстве, как это делает сетчатка глаза?

Нейрохирург-новатор, живший в США и Канаде, Уайлдер Пенфилд, обнаружил в 1930-х годах, что его пациенты ощущают прикосновения к различным частям своего тела, когда он раздражает электричеством некоторые области их мозга, относящиеся к тому, что теперь называется *соматосенсорной корой* (рис. 8.3)^[104]. Он также обнаружил, что его пациенты совершают произвольные движения различными частями тела, когда он стимулирует области мозга, относящиеся к тому, что теперь называется *моторной корой*.

Но означает ли это, что информация, обрабатываемая в этих областях мозга, соответствует осознанию осязания и совершения движений?

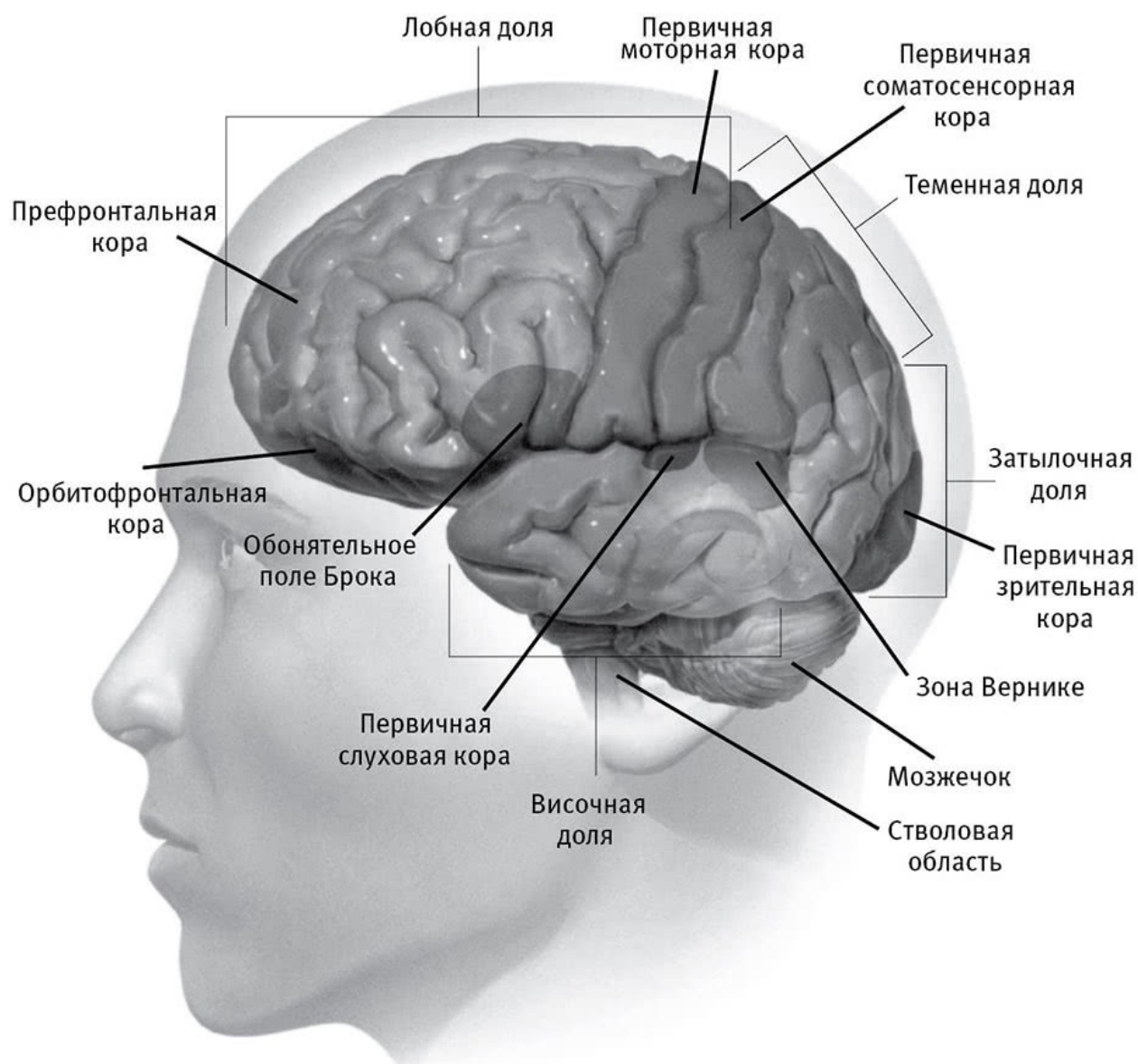


Рис. 8.3

Аудиовизуальные и соматосенсорные области коры головного мозга отвечают за зрение, слух, осязание и двигательные функции, но из этого еще не следует, что именно в этом месте происходит осознание увиденного, услышанного, осязанного или произведенного движения. И в самом деле, последние исследования показывают, что в визуальной области коры головного мозга, как и в мозжечке или в стволе мозга, совершенно не происходит никакого осознания. Изображение предоставлено www.lachina.com

К счастью, современная технология дает нам теперь более надежные зацепки. Хотя мы все еще очень далеки от возможности измерить каждую искорку, пробежавшую по какому-нибудь из примерно 100 миллиардов ваших нейронов, технология чтения мозга развивается очень быстро, возникают все новые методы с пугающими названиями: fMRI, EEG, MEG, ECoG, ePhys^[60], флуоресцентное зондирование напряжений. Функциональная магнитно-резонансная томография заключается в том, чтобы, следя за магнитными свойствами ядер водорода, делать 3D-карту мозга примерно каждую секунду с разрешением в один миллиметр. Электроэнцефалография и магнитоэнцефалография заключаются в измерении электростатического и магнитостатического полей вне головы,

тысячу раз в секунду снимая карту мозга, но с невысоким разрешением, не позволяющим различать детали размером меньше нескольких сантиметров. Пугливому человеку особенно приятно, что все эти методы не инвазивные. Но если вы не против, чтобы вам вскрыли черепную коробку, у вас появляются дополнительные опции. Электрокортикография предполагает размещение, скажем, сотни проводков на поверхности вашего мозга, в то время как электрофизиология подразумевает внедрение микропроводов, иногда тоньше человеческого волоса, глубоко в мозг для записи электрического напряжения, измеряемого в нескольких тысячах точек одновременно. Многие больные эпилепсией вынуждены проводить в больнице по много дней, пока с помощью электрокортикографии выяснится, какая часть мозга вызывает судороги и должна быть удалена, и они любезно соглашались, чтобы, пока суд да дело, неврологи проводили свои эксперименты, касающиеся сознания. Наконец, флуоресцентное зондирование напряжения состоит в том, чтобы генетическими манипуляциями над нейронами заставить их испускать вспышки света, когда по ним проходит сигнал, что позволяет следить за их активностью с помощью микроскопа. Из всех методов именно последний обеспечивает самый быстрый доступ к наибольшему количеству нейронов, по крайней мере для тех животных, у кого прозрачные мозги – например, для червя нематоды (*Caenorhabditis elegans*) с его 302 нейронами и личинок рыбки данио рерио (*Danio rerio*), у которой их около 100 000.

Хотя Фрэнсис Крик предупреждал Кристофа Коха об опасностях, связанных с изучением сознания, Кристоф отказался сдаваться, и время показало его правоту. В 1990 году они вместе написали основополагающую статью о том, что они называли “нейронными коррелятами сознания” (NCC), в которой они выясняют, какие конкретно процессы в мозгу соответствуют осознанным ощущениям. На протяжении тысячелетий мыслители могли следить за обработкой информации в мозгу только через посредство собственного субъективного опыта и поведения. Крик и Кох указывали, что новые методы исследования процессов внутри мозга внезапно предоставили независимый доступ к этой информации, позволяя ученым исследовать, как должна обрабатываться информация, чтобы возникали определенные осознанные ощущения. Что и говорить, теперь методы измерений, основанные на новых технологиях, превратили поиски нейронных коррелятов сознания в настоящий мейнстрим нейробиологии, и тысячи публикаций по этой теме поступают в самые престижные из научных журналов⁽¹⁰⁵⁾.

Какие выводы мы можем сделать? Чтобы получить вкус к детективному поиску нейронных коррелятов, давайте сначала выясним, обладает ли сознанием ваша сетчатка, или же она просто зомби-система, которая записывает визуальную информацию, обрабатывает ее и посылает вниз по течению куда-то еще в вашем мозгу, где и рождается субъективное переживание визуального. Вы можете указать, какой квадрат на левой половине рис. 8.4 темнее, а какой светлее – тот, что помечен буквой А, или тот, что помечен буквой В? Квадрат А, правда? А вот и нет, на самом деле оба квадрата одинаково светлые, в чем можно убедиться, рассматривая их через небольшие просветы между пальцами, закрывающими большую часть картинки. Это доказывает, что сетчатка не может быть всецело ответственной за ваши визуальные ощущения, иначе оба квадрата выглядели бы одинаково.

А теперь посмотрите на правую половину рисунка. Что вы видите: профили двух женщин или вазу? Если вы всматриваетесь достаточно долго, в вашем сознании будет возникать то один образ, то другой, чередуясь примерно через равные интервалы времени, хотя информация, получаемая вашей сетчаткой, остается все это время одной и той же. Проводя измерения тех процессов, которые происходят в вашем мозгу, когда меняется ощущение, можно выяснить, откуда же берется различие, – и дело явно не в сетчатке, в которой ничего не меняется.

Смертельный удар гипотезе о сознательной сетчатке наносит метод, получивший название “непрерывного подавления мерцанием”, который был открыт Кристофом Кохом,

Станисласом Деаном и их коллегами: как выяснилось, если одному вашему глазу показывать нечто сложно организованное и быстро меняющееся, то это будет отвлекать ваши системы визуального восприятия до такой степени, что неподвижное изображение, показываемое другому глазу, совершенно не сможет достичь вашего сознания^[106]. Иначе говоря, на вашей сетчатке может находиться изображение, которое не вызывает у вас никаких визуальных ощущений, и в то же время вы можете испытывать визуальные ощущения тогда (например, во сне), когда никакого изображения на вашей сетчатке нет. Это доказывает, что две ваши сетчатки не в большей степени способствуют возникновению у вас визуального сознания, чем видеокамера, хотя в них производятся сложнейшие вычисления с участием более ста миллионов нейронов.

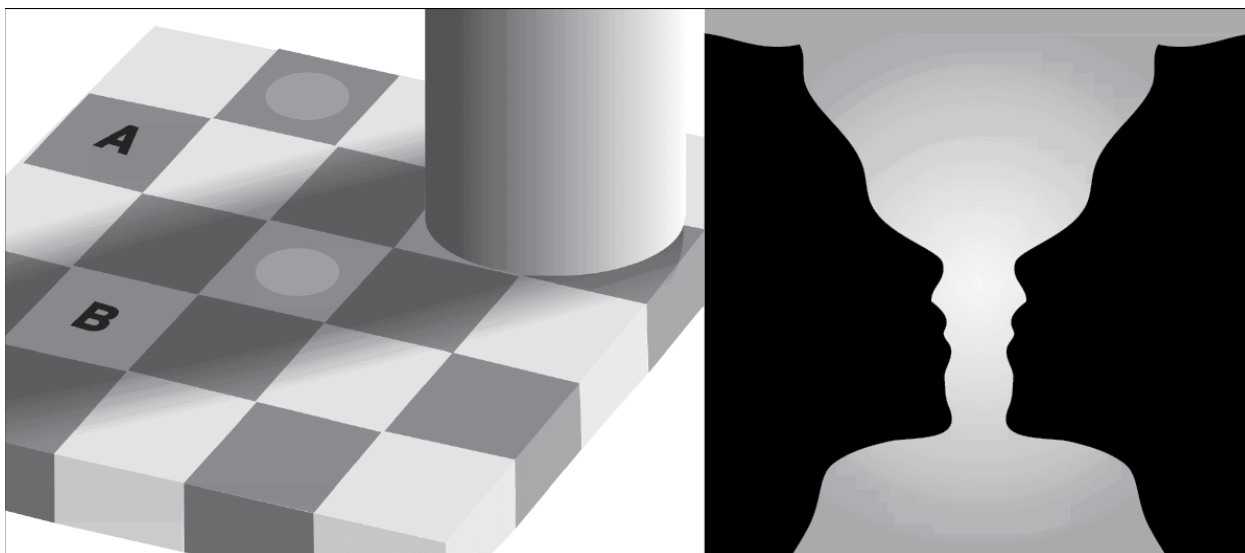


Рис. 8.4

Какая из клеток темнее – А или В? Что вы видите на правой половине рисунка – вазу, силуэты двух женщин или и то и другое по очереди? Иллюзии подобные этим доказывают, что зрительное осознание не может происходить в глазу или где-то в начале вашей зрительной системы, потому что увиденное зависит далеко не только от того, что было на картинке.

Исследователи нейронных коррелятов используют непрерывные подавления мерцанием, неустойчивые визуальные / слуховые иллюзии и другие трюки также и для того, чтобы определить, какие области мозга отвечают за каждое из ваших осознанных ощущений. Основная стратегия заключается в том, чтобы сравнить, что делают ваши нейроны в двух идентичных ситуациях, когда по существу всё (в том числе и сигналы от органов чувств) одно и то же, а осознанные ощущения различаются. Те участки головного мозга, которые в измерениях обнаруживают различное поведение, и определяются как нейронные корреляты сознания.

Такие исследования доказали и то, что никаких следов вашего сознания нет в кишечнике, несмотря на то что там располагается обширная нервная сеть с полумиллиардом нейронов, которые участвуют в вычислениях, как в регуляции переваривания пищи; все же ощущения, вроде чувства голода и тошноты, возникают в мозгу. Аналогичным образом нет никаких признаков сознания и в стволе головного мозга, нижней его части, которая подключается к спинному мозгу и управляет дыханием, сердечным ритмом и артериальным давлением. Самое поразительное, что ваше сознание не зависит, очевидно, и от вашего мозжечка (рис. 8.3), который содержит около двух третей всех нейронов: пациенты с поражениями мозжечка страдают от расстройства речи, а своими неуклюжими движениями они похожи на пьяных, но по-прежнему остаются в полном сознании.

Вопрос о том, какие же части мозга ответственны за возникновение сознания, остается открытым и служит поводом для непрекращающихся споров. Некоторые недавние исследования свидетельствуют о том, что ваше сознание проживает главным образом в “горячей зоне”, с участием таламуса (примерно в середине вашего мозга) и задней части коры (внешнего слоя мозга, состоящего из скомканного шестислойного листа, который, будучи расправлен, сравнился бы по площади с большой обеденной скатертью)^{107}. Из этого же исследования, как это ни удивительно, следует, что первичная зрительная кора в затылочной части головы отлична от всего, что тут еще есть, тем, что так же бессознательна, как глазные яблоки и сетчатки.

Когда же возникло сознание?

До сих пор мы искали экспериментальные подсказки, которые помогли бы нам распознать, какие типы обработки информации можно считать осознанными и где сознание возникает. Но *когда* оно возникает? Будучи ребенком, я был уверен, что мы осознаем события в тот же самый миг, как они происходят, абсолютно без какой-либо задержки во времени. И хотя часто мне так кажется и до сих пор, это явно субъективное чувство, которое не может быть правильным, поскольку моему мозгу требуется время, чтобы обработать ту информацию, которая поступает к нему от моих органов чувств. Исследователи нейронных коррелятов тщательно измеряли, сколько именно времени на это требуется, и окончательный итог, подведенный Кристофом Кохом, сводится к тому, что проходит около четверти секунды от момента, когда свет от какого-то сложного объекта проходит через глаз, до тех пор, пока вы осознанно его воспримете таким, каков он есть^{108}. Это означает, что если вы за рулем своего автомобиля едете по шоссе со скоростью в пятьдесят пять миль в час и вдруг видите белку в нескольких метрах перед машиной, то предпринимать что-нибудь уже слишком поздно: вы ее уже переехали!

Таким образом, ваше сознание живет в прошлом, оно отстает от внешнего мира, по оценке Кристофа Коха, примерно на четверть секунды. Любопытно, что вы часто можете среагировать быстрее, чем осознаете, что происходит, и это, между прочим, доказывает, что обработка информации, нужной для самых быстрых реакций, должна быть бессознательной. Например, если к вашему глазу приближается какой-то посторонний предмет, моргательный рефлекс опускает веко в течение всего лишь десятой доли секунды. Все происходит так, как будто одна из систем вашего мозга получает угрожающую информацию от зрительной системы, производит вычисление, показывающее, что вашему глазу грозит опасность получить удар, и сразу шлет мейл мышцам глаза с инструкцией совершить моргание, одновременно информируя другим мейлом часть мозга, заведующую сознанием: “Эй, мы будем моргать!”. К моменту получения этого мейла и включения его в ваш сознательный опыт, моргание уже завершилось.

Действительно, система, читающая все эти мейлы, постоянно подвергается бомбардировке сообщениями из всех частей вашего тела, и некоторые из этих сообщений отстают больше, чем другие. Нервным сигналам от ваших пальцев требуется больше времени, чтобы добраться до мозга, чем нервным сигналам от лица, потому что расстояние от пальцев больше. На анализ изображений у вас уходит больше времени, чем на анализ звуков, потому что это сложнее: именно поэтому старт олимпийской гонке дается гонгом, а не вспышкой. И все же: если вы касаетесь пальцами своего носа, то осознанные ощущения прикосновения пальцев к носу и носа к пальцам возникнут одновременно, а если вы хлопаете в ладоши, то и вид смыкающихся ладоней, и звук хлопка, и ощущение шлепка на коже также родятся в вашем сознании в одно и то же мгновение^{109}. А это означает, что полного сознательного переживания вами события не

произойдет до тех пор, пока самый последний мейл не прибудет на самой медленной из улиток и не будет проанализирован.

Знаменитая серия NCC-экспериментов, начало которой положил физиолог Бенджамин Либет, показала, что действия, которые можно выполнять бессознательно, не ограничиваются быстрыми реакциями вроде моргания или отбивания ракеткой шарика в пинг-понге, иногда сюда попадают и некоторые решения, принятие которых свидетельствует, можно подумать, о свободе воли – и тем не менее, измерение иногда скажет, какое решение принято, еще прежде, чем вы сами его осознаете^{110}.

Теории сознания

Мы только что убедились, что, хотя мы до сих пор не понимаем, что такое сознание, у нас накопилось удивительно большое количество экспериментальных данных о различных его аспектах. Но все эти данные получены для мозга, – чему они могут нас научить, если говорить о сознании машин? Здесь требуется массивная экстраполяция за пределы нашей нынешней экспериментальной базы. Другими словами, здесь требуется теория.

Почему теория?

Чтобы понять почему, давайте сравним теории сознания с теориями гравитации. Ученые начали серьезно относиться к теории гравитации Ньютона потому, что они получили от нее больше, чем в нее требовалось вложить: простые уравнения, легко уместящиеся на салфетке, могут дать точное предсказание исхода любого эксперимента с силой тяжести, когда-либо проводившегося. Поэтому они так же серьезно отнеслись к его предсказаниям по поводу явлений, далеко выходящих за пределы той области, где теория проверялась, и их самые смелые экстраполяции, как оказалось, подтверждают свою пригодность даже для движения галактик и скоплений в миллионы световых лет в поперечнике. Однако прогнозы чуточку разошлись для движений Меркурия вокруг Солнца. Ученые тогда начали принимать серьезно улучшенную Эйнштейном теорию гравитации, общую теорию относительности, потому что она была, возможно, даже более элегантной и экономичной и к тому же правильно предсказывала те явления, где теория Ньютона оказалось неверной. Они последовательно и серьезно экстраполировали ее предсказания далеко за пределы той области, где ее проверяли, на явления столь экзотические, как черные дыры и гравитационные волны, в самую ткань пространственно-временного континуума и расширения нашей Вселенной, от самого ее рождения в огненном бурлящем вареве – и все со временем подтверждало эксперимент.

Аналогично, если какая-то математическая теория сознания, чьи уравнения уместятся на салфетке, будет давать успешные предсказания для результатов всех тех экспериментов, которые мы проводили со своим мозгом, то мы начнем серьезно относиться не только к самой теории, но и к ее прогнозам относительно сознания вне мозга – например, в машинах.

Сознание в физической перспективе

Хотя некоторые теории сознания возникли еще в глубокой древности, самые современные из них основываются на нейробиологии и нейробиологии, пытаюсь объяснить и предсказать осознанность с точки зрения событий, затрагивающих нейроны и происходящих в мозгу^[111]. Хотя некоторые из этих теорий дали определенные успешные прогнозы для нейронных коррелятов сознания, они не могут делать прогнозы о сознании машин и не стремятся к этому. Чтобы сделать скачок от человеческого мозга к машинам, нам понадобится обобщение нейронных коррелятов сознания – *физические корреляты сознания*, которые мы определим как устойчивые образования движущихся частиц, обладающие сознанием. Потому что если теория правильно предсказывает, что осознается и что не осознается, апеллируя только к физическим строительным блокам – таким, как элементарные частицы или силовые поля, то такая теория сможет делать предсказания не только для мозга, но и для иначе организованного вещества, включая будущие системы искусственного интеллекта. Поэтому давайте встанем теперь на

физическую точку зрения: каким образом должно быть организовано множество частиц, чтобы оно было сознательно?

Но отсюда в действительности немедленно возникает еще один вопрос: каким образом нечто столь сложное, как сознание, может быть построено из чего-то столь простого, как частицы? Я думаю, что так происходит потому, что у самого этого явления имеются свойства, отличные от свойств составляющих его частиц. В физике мы называем такие явления “эмергентными” (от латинского *emergere* – возникать)^[112]. Давайте попытаемся понять, что это такое, анализируя более простое эмергентное явление, чем сознание, – смачиваемость.

Капля воды смачивает, но этого не может ни кристалл льда, ни облако пара, несмотря на то, что молекулярный состав у них у всех один и тот же. Почему? Потому что свойство смачиваемости зависит только от расположения молекул. Абсолютно бессмысленно говорить, что одна молекула воды может что-то намочить: смачивание происходит только тогда, когда большое количество молекул организовано в нечто такое, что мы называем жидкостью. Поэтому все твердые тела, все жидкости и газы представляют собой эмергентные явления: они больше, чем сумма их частей, потому что имеют свойства, лежащие за пределами свойств последних. У них есть какие-то особенности, которых лишены составляющие их частицы.

Так же как и твердые тела, жидкости и газы, так же, я думаю, и сознание – эмергентное явление, свойства которого лежат за пределами свойств составляющих его частиц. Например, погружение в глубокий сон гасит сознание простой перестановкой частиц. Таким же образом мое сознание исчезнет, если я замерзну до смерти, из-за чего составляющие меня частицы расположатся менее удачным образом.

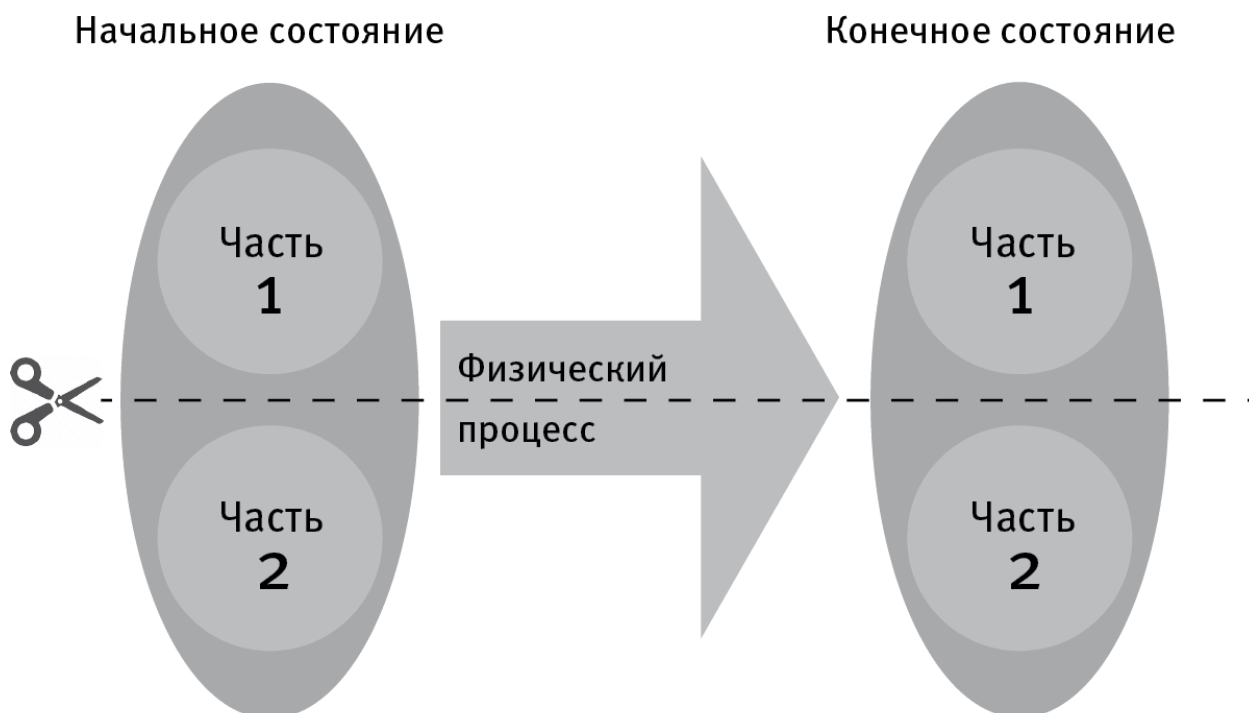
Когда вы соединяете вместе много частиц, образуя что-то новое, от воды до мозга, возникают и новые явления с наблюдаемыми свойствами. Мы, физики, любим изучать эти возникающие эмергентные свойства, которые часто могут быть идентифицированы с помощью небольшого набора чисел, – и вы можете просто пойти и измерить их: какова у возникшего вещества вязкость, какова сжимаемость и так далее. Например, если вещество оказалось настолько вязким, что даже проявляет жесткость, мы называем его твердым телом, в противном случае мы называем его текучим. И если оно текучее, но несжимаемое, то мы называем его жидкостью, в противном случае это газ или плазма, в зависимости от того, насколько хорошо оно проводит электричество.

Сознание как информация

Так могут ли быть аналогичные количественные характеристики, которые бы квантифицировали сознание? Итальянский нейробиолог Джулио Тонони предложил такую характеристику, которую назвал “интегральной информацией” и обозначил греческой буквой Φ (фи): она указывает, как много различных частей системы знают о существовании друг друга (см. рис. 8.5).

Я впервые встретился с Джулио в 2014 году на конференции по физике в Пуэрто-Рико, куда я пригласил его и Кристофа Коха, и он сразу поразил меня как глубоко ренессансная личность, образованная словно бы смешением Галилео Галилея и Леонардо да Винчи. Его мягкие манеры не могут скрыть его феноменального знания искусства, литературы и философии, а слава о его кулинарных талантах опережает его, куда бы он ни ехал. Один всемирно известный тележурналист однажды рассказывал мне, как Джулио всего за несколько минут соорудил салат, вкуснее которого он в жизни не пробовал. Но я очень быстро понял, что за его мягкими манерами скрывается бесстрашный интеллект, который не побоится следовать за данными, куда бы они ни повели его, независимо от предубеждений и табу истеблишмента. Так же, как Галилей создавал свою математическую теорию движения несмотря на растущее давление со стороны

истеблишмента, противящегося отказу от геоцентризма, так и Джулио разработал наиболее точную к настоящему времени математическую теорию сознания, в основе которой его *теория интегральной (интегрированной) информации* (ИИТ).



Φ служит мерой невозможности разделить процесс на две независимые части

Рис. 8.5

Для данного физического процесса, преобразующего с течением времени начальное состояние системы в какое-то новое состояние, его интегральная информация Φ представляет из себя меру невозможности разделить этот процесс на независимые части. Если состояние каждой части определяется только ее прошлым состоянием и не зависит от того, что происходило с другой частью, то $\Phi = 0$: то есть то, что мы называли системой, вовсе ею не является, а представляет собой две независимые системы, не коммуницирующие друг с другом.

Я уже несколько десятилетий утверждаю, что сознание – это ощущение, возникающее при обработке информации определенным сложным образом^{[\[113\]](#)}. Теория интегральной информации вполне согласуется с этим утверждением, заменяя мою расплывчатую фразу “определенным сложным образом” более точным определением: обработка информации должна быть интегрированной, то есть Φ должно быть велико. И Джулио приводит аргумент столь же сильный, сколь и простой: сознательная система должна быть интегрирована в единое целое, потому что если она будет состоять из двух независимых частей, то эти две части будут чувствовать себя как две различные сознательные сущности, а не одна единая. Другими словами, если часть мозга или компьютера, ответственная за сознание, не может коммуницировать со всеми остальными, то все остальные части не смогут участвовать в формировании его субъективного опыта.

Джулио и его коллеги измерили упрощенную версию Φ при помощи электроэнцефалографа, определявшего отклик мозга на стимуляцию магнитным полем. Их “детектор сознания” и в самом деле очень хорошо работает: прибор показывал, что пациент в сознании, если он бодрствовал или дремал (находился в состоянии легкого сна со сновидениями), и что пациент без сознания, если он был под наркозом или в состоянии глубокого сна (без сновидений). Прибор даже установил наличие сознания у двух пациентов, страдающих от синдрома “запертого человека”, которые не могли ни

двигаться, ни общаться каким-либо нормальным способом^[114]. Так что такая технология может оказаться очень перспективной для врачебной практики, чтобы выяснять, какие пациенты находятся в сознании, а какие нет.

Ловля сознания в физике

Теория интегральной информации определена только для дискретных систем, множество состояний которых конечно – это могут быть, например, биты компьютерной памяти или упрощенная модель нейронов, находящихся только в одном из двух состояний: включенном или выключенном, активном или пассивном. К сожалению, это означает, что теорию нельзя применить для большинства традиционных физических систем, параметры которых могут изменяться непрерывно: например, координаты частицы или сила магнитного поля могут принимать любое из бесконечного числа значений^[115]. Если вы будете пытаться применить теорию для таких систем, то получите бесполезный результат, когда Φ равно бесконечности. Квантовая система может быть дискретной, но теория изначально не определена для квантованных систем. Так как же нам заякорить теорию интегральной информации, а также любую другую теорию сознания, использующую теорию информации, на прочном физическом фундаменте?

Мы можем сделать это, основываясь на главе 2, из которой мы узнали о том, как сгустки вещества могут обладать эмергентными свойствами, имеющими отношение к информации. Мы видели, что в качестве устройства, пригодного для хранения информации, может использоваться только нечто такое, у чего есть много долгоживущих состояний. Мы также видели: чтобы быть *компьютрониумом*, то есть субстанцией, пригодной для проведения расчетов, кроме того, требуется сложная динамика – законы физики должны быть в состоянии заставлять эту субстанцию изменяться достаточно сложным образом, чтобы производить произвольную обработку информации. Наконец, мы видели, что нейронные сети, например, оказываются подходящим субстратом для машинного обучения, потому что, просто подчиняясь законам физики, этот субстрат способен трансформировать себя, все больше и больше приспособляясь к проведению нужных вычислений. А теперь мы задаем новый вопрос: что сделает сгусток вещества способным обладать субъективным опытом? Другими словами, при каких условиях сгусток вещества будет в состоянии проделывать следующие четыре вещи:

- 1 помнить;
- 2 считать;
- 3 учиться;
- 4 накапливать переживания.

Первые три вещи мы изучили в главе 2. Теперь к ним добавилась четвертая. Так же, как Марголус и Тоффоли создали термин *компьютрониум* для вещества, которое способно выполнять произвольные вычисления, я хотел бы предложить термин *сентрониум* для абстрактного вещества, которое способно иметь субъективный опыт (от латинского *sentio* – чувствовать, ощущать)^[61].

Но как само сознание может ощущать себя настолько нефизическим, если оно на самом деле – физическое явление? Как может оно чувствовать себя настолько независимым от своего физического субстрата? Я думаю, что так происходит потому, что оно и в самом деле не зависит от своего физического субстрата, служащего ему лишь ложем! Нам уже знакомо много красивых примеров независимости от субстрата, рассмотренных в главе 2. В их числе – волны, воспоминания и вычисления. Мы видели, как во всех этих примерах не просто возникает нечто большее, чем сумма участвующих в явлении частей (то есть эмергентность), но, скорее, независимость явления от своих частей, когда оно начинает жить собственной жизнью. Например, мы видели, что

симулированный разум будущего или персонаж компьютерной игры никоим образом не сможет распознать, какая операционная система при этом используется – Windows, Mac OS, Android на телефоне или еще какая-нибудь другая, потому что его функционирование совершенно не будет зависеть от субстрата. Не сможет он распознать, как именно реализованы логические гейты его компьютера – это транзисторы, оптические схемы или какой-то другой “хард”. Или какие использовались фундаментальные законы физики – они могут быть какими угодно, лишь бы только позволяли создание универсального компьютера.

В целом, я думаю, сознание – это физическое явление, которое воспринимает себя нефизическим потому, что оно подобно волне и вычислению: его свойства не зависят от его конкретного физического субстрата. Это логически следует из идеи представить сознание как информацию. Такое представление приводит к еще одной радикальной идее, которая мне решительно нравится: если сознание – это ощущение, возникающее при обработке информации определенным образом, то оно обязано быть не зависящим от субстрата – ведь значение имеет структура обработки информации, а не структура того вещества, в котором эта обработка производится. Другими словами, сознание оказывается дважды не зависящим от субстрата!

Как мы видели, физика описывает возмущения пространственно-временного континуума, которые соответствуют движущимся частицам. Когда конгломераты определенным образом организованных частиц подчинены определенным принципам, они порождают явления, которые мы называем эмергентными и которые очень мало зависят от порожденного частицами субстрата, и эти явления совершенно иначе воспринимаются. Яркий пример тому – обработка информации, или компьютеризм. Но теперь мы можем перенести эту идею на новый уровень: если обработка информации ведется в соответствии с определенными принципами, то результатом может оказаться эмергентное явление более высокого уровня, которое мы называем сознанием. Это помещает ваши субъективные переживания не на один, а на два уровня выше материи. Ничего удивительного, ваш дух ощущает себя нефизическим!

Но тогда возникает вопрос: каковы должны быть принципы, следуя которым, обработка информации становится сознанием? Я не претендую на знание всех *достаточных* условий, гарантирующих сознание, но есть четыре *необходимых* условия, на которые я делаю ставку и посвящаю им свои исследования:

Принцип	Определение
Информационный принцип	Сознательная система располагает значительной емкостью хранилищ информации
Динамический принцип	Сознательная система располагает значительной способностью переработки информации
Принцип независимости	Сознательная система обладает значительной независимостью от остального мира
Интегральный принцип	Сознательная система не может состоять из почти независимых частей

Как я уже говорил, по-моему, сознание – это то, что испытывает информация, когда обрабатывается определенным образом. А значит, чтобы быть сознательной, система должна иметь возможность хранить и обрабатывать информацию, что подразумевают

первые два принципа. Обратите внимание, что память не должна быть долгосрочной: я очень рекомендую посмотреть трогательное видео Клайва Уэринга о человеке, который выглядит совершенно сознательным, даже несмотря на то, что его воспоминания длятся меньше минуты⁽¹¹⁶⁾. Я также думаю, что сознательная система должна быть достаточно независимой от всего остального мира, потому что в противном случае у нее не возникнет субъективного ощущения самостоятельного существования. Наконец, я думаю, что сознательная система должна быть интегрирована в единое целое, как утверждал Джулио Тонони, потому что если бы она состояла из двух независимых частей, то они чувствовали бы себя как две отдельные сознательные сущности, а не одна. Первые три принципа предполагают *автономию*: система способна хранить и обрабатывать информацию без заметного вмешательства извне, определяя тем самым свое собственное будущее. Все четыре принципа вместе означают, что система как целое автономна, но это не относится ни к одной из ее частей.

Если эти четыре принципа верны, то значит, для нас готова новая работа: нам надо найти строгие математические теории, в которых эти принципы были бы воплощены, и провести их экспериментальную проверку. Нам необходимо также проверить, не нужны ли какие-нибудь дополнительные принципы. Независимо от того, верна ли теория интегральной информации, исследователям надо попытаться разработать конкурирующие теории и проверить их все в самых совершенных экспериментах.

Споры о сознании

Мы уже обсуждали многолетние споры о том, стоит ли заниматься изучением сознания, не оборачивается ли оно с неизбежностью ненаучной чепухой и пустой тратой времени. Вдобавок к этому недавно начались споры на самом переднем крае исследований сознания – давайте изучим те из них, которые я считаю наиболее поучительными.

Теория интегральной информации Джулио Тонони принесла ее автору не только похвалы, но и немало критики, часть которой была весьма ядовитой. Скотт Ааронсон недавно так высказался в своем блоге: “На мой взгляд, тот факт, что теория интегральной информации ошибочна, – причем явно ошибочна, по причинам, прослеживаемым к самому ее ядру, – дает ей право занять место среди лучших 2 % всех математических теорий сознания из когда-либо предложенных. Почти все конкурирующие теории сознания, кажется мне, были настолько расплывчатыми, пустыми и бесформенными, что они могут только мечтать об ошибочности”^[117]. К чести Ааронсона и Тонони надо сказать, что они ни разу не переступили черту дозволенного, когда я видел их во время обсуждения теории интегральной информации на недавнем семинаре в Нью-Йоркском университете: они вежливо выслушивали аргументы друг друга. Ааронсон показал, что у некоторых простых сетей логических гейтов чрезвычайно высокий показатель интегральной информации (Ф), и поскольку они очевидно не обладают сознанием, теория не верна. Тонони возражал, что если бы такие сети были построены, они *были бы* сознательны, и что утверждение Ааронсона об обратном антропоцентрически предвзято – примерно в той же степени, в какой утверждение хозяина скотобойни по поводу того, что у животных нет сознания, лишь на том основании, что они не могут говорить и совсем не похожи на людей. Мой анализ, с которым они оба согласились, заключался в том, что их разногласия относительно интегральной информации сводятся к тому, является ли интеграция лишь *необходимым* условием сознания (с чем Ааронсон соглашался), или это также и *достаточное* условие (на что претендовал Тонони). Последнее безусловно значительно более сильное и спорное утверждение, которое, я надеюсь, скоро можно будет проверить экспериментально^[118].

Немало споров вызывает и другой вывод теории интегральной информации, в соответствии с которым архитектура современных компьютеров не может быть сознательной, потому что то, как соединяются их логические гейты, дает слишком низкий уровень интеграции^[119]. Другими словами, если вы в будущем зальетесь в мощного робота, который точно имитирует каждый ваш нейрон и каждый синапс, и даже если этот цифровой клон будет выглядеть, ходить, говорить и вести себя не так как вы, то все равно, как утверждает Тонони, он будет бессознательным зомби без субъективных ощущений, и все ваши попытки найти субъективное бессмертие закончатся чистым разочарованием^[62]. Это утверждение было оспорено Дэвидом Чалмерсом и профессором Мюрреем Шанаханом, специалистом в области искусственного интеллекта, которые предложили рассмотреть, что будет происходить, если вы двинетесь в противоположном направлении и будете постепенно заменять нейронные цепи в своем мозгу гипотетическим цифровым оборудованием, идеально их симулирующим^[120]. Хотя, если мы поверим Тонони, замены никак не затронут ваше *поведение*, поскольку, по предположению, симуляция идеальна, ваши ощущения должны измениться от осознанных в начале к неосознанным в конце. Но что будет происходить в промежутке, по мере того как все больше и больше частей подвергаются замене? Когда части вашего мозга, отвечающие за осознанное восприятие верхней половины поля зрения, будут заменены, заметите ли вы, что часть вашего визуального пейзажа внезапно пропадет, при том что вы таинственным образом все равно знаете, что там происходит, как утверждают

пациенты со “слепыми пятнами” в поле зрения^[121]? Это вызвало бы огромные сложности, потому что если вы осознаете разницу, то вы можете сказать о ней своим друзьям, когда они об этом спросят, что противоречило бы предположению об отсутствии каких-либо изменений в вашем поведении. Единственная логически непротиворечивая возможность, совместимая с этим предположением, заключается в том, что в точности в тот самый момент, когда какая-то вещь исчезает из вашего сознания, ваш ум таинственно изменяется либо так, чтобы вынудить вас лгать, отрицая изменения чувственного восприятия, либо так, чтобы вы вдруг сразу забыли, как все выглядело раньше.

С другой стороны, Мюррей Шанахан признает, что то же самое рассуждение, использующее тактику постепенной замены, может быть применено к любой теории, утверждающей, что вы можете действовать сознательно, не будучи в сознании, поэтому возникает большой соблазн сделать вывод, что действовать сознательно и быть в сознании – одно и то же и что поэтому только наблюдаемое извне поведение имеет значение. Но тогда вы попадаете в ловушку: теория заставляет вас принять ее предсказание, что во сне вы должны быть без сознания, даже если вам самому об этом известно лучше.

Третья волна споров, вызванных теорией интегральной информации, поднялась в связи с вопросом, может ли сознательная сущность быть построена из частей, каждая из которых по отдельности тоже сознательна. Например, может ли общество в целом обрести сознание, так чтобы составляющие его люди не лишились своего? Может ли обладающий сознанием мозг состоять из частей, каждая из которых тоже сама по себе обладает сознанием? Теория интегральной информации отвечает категорическим “нет”, но это не всех убеждает. Например, некоторые пациенты, у которых повреждения резко снижают возможность взаимодействия между полушариями, страдают от синдрома чужой руки, когда правое полушарие их мозга вынуждает левую руку делать что-то такое, чего они сами не хотят и даже не понимают – иногда до такой степени, что используют другую руку, чтобы удерживать ею “чужую”. Как мы можем быть уверены, что в них нет двух отдельных сознаний, одно из которых размещено в правом полушарии, не способном говорить, а другое в левом полушарии, контролирующем речь и утверждающем, что контролирует и оба полушария? Представьте себе будущую технологию, устанавливающую прямую связь между мозгом одного человека и мозгом другого, и представьте себе, что она совершенствуется до тех пор, пока не станет возможной коммуникация между двумя полушариями одного мозга. Наступит ли такой момент, когда два отдельных сознания внезапно исчезнут и вместо них появится одно общее, как предсказывает теория интегральной информации, или переход будет проходить постепенно, так, что индивидуальные сознания будут сосуществовать в каком-то виде и даже формировать совместные переживания?

Еще одна увлекательная дискуссия развернулась вокруг подозрения, что в экспериментах недооценивается истинный уровень нашего осознания. Мы уже видели, что как бы мы ни были уверены в том, что полностью отдаем себе отчет относительно всех свойств окружающих нас предметов – их цвета, формы, взаимного расположения – в то время, когда смотрим на них, эксперименты обнаруживают нашу весьма незначительную способность все их запомнить и потом о них отчитаться^[122]. Некоторые исследователи пытались устранить это несоответствие, предполагая, что иногда у нас может быть “сознание без доступа”, то есть субъективное восприятие вещей, которые оказываются слишком сложными для сознания, чтобы вписаться в рабочую память и использоваться в дальнейшем^[123]. Например, когда у вас случается “слепота от невнимательности”: вы слишком отвлечены на другое, чтобы заметить предмет, который лежит прямо перед вами на виду, это вовсе не означает, что у вас нет осознанного визуального восприятия, просто образ этого предмета не сохранился в вашей рабочей памяти^[124]. Не следует ли отнести к этому скорее как к забывчивости, а не как к слепоте? Другие исследователи не согласны с тем, что мы можем не доверять рассказам людей об их чувственном опыте,

предупреждая о последствиях такого недоверия. Мюррей Шанахан смоделировал клиническое исследование, в ходе которого пациенты сообщают о полном избавлении от боли благодаря новому чудо-лекарству, которое, тем не менее, отклоняется правительственной комиссией: “Пациентам только кажется, что им не больно. Благодаря достижениям нейробиологии мы знаем лучше”^{125}. Но с другой стороны, известны случаи, когда пациенты случайно просыпались во время операции и им давали сильнодействующие препараты, чтобы заставить забыть о пережитом. Должны ли мы и в этом случае доверять их сообщениям, что они не испытывали никакой боли^{126}?

Что может чувствовать искусственный интеллект, обладая сознанием?

Если какая-то будущая система с искусственным интеллектом окажется обладающей сознанием, каковы могут быть ее субъективные переживания? В этом вопросе суть “еще более трудной проблемы” сознания, и это заставляет нас переместиться на второй уровень сложности (рис. 8.1). Мы не просто лишены в настоящее время какой-либо теории, позволяющей ответить на этот вопрос, но мы вообще не уверены в существовании логической возможности полностью на него ответить. В конце концов, как мог бы звучать удовлетворительный ответ? Как вы объясните слепому от рождения человеку, на что похож красный цвет?

К счастью, наша нынешняя неспособность дать полный ответ не мешает нам давать частичные ответы. Разумные инопланетяне, изучая органы чувств человека, придут, вероятно, к заключению, что цвета – это такие квалиа, которые воспринимаются связанными с каждой точкой на плоской поверхности (нашего поля зрения), в то время как звуки не воспринимаются пространственно локализованными, а боль – это квалиа, которая воспринимается привязанной к определенным частям нашего тела. Открыв, что на нашей сетчатке располагаются три типа светочувствительных колбочек, они могут прийти к заключению, что мы воспринимаем только три основных цвета и что все другие цветовые квалиа возникают в результате их сочетания. Измерив, сколько времени требуется нейронам для передачи информации через мозг, они могут заключить, что мы способны переживать не более десяти осознанных мыслей или чувственных переживаний в секунду и что, когда мы смотрим по телевизору фильмы, кадры которых сменяют друг друга по 24 раза в секунду, мы не видим каждый кадр в отдельности как неподвижную картинку, но складываем их в непрерывное движение. Измерив скорость выделения адреналина у нас в крови и время, которое он способен оставаться там не разлагаясь, они могли предсказать, что мы способны за считанные секунды впадать в гнев, испытывая приступы длительностью в несколько минут.

При помощи аналогичных, основанных на физике соображений мы можем построить некоторые ученые догадки относительно определенных аспектов самочувствия искусственного сознания. Прежде всего, пространство возможных переживаний искусственного интеллекта несопоставимо огромно в сравнении с тем, что можем испытывать мы, люди. У нас есть по одному классу квалиа для каждого из наших чувств, а у искусственного интеллекта может быть значительно больше типов датчиков и способов внутреннего представления информации, поэтому мы не должны попадать в ловушку, предполагая, что чувствовать себя искусственным интеллектом – это примерно то же самое, что чувствовать себя человеком.

Во-вторых, искусственное сознание, упакованное в коробочку размером с мозг человека, может испытывать в миллионы раз больше переживаний в секунду, чем мы, так как электромагнитные сигналы распространяются со скоростью света, – то есть в миллионы раз быстрее, чем распространяются сигналы вдоль нейронов. Однако чем больше искусственный интеллект, тем медленнее должны быть его глобальные мысли, чтобы информации хватило времени достичь каждой из его частей, как мы видели в главе 4. Поэтому мы ожидаем, что у искусственного интеллекта “Gaia” размером с Землю будет только около десяти осмысленных переживаний в секунду, как у человека, а искусственный интеллект размером с галактику может пронзить лишь одна глобальная мысль в 100 000 лет, – то есть не более сотни переживаний за все время существования нашей Вселенной до настоящего времени! Поэтому, казалось бы, у большого искусственного интеллекта должен быть мощный стимул делегировать вычисления маленьким подсистемам, способным выполнять их, чтобы не терять времени, – подобно тому, как наше сознание делегировало моргательный рефлекс в маленькую подсистему,

быструю и бессознательную. Хотя мы уже видели, что сознательная обработка информации в нашем мозгу – это всего лишь верхушка нашего в общем-то бессознательного айсберга, мы должны ожидать еще больших крайностей у больших интеллектуальных систем будущего: представляя собой единое сознание, они, вероятно, должны будут оставить без внимания почти всю обработку информации, происходящую внутри них. Кроме того, хотя осознанные переживания, которые испытывает такая система, могут быть чрезвычайно сложными, они при этом развиваются черепашими шагами в сравнении с быстрой деятельностью ее меньших частей.

Неудивительно, что в голову в связи с этим приходит упоминавшаяся выше полемика по поводу того, могут ли части сознательной сущности также обладать сознанием. Теория интегральной информации отрицает такую возможность, и это означает, что если какая-то будущая астрономически большая система с искусственным интеллектом будет обладать сознанием, то почти вся обработка информации внутри нее будет проходить бессознательно. Отсюда следует, что если цивилизация меньших сущностей с искусственным интеллектом в процессе улучшения своих коммуникативных способностей придет к точке, когда возникает единый коллективный обладающий сознанием ум, каждое из их намного более быстрых индивидуальных сознаний внезапно угаснет. Если предсказание теории интегральной информации ошибочно, то, с другой стороны, коллективный разум сможет сосуществовать с целой армией обладающих сознанием меньших умов. Тогда мы даже можем себе представить иерархию вложенных друг в друга сознаний всех уровней, от микроскопического до космического.

Как мы видели выше, бессознательная обработка информации в нашем человеческом мозгу представляется связанной с легким, быстрым и автоматическим способом мышления, что психологи называют “системой 1”^[127]. Например, система 1 может сообщить вашему сознанию, что проведенный ею сложный анализ визуальной информации, полученной от органов чувств, выявил в приближающемся человеке вашего лучшего друга. При этом она никоим образом не сообщает, какие вычисления ею проводились. Если такая связь между системами и сознанием подтвердится, то возникнет соблазн распространить эту терминологию на искусственный интеллект, называя системой 1 все быстрые рутинные задачи, делегированные бессознательным модулям. Трудоемкая, медленная, требующая усилий и глобального контроля со стороны всего искусственного интеллекта – это система 2. У нас, людей, также есть осознанные переживания, связанные с тем, что я называю “системой 0”: никак не обрабатываемое пассивное восприятие происходит вокруг вас, когда вы сидите не двигаясь и даже не думаете, а просто наблюдаете. Системы 0, 1 и 2 кажутся непрерывно усложняющимися, и даже поразительно, что только та, что посередине, представляется бессознательной. Теория интегральной информации объясняет это тем, что пассивная сенсорная информация в системе 0 хранится в решетчатых структурах с очень высокой интеграцией, в то время как система 2 имеет высокий уровень интеграции из-за петель обратной связи, где вся информация, которой вы располагаете сейчас, может повлиять на будущие состояния вашего мозга. С другой стороны, именно предсказание сознания в решетчатых структурах вызвало критику Скотта Ааронсона в эпизодах, о которых шла речь выше. Подводя итог, можно сказать, что если теория, дающая решение довольно трудной проблемы сознания, сможет в один прекрасный день пройти через строй строгих экспериментальных проверок так, что мы начнем серьезно относиться к ее предсказаниям, то тем самым будет значительно сужен круг возможных вариантов решения еще более сложной проблемы – относительно того, что может испытывать в будущем обладающий сознанием искусственный интеллект.

Некоторые аспекты нашего субъективного опыта позволяют проследить весь путь нашей эволюции, например связать наши эмоциональные порывы с требованиями самосохранения (потребность в еде и питье, уклонение от опасности быть убитым) и размножения. Это означает, что, вероятно, возможно создать такой искусственный

интеллект, которому не знакомы такие квалиа, как голод, жажда, страх или сексуальное влечение. Как мы видели в предыдущей главе, если очень умный искусственный интеллект запрограммирован так, чтобы он стремился к достижению практически любой достаточно амбициозной цели, для него весьма вероятно и стремление к самосохранению – для того, чтобы иметь возможность этих целей достигать. Если искусственный интеллект существует в сообществе аналогичных систем, у них может отсутствовать присущий человеку сильный страх смерти: делая регулярные резервные копии, максимум, что они могут потерять, это содержимое памяти, накопленное с момента последнего резервного копирования, при условии, что сохраненные копии пойдут в дело. Кроме того, возможность легко копировать информацию и программное обеспечение между различными системами искусственного интеллекта, вероятно, снизит остроту чувства индивидуальности, так свойственного нашему человеческому сознанию: между мной и вами было бы меньше различий, если бы мы могли легко делиться своими воспоминаниями и копировать свои способности, поэтому группа связанных друг с другом интеллектуальных систем может чувствовать себя единым организмом с коллективным разумом.

Будет ли искусственное сознание считать, что обладает свободой воли? Обратите внимание, что, хотя философы потратили тысячелетия на препирательства на тему, есть ли у нас свобода воли, и не приблизились к согласию даже по поводу того, как правильно поставить вопрос^{128}, я ставлю другой вопрос, на который, полагаю, значительно легче ответить. Позвольте мне попытаться убедить вас, что ответ на него прост: “Да, всякий, кто сознательно принимает решения, субъективно чувствует, что обладает свободой воли, независимо от его биологического или искусственного происхождения”. Спектр решений заполняет промежуток между двумя крайностями:

- 1) Вы знаете точно, почему делаете именно этот выбор.
- 2) Вы не знаете, почему сделали именно такой выбор – кажется, это был просто случайный каприз.

Дискуссии по поводу свободы воли обычно вращаются вокруг попыток примирить наше поведение, нацеленное на принятие некоторых решений для достижения определенных целей, с законами физики. И если вам надо выбрать одно из двух приведенных ниже объяснений ваших поступков, то какое из них правильное: “Я пригласил ее на свидание, потому что она действительно мне нравится” или “Элементарные частицы, из которых состоит мое тело, вынудили сделать это, двигаясь в соответствии с законами физики”? Но в предыдущей главе мы уже увидели, что годятся оба: то, что ощущается как целенаправленное поведение, может возникать из не знающих целей детерминистских законов физики. Говоря более конкретно, когда какая-то система (головной мозг или искусственный интеллект) принимает решение типа 1, она вычисляет его, используя некоторый детерминистский алгоритм, и причина, по которой она чувствует, что приняла решение, заключается в том, что она и в самом деле приняла решение, когда производила вычисления. Кроме того, как подчеркивал Сет Ллойд^{129}, в информатике есть знаменитая теорема, которая утверждает, что почти для всех вычислений не существует более быстрого способа определить их результат, чем просто провести эти вычисления. А значит, вы обычно не можете узнать, что вы решите делать через секунду, быстрее, чем через секунду, и это помогает вам укрепить свою уверенность в свободе своей воли. Напротив, когда система (головной мозг или искусственный интеллект) принимает решение типа 2, она просто приходит к мнению, что надо основывать его на результате вычислений в какой-то подсистеме, которая в данном случае действует как генератор случайных чисел. И в мозгу, и в компьютере эффективная замена случайным числам легко создается путем усиления шума. Независимо от того, на какую часть спектра между типом 1 и типом 2 приходится решение, как биологическое, так и искусственное сознание придут к заключению, что их воля свободна: они чувствуют, что

это именно они принимают решение, и они не в состоянии предсказать с уверенностью, каким оно будет, до того момента, пока не закончат его продумывать.

Некоторые люди говорят мне, что причинность их унижает, что она лишает смысла их мыслительные процессы и превращает их “просто” в машины. Я нахожу эти жалобы необоснованными и абсурдными. Во-первых, нет ничего “простого” в человеческом мозге, который, насколько я могу судить, представляет собой наиболее сложный тип физического объекта в нашей Вселенной. Во-вторых, какие альтернативы они бы предпочли? Неужели они хотят, чтобы решения принимались их собственными мыслительными процессами (вычислениями, выполняемыми в их мозгу)? Их субъективное ощущение свободы воли – это именно то, как их вычисления ощущаются изнутри: они не знают результата вычислений до тех пор, пока вычисление не завершено. То есть это то же самое, что сказать: вычисление и есть решение.

Давайте вернемся к исходной точке этой книги: каким бы вы хотели видеть будущее жизни? В предыдущей главе вы рассмотрели, как различные культуры по всему миру стремятся к будущему полному положительных ощущений, но в поисках определения, что же считать положительным и как согласовать все то, что было бы полезно для разных форм жизни, возникают жесточайшие споры. Давайте же не позволим всем этим спорам отвлекать нас от слона в комнате: не может быть никаких положительных ощущений, когда нет никаких ощущений вовсе, то есть когда нет никакого сознания. Другими словами, без сознания не будет ни счастья, ни добра, ни красоты, ни смысла или цели – только пропащее астрономическое пространство. А значит, когда люди спрашивают о смысле жизни так, словно это дело космоса – сообщать смысл нашему существованию, они выворачивают суть наизнанку: *не наша Вселенная сообщает смысл сознательным существам, но сознательные существа сообщают смысл нашей Вселенной*. А значит, первым пунктом в нашем списке благих пожеланий на будущее должно стоять сохранение (а по возможности и расширение) биологического и/или искусственного сознания в нашем космосе, а не борьба за его истребление.

Если мы преуспеем в этом деле, то как мы, люди, будем чувствовать себя в условиях сосуществования со все более умными машинами? Прогресс искусственного интеллекта, представляющийся неумолимым, беспокоит ли он вас, и если да, то почему? В главе 3 мы видели, как относительно легко может вооруженная искусственным интеллектом технология удовлетворить большинство базовых нужд, касающихся дохода и безопасности, при наличии на то политической воли. Однако быть хорошо накормленным и одетым, иметь уютное жилище и возможность предаваться развлечениям – это еще не все. Если нам будет гарантировано, что искусственный интеллект позаботится обо всех наших практических нуждах и пожеланиях, не будем ли мы, тем не менее, страдать от отсутствия смысла и цели в нашей жизни, как ухоженные животные в зоопарке?

Для нас, людей, типично исходить в своей самооценке из идеи *человеческой исключительности*, из представления о том, что мы – самые умные существа на планете и оттого уникальные и превосходные. Появление искусственного интеллекта вынудит нас отказаться от этой мысли и стать поскромнее. Но, наверное, нам придется это сделать в любом случае: в конце концов, высокомерная уверенность в своем превосходстве над другими (личностями, этническими группами, биологическими видами) уже приводила к ужасным проблемам в прошлом, и пришла пора от нее отказаться. В самом деле, идея человеческой исключительности не только становилась причиной бед, но и кажется совершенно ненужной для процветания: если нам суждено открыть мирную внеземную цивилизацию, далеко превзошедшую нас в науках, искусствах и во всем остальном, что для нас важно, это, наверное, не сможет помешать людям продолжать находить в своих жизнях смысл и ставить перед собой цели. Мы можем сохранять верность своим семьям, друзьям и более многочисленным сообществам и продолжать заниматься всеми теми делами, которыми считали нужным и важным заниматься и раньше, не потеряв ничего, кроме своей заносчивости.

Планируя наше будущее, давайте подумаем о смысле не только для своих жизней, но и для самой Вселенной. В этом отношении два моих любимых физика, Стивен Вайнберг и Фримен Дайсон, представляют два диаметрально противоположных взгляда. Стало знаменитым высказывание Вайнберга, удостоенного Нобелевской премии за важнейший вклад в создание стандартной модели элементарных частиц: «Чем лучше мы понимаем Вселенную, тем бессмысленнее она кажется»^{130}. Дайсон, со своей стороны, гораздо оптимистичнее, как мы могли убедиться в главе 6: хотя он тоже считает, что Вселенная *была* бессмысленной, она стала все больше и больше наполняться смыслом

с возникновением жизни, но вершина осмысленности все еще впереди – когда жизнь сможет распространиться в космосе. Он закончил свою программную статью 1979 года так: “Какая из Вселенных, моя или Вайнберга, ближе к истине? Когда-нибудь, и не так уж не скоро, мы узнаем”⁽¹³¹⁾. Если когда-нибудь наша Вселенная навсегда вернется к своему бессознательному состоянию из-за того, что мы истребим жизнь на Земле, или из-за того, что мы позволим бессознательным зомби с искусственным интеллектом захватить нашу Вселенную, то Вайнберг возьмет верх окончательно и бесповоротно.

С этой точки зрения, хотя наше внимание в этой книге по большей части было обращено к будущему разума, будущее сознания куда как более важно, потому что это то, что создает смыслы. Когда речь заходит об этом различии, философы любят переходить на латынь, противопоставляя *sapientia* (способность разумно мыслить) и *sentientia* (способность к субъективному восприятию квалиа). Мы, люди, выстроили свою идентичность на идее принадлежности к виду *Homo sapiens* – самых разумных существ в природе. Сейчас мы готовимся к встрече с более разумными машинами и потому должны быть скромнее. Я предлагаю людям пройти ребрендинг и стать *Homo sentiens*.

Подведение итогов

- Нет общепринятого и не вызывающего споров определения сознания. Я использую достаточно широкое и не антропоцентрическое: *сознание = субъективные переживания*.

- Сознательны ли системы искусственного интеллекта в этом смысле – это принципиально важно для ответа на самые острые вопросы, возникающие в связи с развитием искусственного интеллекта вообще. Могут ли такие системы страдать? Могут ли они иметь права? Не окажется ли заливка человеческого разума в компьютерную память самоубийством? Не станет ли космос будущего, кишаший системами искусственного интеллекта, зомби-апокалипсисом?

- Проблема понимания разума не должна смешиваться с тремя различными проблемами понимания сознания: “довольно трудной проблемой” определения, какая физическая система обладает сознанием, “еще более трудной проблемой” предсказания квалиа и “по-настоящему трудной проблемой” – почему что-то вообще оказывается сознательным.

- “Довольно трудная проблема” научна, так как теория, предсказывающая, какие из процессов, протекающих у вас в мозгу, вами осознаются, может быть проверена экспериментально и фальсифицирована, притом что пока остается неясным, как наука может разрешить две более сложные проблемы.

- Нейробиологические эксперименты показывают, что многие действия и связанные с ними области мозга не осознаются, а большая часть наших осознанных переживаний складывается из апостериорного суммирования значительного количества бессознательной информации.

- Для того чтобы предсказывать появление сознания у машин, требуется теория. Сознание, по-видимому, связано не с какой-то особой разновидностью полей или частиц, а с особой разновидностью информационных процессов, в большой степени автономных и интегральных, так что вся система в целом автономна, но каждая ее часть в отдельности – нет.

- Сознание ощущается исключительно нефизическим потому, что оказывается дважды независимым от субстрата: если сознание – это самоощущение, возникающее при обработке информации определенным образом, то значение имеет только структура собственно информационного процесса, но не структура вещества, обеспечивающего этот процесс.

- Если искусственное сознание возможно, то пространство возможных переживаний искусственного интеллекта в сравнении с человеческим должно быть несопоставимо велико, охватывая широчайший спектр квалиа и временных масштабов – включая ощущение свободы воли.

- Поскольку без сознания невозможен никакой смысл, то не наша Вселенная дает смысл сознающим существам, а сознающие существа дают смысл нашей Вселенной.

- Так как людям предстоит уступить место умнейших существ во Вселенной машинам, мы должны свыкнуться с новой ролью – *Homo sentiens*, а не *Homo sapiens*.

Эпилог

Сказание о команде FLI

Самая печальная сторона жизни сейчас заключается в том, что наука обретает знания быстрее, чем общество обретает мудрость.

Айзек Азимов

Итак, мой дорогой читатель, мы добрались до конца книги, изучив происхождение и судьбы разума, целей и смысла. Так как же воплотить эти идеи в действия? Что конкретно мы должны сделать, чтобы наше будущее было как можно лучше? Именно этот вопрос я задаю себе прямо сейчас, сидя у окна, по пути из Сан-Франциско в Бостон 9 января 2017 года, возвращаясь с конференции по искусственному интеллекту, которую мы только что провели в Асиломаре, поэтому позвольте мне закончить эту книгу, делясь с вами своими впечатлениями.

Мейя задремала в соседнем кресле, утомленная бессонными ночами, наполненными хлопотами и приготовлениями. О, что за неделька это была! Нам удалось на несколько дней собрать вместе на сиквел в Пуэрто-Рико почти всех, упоминавшихся в этой книге, включая Илона Маска, Ларри Пейджа, ведущих исследователей по искусственному интеллекту из университетов и компаний, таких как DeepMind, Google, Facebook, Apple, IBM, Microsoft и Baidu, а также многих экономистов, юристов, философов и других блестящих мыслителей (см. рис. 9.1). Результат превзошел мои самые смелые ожидания, и я пребываю на вершине оптимизма, какого не знал долгое время, по поводу будущего жизни. Здесь, в эпилоге, я хочу объяснить вам почему.

Рождение FLI

С тех пор как в возрасте четырнадцати лет я узнал о гонке ядерных вооружений, меня тревожило, что мощь наших технологий растет быстрее, чем мудрость, с которой мы управляем ими. Поэтому я решил вставить главу об этой проблеме в мою первую книгу *Our Mathematical Universe*, несмотря на то, что в основном она была о физике. К началу 2014 года я дал новогоднее обещание, что больше не стану жаловаться ни на что, не подумав сначала самым серьезным образом над тем, как лично я могу что-то сделать ради исправления ситуации, и я держал свое обещание, пока ездил, представляя книгу, на протяжении всего января: мы с Мейей провели в это время не один мозговой штурм, решая, как нам создать некоммерческую организацию, служащую для всемерного улучшения будущего жизни путем адекватного управления технологической средой.

Мейя настаивала на том, что мы должны дать этой организации какое-то оптимистическое название, как можно больше отличающееся от чего-нибудь вроде “Институт кошмара и ужаса” или “Давайте побеспокоимся о будущем”. Так как название “Институт будущего человечества” было уже занято, мы сошлись на “Институте будущего жизни” (Future of Life Institute, FLI), у которого было дополнительное преимущество, заключающееся в возможности расширительных толкований. 22 января мы должны были проводить презентацию книги в Санта-Крузе, и во время ужина с нашим старым другом Энтони Агирре, глядя, как калифорнийское солнце опускается в Тихий океан, мы уговорили его объединить наши силы. Он не только один из умнейших людей, кого я знаю, и всегда полон идей, но у него уже есть опыт ведения дел одной из некоммерческих организаций, Института основополагающих вопросов (Foundational Questions Institute, <http://fqxi.org>), созданного нами более десятилетия назад.

На следующей неделе тот же тур привел меня в Лондон. Поскольку будущее искусственного интеллекта владело моим сознанием, я связался с Демисом Хассабисом, который любезно пригласил меня посетить штаб-квартиру DeepMind. Я был потрясен, увидев, насколько они выросли с того времени, как посещали меня в MIT двумя годами раньше. Google только что купил их за 650 миллионов долларов, и, оценивая впечатляющий интеллектуальный пейзаж их офиса, наполненного самыми блестящими умами, которые устремились к дерзкой цели Хассабиса “разрешить задачу разума”, я преисполнился висцеральным предчувствием реально возможного успеха.

Следующим вечером я говорил со своим другом Яном Таллином по Skype, созданному при его участии. Я объяснял ему задачи FLI, а час спустя он решил дать нам шанс, вкладывая в наш фонд по 100 000 долларов ежегодно! Ничто не трогает меня больше, чем доверие, оказываемое мне в пределах, которых я, как мне кажется, не заслужил. И мне было особенно приятно, когда год спустя, на конференции в Пуэрто-Рико, о которой шла речь в главе 1, он пошутил, что никогда еще не делал таких удачных инвестиций.

Следующий день мой издатель оставил свободным, и я решил посвятить его посещению Лондонского музея науки. Будучи так долго одержим идеей прошлого и будущего разума, я вдруг оказался посреди физических иллюстраций своих мыслей. Они собрали фантастическую коллекцию, свидетельствующую о росте нашего знания, от паровоза “Ракета” Роберта Стефенсона до автомобиля Форд-Т, копии в натуральную величину лунного посадочного модуля Аполлона-11 и компьютеров – от разностной машины Чарльза Бэббиджа до современных моделей. Здесь были представлены также эксперименты, посвященные выяснению природы разума: и реплика опыта Гальвани с лягушачьими лапками, и электроэнцефалограф, и оборудование для функциональной магнитно-резонансной томографии.



Рис. 9.1

Наша конференция в Асиломаре стала сиквелом пуэрто-риканской. Она проходила в январе 2017 года и собрала вместе многих исследователей искусственного интеллекта и представителей других, смежных с этой, исследовательских областей. В верхнем ряду стоят слева направо: Патрик Лин, Дэниел Вельд, Эриел Конн, Нэнси Чан, Том Митчел, Рэй Курцвейл, Дэниел Дьюи, Маргарет Боден, Питер Норвиг, Ник Хэй, Моше Варди, Скотт Зискинд, Ник Бострём, Франческа Росси, Шейн Легг, Мануэла Велозо, Дэвид Марбл, Катя Грейс, Ираклий Беридзе, Марти Тененбаум, Жиль Пратт, Мартин Рис, Джошуа Грин, Мэтт Шэрер, Энджела Кейн, Амара Анхелика, Джефф Мор, Мустафа Сулейман, Стив Омохундро, Кейт Кроуфорд, Виталик Бутерин, Ютака Мацуо, Стефано Эрмон, Майкл Веллман, Бас Стойнебринк, Вендель Валлах, Аллен Дафо, Тоби Орд, Томас Диттерих, Дэниел Канеман, Дарио Амодей, Эрик Дрекслер, Томазо Поджо, Эрик Шмидт, Педро Ортега, Дэвил Лик, Шон О'х Эйгертейг, Оувейн Эванс, Яан Таллин, Анка Драган, Шон Легассик, Тоби Уолш, Питер Асаро, Кей Фирт-Баттерфилд, Филип Сабеш, Пол Меройа, Барт Зельман, Такер Дэви, Джейкоб Стейнхардт, Моше Лукс, Джош Тененбаум, Том Грубер, Эндрю Ын, Карэм Аюб, Крейг Кэлхун, Перси Лян, Хелен Тонер, Дэвид Чалмерс, Ричард Саттон, Клаудиа Пассос-Феррейра, Янош Крамар, Уильям Макаскил, Элиезер Юдковски, Брайан Цибарт, Хью Прайс, Карл Шульман, Нил Лоренс, Ричард Маллах, Юрген Шмидхубер, Дилип Джордж, Джонатан Ротберг, Ноа Ротберг; В переднем ряду: Энтони Агирре, Соня Закс, Лукас Перри, Джеффри Закс, Винцент Коницер, Стив Гуз, Виктория Краковна, Оуэн Коттон-Барратт, Даниела Рус, Дилан Хадфилд-Менель, Верити Хардинг, Шивон Зилис, Лоран Орсо, Рамана Кумар, Нейт Соареш, Эндрю Макафи, Джек Кларк, Анна Саламон, Лун Оуян, Эндрю Крич, Пол Кристиано, Иошуа Бенджо, Дэвид Сэнфорд, Катрин Олссон, Джессика Тэйлор, Мартина Кунц, Кристинн Ториссон, Стюарт Армстронг, Янн Лекюн, Александр Тамас, Роман Ямпольский, Марин Солачич, Лоуренс Краусс, Стюарт Рассел, Эрик Бриньоулфссон, Райан Кало, Сяолань Сюэ, Мейя Чита-Тегмарк, Кент Уокер, Хизер Рофф, Мередит Уиттейкер, Макс Тегмарк, Адриан Веллер, Хосе Эрнандес-Оральо, Эндрю Майнард, Джон Херинг, Абрам Демски, Николас Берггруен, Грегори Боннет, Сэм Харрис, Тим Хван, Эндрю Снайдер-Бити, Марта Халина, Себастьян Фаркар, Стивен Кейв, Ян Лейке, Таша Мак-Коли, Джозеф Гордон-Левитт.

Позже также прибыли: Гуру Банавар, Демис Хассабис, Рао Кабхампати, Илон Маск, Ларри Пейдж, Энтони Ромеро.

Я очень редко плачу, но, выходя из музея, я не мог сдержаться, и уже в метро, двигаясь по туннелю, полному пешеходов, к станции Южный Кенсингтон, я не мог остановиться. Мимо шли люди, спешившие по своим делам, в блаженном неведении о том, чем были наполнены мои мысли. Сначала мы, люди, обнаружили, как можно скопировать некоторые естественные процессы при помощи машин, создавая свой собственный ветер и молнии и свою собственную механическую лошадиную силу. Постепенно мы начали понимать, что наши тела – это тоже машины. А потом открытие нервных клеток положило начало размыванию грани между телом и разумом. Тогда мы начали строить машины, которые могли бы превзойти нас не только по части мускулатуры,

но и по части разума. И поэтому параллельно с постижением собственной природы мы неизбежно устареваем? В этом есть что-то, наполненное поэзией трагедии.

Эта мысль пугала меня, но в то же время укрепляла в стремлении следовать своему новогоднему обету. Я чувствовал, что в нашей команде не хватает еще одного члена, который мог бы возглавить компанию волонтеров-идеалистов. Логичным образом выбор пал на Викторию Краковну, блестящую магистрантку из Гарварда, только что получившую серебряную медаль на Международной олимпиаде по математике и основавшую Цитадель, которая объединила дюжину юных идеалистов, мечтающих играть более важную роль не только в своей собственной жизни, но и в жизни всего мира. Мы пригласили ее к себе пять дней спустя и рассказали о наших планах. И прежде, чем наши суши пришли к финишу, FLI был рожден.

Пуэрто-риканское приключение

Рождение института было ознаменовано началом удивительного приключения, которое продолжается до сих пор. Как уже рассказывалось в главе 1, мы стали регулярно собирать у себя дома по несколько десятков идеалистически настроенных студентов, профессоров и других мыслящих людей из числа наших соседей. Лучшие идеи становились проектами – первым среди них была та самая колонка об искусственном интеллекте, написанная вместе со Стивеном Хокингом, Стюартом Расселом и Фрэнком Вильчеком, о которой я рассказал в главе 1 и которая помогла разжечь публичные дебаты. Параллельно с первыми шагами, необходимыми при создании новой организации (регистрация, найм консультантов и запуск веб-сайта), мы провели веселую презентацию, посвященную рождению института, в переполненной аудитории MIT, в ходе которой Алан Алда побеседовал о технологиях будущего с ведущими экспертами.



Рис. 9.2

Яан Таллин, Энтони Агирре, ваш покорный слуга, Мейя Чита-Тегмар и Виктория Краковна отмечают создание FLI 23 мая 2014 года.

Всю остальную часть года мы посвятили организации конференции в Пуэрто-Рико, которая, как говорилось в главе 1, ставила своей целью привлечение ведущих мировых экспертов в области искусственного интеллекта к дискуссии о том, как обеспечить его дружелюбность. Мы стремились главным образом к тому, чтобы в разговорах о безопасности центр тяжести был перенесен с предупреждений об угрозах на привлечение к работе: надо было прекращать пререкаться о том, по какому поводу следует больше всего беспокоиться, и искать области общего согласия, дающие возможности объединять усилия в конкретных исследовательских проектах, в которых шанс положительного результата был бы максимален. Для начала мы собирали наиболее перспективные идеи в исследованиях безопасности искусственного интеллекта по всему миру и просили наше сообщество присылать нам свои отклики на те из них, которые мы включали в наш растущий список. С помощью Стюарта Рассела и группы молодых трудолюбивых волонтеров, среди которых выделились Дэниел Дьюи, Янош Крамар и Ричард Маллах, мы определили среди них исследовательские приоритеты, собрав их в

отдельный документ, который вынесли на обсуждение в ходе пленарного заседания конференции^[132]. Достигнув общего согласия относительно наиболее ценных, достойных проведения исследований, мы надеялись поощрить людей к тому, чтобы действительно провести такие исследования. Особенно желательным было бы для нас, если бы нам удалось убедить кого-нибудь поддерживать такие исследования финансово, так как пока у нас не было фактически никакой поддержки со стороны правительственных финансовых учреждений.

Появился Илон Маск. 2 августа наши радары засекли его знаменитый твит “Стоит почитать Бострёма о сверхразуме. Нам надо быть предельно осторожными в отношении искусственного интеллекта. Потенциально эта штука опаснее бомбы”. Я связался с ним, чтобы рассказать о наших усилиях, а через несколько недель мне довелось поговорить с ним по телефону еще раз. Хотя я очень нервничал и его звездность меня подавляла, результат оказался потрясающим: он согласился войти в научный совет FLI, принять участие в нашей конференции и даже, может быть, профинансировать первую исследовательскую программу, посвященную проблемам безопасности искусственного интеллекта, запуск которой планировалось анонсировать на пуэрто-риканской конференции. Это наэлектризовало всех в институте, мы удвоили наши усилия, чтобы провести достойную конференцию, выявить на ней актуальные исследовательские направления и выстроить поддерживающее их сообщество.

В конце концов я встретился с Илоном Маском и лично, когда он два месяца спустя приехал в MIT для участия в симпозиуме по космонавтике. Очень странно было оказаться с ним лицом к лицу наедине в маленькой зеленой комнате всего через несколько мгновений после того, как он, словно рок-звезда, стоял, окруженной тысячами завывающих от восторга студентов MIT, но прошло еще несколько минут, и я уже не мог думать ни о чем, кроме нашего совместного проекта. Он мне сразу понравился. Я чувствовал излучаемую им искренность и был вдохновлен его неподдельной заботой о долгосрочном будущем человечества, а также тем, как решительно он превращал свои устремления в конкретные дела. Он хотел, чтобы люди продолжали исследовать Вселенную и осваивать ее. Поэтому он создал космическую компанию. Он поддерживал идею возобновляемой энергии, и поэтому начал производство солнечных батарей и электромобилей. Высокий, привлекательный внешне, красноречивый и невероятно просвещенный – совсем не удивительно, почему его так внимательно слушают.

К сожалению, в тот же раз я получил и неприятный урок: СМИ иногда могут сеять страх и раздор среди людей. Часовое публичное выступление Маска было посвящено увлекательнейшему рассказу об исследовании космоса, и я был убежден, что из него можно будет сделать потрясающее ТВ-шоу. В самом конце какой-то студент задал ему не имеющий отношения к теме выступления вопрос об искусственном интеллекте. Ответ Маска содержал фразу: “С искусственным интеллектом мы вызываем дьявола”, – и только об этом и сообщалось в СМИ. Причем практически вне всякого контекста. Меня поражало, как многие журналисты неосознанно делали *прямо противоположное* тому, чего мы пытались достичь нашей конференцией в Пуэрто-Рико. В то время как мы хотели создать в сообществе базовый консенсус, находя почву для объединения, СМИ стремились подчеркивать то, что нас разделяет. Чем больше они сообщали о наших разногласиях, тем более высокое место они занимали в рейтинге Нильсена и тем быстрее рос рекламный бюджет. Более того, при том что мы хотели помочь людям, придерживающимся противоположных мнений, объединиться, понять друг друга и начать действовать вместе, сообщения СМИ вызвали в людях с противоположными мнениями раздражение друг против друга, усугубляя их взаимное непонимание публикацией только самых провокационных высказываний, – всегда лишенных контекста. По этой причине мы забанили журналистов на своей конференции в Пуэрто-Рико и установили там правило Чатем-хауса, которое запрещало участникам конференции рассказывать во время и после нее кому бы то ни было кто что говорил^[63].

Хотя конференция в Пуэрто-Рико прошла успешно, нам этот успех дался совсем не легко. Пошедший обратный отсчет требовал скрупулезной подготовительной работы, включавшей, например, мои переговоры по телефону и Skype, чтобы собрать критическую массу экспертов, достаточную для притяжения всех остальных будущих участников; в ней были и свои драматические моменты – например, когда я встал в 7 утра 27 декабря, чтобы сквозь хрипы и треск ненадежного соединения услышать голос Маска, говорившего мне из Уругвая: “Я не думаю, что это может сработать”. Его беспокоило, что программа исследований по AI-безопасности может родить ложное чувство успокоенности, и безрассудные ученые будут заниматься своими разработками, отговариваясь от необходимых мер безопасности лишь пустыми словами. Но тогда, несмотря на то что связь то и дело прерывалась, мы смогли подробно поговорить о колоссальных преимуществах переноса этой тематики в центр нашего предметного поля и объединения всех, кто занимается искусственным интеллектом, вокруг проблем безопасности. Когда разговор окончательно прервался, он послал мне один из самых моих любимых мейлов: “Звонки до меня больше не проходят. Но все вроде бы решено. Я счастлив буду поддержать исследования 5 миллионами на ближайшие три года. Или все-таки надо десять?”.

Через четыре дня начался 2015 год, для нас с Мейей начало было хорошим: мы встречали Новый год, избавленные от приготовлений, танцуя на пуэрто-риканском пляже при свете фейерверков. Конференция тоже началась большим успехом: сложился примечательный консенсус участников по поводу того, что вопросы AI-безопасности требуют существенно большего внимания, и, опираясь на дальнейшее содействие участников конференции, мы смогли довести до окончательного совершенства итоговый документ об исследовательских приоритетах, над которым так много работали. Мы распространили среди участников проект коллективного письма из главы 1 о необходимости проводить исследования безопасности и были рады, что почти все участники конференции подписали его.

У нас с Мейей прошла очень вдохновившая нас беседа с Маском в нашей комнате в отеле, во время которой он благословил наш подробный план программы грантов. Мейю очень тронуло, насколько ясно и просто он рассказывал о своей частной жизни и какой интерес проявил к нашей. Он спросил нас о нашей встрече, и обстоятельный рассказ Мейи очень ему понравился. На следующий день мы записывали интервью с ним о AI-безопасности и о том, почему он поддерживает такие исследования. Все шло как по маслу^{133}.

Высшая точка нашего собрания – объявление о дотации Маска, оно было назначено на воскресенье, 4 января 2015 года, на 7 часов вечера. Я был так взволнован, что полночи проворочался без сна. А потом, всего за тринадцать минут до нашей предполагавшейся встречи, после чего мы должны были пойти на собрание, вдруг все рухнуло! Позвонил секретарь Маска и сказал, что Маск, похоже, не сможет прийти и сделать свое объявление. Мейя говорит, что она никогда не видела меня таким потрясенным и разочарованным. Но Маск все же пришел, и я слышал, как секунда отсчитывается за секундой, приближая начало собрания, пока мы сидели и разговаривали. Он объяснял, что до запуска SpaceX остается всего два дня, для них это жизненно важный этап, и они очень рассчитывают произвести первый в истории возврат отстрелившейся первой ступени на самоуправляемую платформу. Из-за этого команда SpaceX очень не хочет, чтобы внимание медиа отвлекалось на какие-то посторонние события с его участием. Энтони Агирре, хладнокровный и рассудительный как всегда, заметил, что никто не хочет привлекать к предстоящему собранию внимания медиа, ни Маск, ни наше сообщество. В результате собрание началось с опоздания в несколько минут, я должен был его вести, и у нас уже сложился план: сумма дотации не будет оглашена, чтобы объявление не содержало в себе новостного повода; более того, я, со ссылкой на Чатем-хаус, призвал присутствующих держать объявление Маска в секрете на протяжении как минимум

девяти дней, пока ракета не достигнет станции, независимо от успеха с возвращением первой ступени; сам Маск добавил, что будет нуждаться в значительно более продолжительном молчании, если ракета взорвется на старте.

Обратный отсчет дошел до нуля. Сверхразумный президиум сидит на сцене на своих стульях: Элиезер Юдковски, Илон Маск, Ник Бострём, Ричард Маллах, Мюррей Шанахан, Барт Зельман, Шейн Легг, Вернор Виндж. Аплодисменты в зале постепенно стихают, но президиум продолжает сидеть: я сказал им, чтобы они оставались на местах, хотя и не объяснил зачем. Мейя позже говорила мне, что ее пульс в тот момент достиг стратосферы, а рукой под столом она вцепилась в успокаивающую руку Виктории Краковны. Я улыбался. Я знал, что ради этого момента мы так долго работали, так надеялись и так ждали его.

Я счастлив, что на нашей встрече мы достигли общего согласия относительно необходимости продолжения исследований в области создания дружественного искусственного интеллекта, говорил я, и что сейчас мы определили так много конкретных исследовательских направлений, в которых можем двигаться прямо сейчас. Но во время этой сессии было также много разговоров о рисках, добавил я, и поэтому было бы хорошо закрепить наш успех, взглянуть на будущее с большим оптимизмом, прежде чем мы направимся к бару и приступим к ожидающему нас банкету. “И поэтому я передаю свой микрофон ... Илону Маску!” – я чувствовал, как творится история в тот момент, когда Маск взял протянутый микрофон и объявил, что будет поддерживать исследования по AI-безопасности крупными дотациями. Неудивительно, что его выступление завершилось овациями. Как и планировалось, он не назвал суммы, но я знал, что она составит 10 миллионов долларов, о которых мы условились раньше.

После конференции мы с Мейей отправились навестить наших родителей в Швецию и Румынию. В Стокгольме мы с отцом, затаив дыхание, смотрели прямую трансляцию старта ракеты. Попытка посадки завершилась неудачно; как эвфемистично определил ее Маск, – это был RUD, незапланированное быстрое разрушение (rapid unscheduled disassembly), и на то, чтобы благополучно вернуть первую ступень на океанскую платформу после старта ракеты, его команде понадобилось еще пятнадцать месяцев усилий^{134}. Однако все запущенные им спутники благополучно достигли орбиты, как и наша программа грантовой поддержки твитом Маска достигла миллионов его фолловеров^{135}.

Перенос AI-безопасности в центр исследовательского поля

Ключевая цель пуэрто-риканской конференции состояла в том, чтобы сосредоточить интерес исследователей искусственного интеллекта на теме AI-безопасности, и было исключительно занятно видеть, как это происходит: процесс разделился на великое множество шагов. Сначала была сама встреча, на которой многие исследователи почувствовали себя значительно увереннее и спокойнее, когда обнаружили, что их дело собирает вокруг себя целое сообщество специалистов. Слова поддержки от многих из участников были для меня очень важны. Специализирующийся на AI профессор Корнеллского университета Барт Зельман прислал мне мейл: “Честно признаюсь: я никогда не видел лучше организованного, более восхитительного и воодушевляющего научного собрания”.

Следующий этап начался 11 января, когда Илон Маск написал в своем Twitter: “Ведущие разработчики искусственного интеллекта подписывают открытое письмо, призывающее к проведению исследований по AI-безопасности”^{136}, с линком на открытую для подписей страницу с письмом, где очень скоро значилось уже более восьми тысяч подписей, включая имена самых известных создателей интеллектуальных систем мира. Внезапно оказалось очень трудно отмахиваться от тех, кто призывает побеспокоиться об AI-безопасности, как от неспециалистов, потому что из этого немедленно бы следовало, что и самые крутые специалисты по искусственному интеллекту тоже не специалисты. О письме сообщили СМИ, и сделали это так, что мы порадовались своему решению не пускать их на конференцию. Хотя самым алармистским словом в нашем письме было “ловушка”, но, тем не менее, оно дало толчок к появлению, например, такого заголовка: “Илон Маск и Стивен Хокинг подписывают открытое письмо в надежде предотвратить восстание роботов”, проиллюстрированного убийственными терминаторами. Из сотен отслеженных нами статей больше других нам понравилась та, в которой, иронизируя над прочими, автор писал: “Использование в статье образов скелетоподобных андроидов, пинающих ногами человеческие черепа, превращает сложную, трансформирующую мир технологию в карнавальное слайд-шоу”^{137}. К счастью, было также много трезвых статей, но теперь они создали для нас новую проблему: сдерживать поток новых подписей, вручную верифицируя их ради сохранения доверия к нашему делу и избавляясь от пранков вроде “HAL 9000”, “Терминатор”, “Сара Коннор” или “Скайнет”. Ради этого и ради наших будущих открытых писем Виктория Краковна и Янош Крамар помогли организовать бригаду волонтеров, среди которых были Джесси Галеф, Эрик Гастфренд и Реватхи Винотх Кумар, работавшие посменно, так что когда Реватхи в Индии отправлялась спать, она передавала бразды правления Эрику, и так далее.

Третий этап переноса проблемы AI-безопасности в центр внимания исследователей начался четыре дня спустя, когда Маск выложил в своем твиттере ссылку на наше объявление о его дотации в 10 миллионов долларов на исследования по AI-безопасности^{138}. А неделю спустя мы запустили онлайн-портал, где исследователи со всего мира могли подать заявку на конкурс для получения своей части из этой суммы. Мы смогли подстегнуть работу системы подачи заявок, раскрутив ее так быстро, только потому, что мы с Энтони Агирре все предыдущее десятилетие запускали подобные конкурсы для распределения грантов по физике. Калифорнийский благотворительный фонд “Open Philanthropy Project”, заинтересованный в предоставлении грантов проектам, которые обещают высокоэффективную отдачу, великодушно согласился пополнить дотацию Маска, дав нам возможность увеличить количество грантов. Мы не знали, сколько может быть заявок, поскольку тема исследования была новой, а срок проведения конкурса – коротким. Но итог нас потряс: к нам обратились около трех сотен команд со всего мира, претендующих в общей сложности на 100 миллионов долларов. Группа

специалистов по искусственному интеллекту тщательно изучила все предложения и выбрала тридцать семь заявок-победителей, которым выделялось финансирование на срок до трех лет. Когда мы объявили список победителей, медиа впервые отреагировали на итоги нашей деятельности довольно уравновешенно и обошлись без изображений роботов-убийц. Наконец-то стало ясно, что исследования по AI-безопасности – не пустые разговоры: есть нужное дело, и его предстоит выполнить, и есть много исследовательских команд, закатывающих рукава, чтобы объединить свои усилия.

Четвертый этап органично протекал сам собой на протяжении следующих двух лет. Тучи технических публикаций и десятки семинаров по AI-безопасности во всем мире, обычно в составе общей тематики мейнстримных конференций по искусственному интеллекту. Самые настойчивые тратили годы, чтобы привлечь сообщество к исследованиям по AI-безопасности, и без успеха, а теперь все действительно завертелось. Многие из публикаций финансировались нашей программой, и мы в FLI старались помогать в организации и финансировании многих таких семинаров, но все большая их часть проходила на средства самих исследователей, которые вкладывали свои деньги и не жалели своего времени. В результате все больше ученых узнавало об исследованиях по AI-безопасности от своих коллег, обнаруживая, что кроме пользы они могут приносить радость, благодаря решению интересных математических и вычислительных задач, возникающих в связи с ними.

Сложные уравнения не всех радуют, конечно. Спустя два года после нашей конференции в Пуэрто-Рико мы готовились к конференции в Асиломаре и проводили в связи с этим технический семинар, где наши победители, выигравшие грант FLI, могли продемонстрировать результаты своих исследований, и мы рассматривали слайд за слайдом с наполняющими их математическими формулами на большом экране. Моше Варди, профессор по искусственному интеллекту из Райсовского университета, пошутил: он понял, что мы преуспели в превращении AI-безопасности в исследовательскую область, коль скоро конференции стали скучными.

Этот впечатляющий рост числа работ по AI-безопасности не ограничивался университетской наукой. Amazon, DeepMind, Facebook, Google, IBM и Microsoft – все эти компании запустили программы промышленного партнерства по созданию дружественного искусственного интеллекта⁽¹³⁹⁾. Новые внушительные пожертвования на исследования по AI-безопасности позволили существенно расширить их в наших крупнейших некоммерческих дочерних организациях: Machine Intelligence Research Institute в Беркли, Future of Humanity Institute в Оксфорде и Центр по изучению экзистенциальных рисков в Кембридже (Великобритания). Новые пожертвования в 10 миллионов долларов и больше стимулировали дальнейшие усилия в направлении создания дружественного искусственного интеллекта: Leverhulme Centre for the Future of Intelligence в Кембридже, K & L Gates Endowment for Ethics and Computational Technologies в Питтсбурге и Ethics and Governance of Artificial Intelligence Fund в Майами. Последний, но не менее важный, чем прочие в этом ряду, – проект на миллиард долларов при участии Илона Маска наряду с другими предпринимателями для запуска OpenAI, некоммерческой компании в Сан-Франциско по разработке дружественного искусственного интеллекта. Исследования по AI-безопасности пришли, чтобы остаться.

Синхронно с этим исследовательским всплеском пришла и волна мнений, выраженных как отдельными людьми, так и целыми коллективами. Промышленное “Partnership on AI” опубликовало свои основополагающие принципы. Длинные доклады со списками рекомендаций были опубликованы американским правительством, Стэнфордским университетом и IEEE (крупнейшим в мире объединением технических специалистов), вместе с десятками других докладов и программных документов отовсюду⁽¹⁴⁰⁾.

Мы стремились всецело способствовать содержательной дискуссии среди участников конференции в Асиломаре и выяснить, что именно может объединить эти разнообразные

сообщества, если таковое вообще возможно. Лукас Перри взял на себя героическую задачу читать все те документы, которые нам присылали, и извлекать из них главное. В марафоне усилий, инициированных Энтони Агирре и завершенных серией длинных телеконференций, наша команда FLI попыталась сгруппировать предложения по близости высказанных позиций и, избавив их от бюрократического пустословия, превратить в единый список лаконичных принципов, прибавив к ним также некоторые неопубликованные, но влиятельные мнения, высказанные более или менее неофициально в ходе переговоров или при других обстоятельствах. Но список все еще содержал многочисленные двусмысленности и противоречия и оставлял слишком много возможностей для толкований, так что за месяц до начала конференции мы распространили его среди участников, чтобы они выразили свои мнения и дали предложения по улучшению имеющихся принципов или замене их новыми. В результате к началу работы конференции у нас был основательно пересмотренный список принципов, с которым можно было начинать работать.

В Асиломаре список вновь подвергся улучшениям, причем в два этапа. Во-первых, принципы обсуждали небольшие группы наиболее заинтересованных участников (рис. 9.4), которые предлагали подробные уточнения, отклики, новые принципы или новые версии старых. Во-вторых, мы опросили всех участников, чтобы определить уровень поддержки для каждой версии каждого принципа.

Эта коллективная работа была скрупулезной и изнурительной. Мы с Энтони и Мейей до предела сократили время, отведенное на сон и приемы пищи, чтобы успеть учесть все необходимое для наших последующих этапов. Но эта работа захватывала. После такого подробного, многоступенчатого, иногда вызывающего бурные дискуссии и опирающегося на самый широкий спектр откликов обсуждения мы были поражены тем единодушным согласием, которое было выражено по поводу большинства этих принципов в ходе нашего заключительного опроса – до 97 % поддержки по некоторым позициям. Этот консенсус позволил нам установить высокую планку для включения в окончательный список: мы оставили в нем только те принципы, с которыми согласились по крайней мере 90 % участников конференции. Хотя это означало, что некоторые довольно часто предлагавшиеся принципы были сняты в последнюю минуту, в том числе некоторые из тех, что очень нравились мне самому^{141}, но это позволяло большинству участников не испытывать никаких колебаний при одобрении всего подписанного ими списка, который по окончании был распространен по аудитории. И вот результат.



Рис. 9.4

Величайшие умы, разбившись на группы, продумывают принципы разработки искусственного интеллекта в Асиломаре.

Асиломарские принципы разработки искусственного интеллекта

Искусственный интеллект уже создал несколько дружественных инструментов, которыми каждый день пользуются люди во всем мире. Его дальнейшее развитие, основанное на следующих принципах, будет расширять возможности людей и оказывать им помощь на протяжении следующих десятилетий и даже столетий.

Исследовательские вопросы

§ 1 Исследовательская цель: Цель, к достижению которой направлены исследования искусственного интеллекта, состоит не в том, чтобы создать неуправляемый разум, а в том, чтобы разум был дружелюбным.

§ 2 Финансирование исследований: Инвестиции в искусственный интеллект должны сопровождаться финансированием исследований, обеспечивающих его дружелюбность и включающих неудобные вопросы из разных областей знания: информатики, экономики, права, этики и социологии, в том числе:

(а) Как мы можем обеспечить высокую надежность систем с искусственным интеллектом, так чтобы они выполняли то, чего мы от них хотим, без опасности нанесения вреда из-за сбоев или злонамеренного взлома?

(b) Как мы можем добиваться роста нашего процветания благодаря автоматизации при сохранении человеческих ресурсов и целесообразности их жизней?

(с) Как нам обновить наше законодательство, чтобы оно стало более справедливым и эффективным, чтобы оно учитывало существование искусственного интеллекта, и как управлять рисками, связанными с ним?

(d) Как искусственный интеллект должен быть увязан с нашей системой ценностей, и каким должен быть его юридический и этический статус?

§ 3 Научно-политическое звено: Между исследователями искусственного интеллекта и политиками должен быть конструктивный и содержательный диалог.

§ 4 Исследовательская культура: Среди исследователей и разработчиков искусственного интеллекта следует поддерживать атмосферу сотрудничества, доверия и прозрачности.

§ 5 Исключение технологической гонки: Группы, разрабатывающие системы искусственного интеллекта, должны активно сотрудничать и не допускать получения решающего преимущества за счет снижения стандартов безопасности.

Ценности и вопросы этики

§ 6 Безопасность: Системы с искусственным интеллектом должны быть безопасными и хорошо защищенными на протяжении всей операциональной жизни и проверяемыми в этом отношении, когда проверка осуществима.

§ 7 Прозрачность при неисправностях: Если система с искусственным интеллектом причиняет вред, всегда должна быть возможность выяснить причину.

§ 8 Прозрачность при использовании в судебной практике: Привлечение автономной системы к принятию судебных решений должно сопровождаться удовлетворительными обоснованиями, поддерживаемыми компетентными специалистами-людьми.

§ 9 Ответственность: Разработчики и создатели продвинутых систем с искусственным интеллектом несут моральную ответственность за последствия их правильного использования, неправильного использования и любых предпринимаемых ими действий, им вменяется в моральную обязанность предвидеть и формировать эти возможные последствия.

§ 10 Приведение ценностей в соответствие: Независимо функционирующие системы с искусственным интеллектом должны быть спроектированы таким образом, чтобы их цели и поступки гарантированно не приходили в противоречие с человеческими ценностями на протяжении всего срока их работы.

§ 11 Человеческие ценности: Системы с искусственным интеллектом должны разрабатываться и действовать таким образом, чтобы это соответствовало идеалам человеческого достоинства, прав, свобод и культурного разнообразия.

§ 12 Неприкосновенность частной жизни: Люди должны иметь право получать доступ к производимым ими данным, редактировать их и контролировать их использование во всех тех ситуациях, когда такие данные анализируются или используются системами с искусственным интеллектом.

§ 13 Свобода и конфиденциальность: Применение искусственного интеллекта с использованием личных данных людей не должно необоснованно ограничивать реальную или воображаемую свободу этих людей.

§ 14 Общая польза: AI-технологии должны приносить пользу и расширять возможности как можно большего числа людей.

§ 15 Общее процветание: Экономические преимущества, созданные при помощи искусственного интеллекта, должны предоставляться всем, использоваться на благо всего человечества.

§ 16 Человеческий контроль: Делегирование принятия решений искусственному интеллекту и его форма определяется людьми; цели, для достижения которых такие решения будут приниматься, ставятся также людьми.

§ 17 Сохранение социального порядка: Власть, возникающая у высокоразвитой системы с искусственным интеллектом благодаря ее возможностям управления, должна поддерживать и развивать социальные и гражданские процессы, от которых зависит общественное благополучие, а не подрывать их.

§ 18 Гонка интеллектуальных систем вооружений: Следует всеми мерами избегать начала гонки вооружений в сфере разработки смертоносных автономных систем.

Долгосрочные аспекты

§ 19 Осторожность в отношении способностей: В отсутствие общего согласия мы должны избегать сильных предположений относительно верхнего предела будущих возможностей искусственного интеллекта.

§ 20 Значение для будущего: Продвинутой искусственный интеллект может стать причиной глубоких изменений в истории жизни на Земле, они должны планироваться и управляться с соразмерной заботой и привлечением необходимых ресурсов.

§ 21 Риски: Риски, связанные с развитием систем искусственного интеллекта, в особенности катастрофические или экзистенциальные, должны регулироваться соразмерными усилиями по планированию и смягчению последствий их ожидаемого воздействия.

§ 22 Последовательное самосовершенствование: Системы искусственного интеллекта, конструктивно предполагающие самосовершенствование или саморазмножение, которые могли бы привести к быстрому росту их качества или количества, должны подвергаться самому строгому контролю в плане безопасности и управляемости.

§ 23 Общее благо: Сверхразум может создаваться только для служения всеми разделяемым этическим идеалам и во благо всего человечества, а не какого-то одного государства или какой-то одной организации.

После того как мы вывесили список этих принципов онлайн, количество подписей под ним стало быстро расти, и сейчас среди них можно видеть имена более тысячи самых известных мыслителей и ведущих разработчиков интеллектуальных систем. Если и вы хотите поставить свою подпись, пройдите по ссылке: <http://futureoflife.org/ai-principles>.

Нас поразило не столько то, как высок был уровень всеобщего согласия относительно этих принципов, сколько сама их сила. Конечно, некоторые могут на первый взгляд показаться такими же спорными, как что-нибудь вроде “Мир, любовь и материнство – бесценны”. Но у многих из них есть настоящие зубы, которые можно заметить, например, если попытаться сформулировать отрицание. Например, утверждение “Сверхразум невозможен!” будет противоречить § 19, а “Проводить исследование, направленное на снижение экзистенциального риска от искусственного интеллекта, – только попусту тратить время!” будет противоречить § 21.

В самом деле, как вы можете самостоятельно убедиться, посмотрев на YouTube наши панельные дискуссии, посвященные долгосрочной перспективе^{142}, Илон Маск, Стюарт Рассел, Рэй Курцвейл, Демис Хассабис, Сэм Харрис, Ник Бострём, Дэвид Чалмерс, Барт Зельман и Яан Таллин – все согласились, что сверхразум, возможно, будет создан и исследования по безопасности важно проводить.

Я надеюсь, что наши Асиломарские принципы разработки искусственного интеллекта станут исходной точкой для более детальных дискуссий, которые в итоге приведут нас к созданию разумных стратегических и политических решений по отношению к искусственному интеллекту. В этом духе наш медиа-директор Эриел Конн работал с Такером Дави и другими членами нашей команды, проводя интервью среди ведущих исследователей по поводу их видения и интерпретаций создаваемых принципов, а Дэвид Стэнли и его команда международных волонтеров FLI переводили эти принципы на ключевые языки мира.

Разумный оптимизм

Как я объяснил в начале этого эпилога, сейчас я испытываю значительно больше оптимизма по отношению к будущему, чем раньше. Я сейчас расскажу почему.

Есть две разные причины, которые способствовали росту моего оптимизма на протяжении последних двух лет. Во-первых, я был свидетелем тому, как сообщество специалистов по искусственному интеллекту складывалось самым конструктивным образом и не боялось браться за решение будущих проблем, нередко привлекая для этого думающих людей из смежных исследовательских областей. После нашей конференции в Асиломаре Илон Маск говорил мне, что приятно удивлен тем, как тема безопасности с обочины исследований искусственного интеллекта переместилась в самый центр этой исследовательской области всего лишь за несколько лет, и это удивляет и меня самого. Сейчас она включает в себя не только краткосрочные задачи из главы 3, которые стало не стыдно обсуждать, но и вопросы, связанные со сверхразумом и экзистенциальными рисками, как мы видим по Асиломарским принципам. Эти принципы никоим образом не могли быть одобрены на конференции в Пуэрто-Рико двумя годами раньше, где самым страшным словом, просочившимся в итоговое открытое письмо, было слово “ловушки”.

Мне нравится наблюдать за людьми, и в какой-то момент, в последнее утро асиломарской конференции, я стоял в боковом проходе аудитории и наблюдал за дискуссией о правовых аспектах искусственного интеллекта. Вдруг, неожиданно для меня самого, в моей душе разлилось что-то теплое и мягкое, я ощутил себя внезапно взволнованным. Это было так непохоже на испытываемое мной в Пуэрто-Рико! Тогда, я помню, я смотрел на AI-сообщество со смесью страха и уважения – нет, не как на команду противника, но как на группу моих коллег, которых мне предстоит в чем-то переубедить. А теперь мне было совершенно очевидно, что мы – одна команда. И хотя, читая эту книгу, вы, наверное, уже поняли, что у меня нет готовых ответов на вопросы о том, как с помощью искусственного интеллекта создать нам всем великое будущее, я рад ощущать себя частью растущего сообщества, внутри которого эти ответы ищут сообща (рис. 9.4).

Вторая причина, по которой растет мой оптимизм, – в расширяющемся опыте нашего FLI. Слезы брызнули у меня из глаз в Лондоне от чувства неизбежности: наступает тревожное будущее, но мы ничего с этим поделать не можем. Прошло три года, и это чувство фатальной предопределенности рассеялось. Даже если небольшая компания разномастных волонтеров смогла развернуть в правильную сторону то, что называли самой важной темой нашего времени, подумайте только, что мы сможем сделать, взявшись все вместе!

Эрик Бриньоулфссон в своем асиломарском выступлении говорил о двух типах оптимизма. Первый – безусловный. Он включает в себя, например, положительную уверенность, что и завтра солнце тоже взойдет. Но кроме него, есть и такой, который он называет разумным оптимизмом – это уверенность, что хорошие вещи будут случаться, если ты хорошо все спланируешь и хорошо ради них потрудишься. Это тот самый тип оптимизма, который я испытываю в отношении будущего жизни.



Рис. 9.4

Растущее исследовательское сообщество занято совместным поиском ответов в Асиломаре.

Так что *вы* можете сделать, чтобы изменить будущее жизни к лучшему в момент вступления в эпоху искусственного интеллекта? По причинам, которые я сейчас объясню, удачный первый шаг поможет вам стать разумным оптимистом, если вы еще им не стали. Чтобы быть успешным разумным оптимистом, очень важно развивать позитивное видение будущего. Когда ко мне в офис приходят студенты MIT и спрашивают моих советов, какую карьеру выбрать, я обычно начинаю расспрашивать их о том, какими они видят себя через десять лет. Если студентка отвечает мне: “Наверное, я буду уже в раковом корпусе или на кладбище после встречи с автобусом”, – я не могу сказать ей ничего хорошего. Ожидать от будущего плохого – это бесперспективный подход к планированию своей карьеры. Посвящать 100 % своих усилий только тому, чтобы избежать болезней и несчастных случаев, – это чудесный рецепт для ипохондрии и паранойи, но никак не для счастья! Я бы посоветовал ей, напротив, отнестись к своим целям с энтузиазмом, и тогда мы сможем обсудить стратегии, которые приведут к ним и помогут избежать ловушек и ухабов.

Эрик Бриньоулфссон указывал, что позитивные ожидания в соответствии с теорией игр дают самое надежное в мире основание для сотрудничества – от заключения брака и слияния компаний до объединения независимых штатов в США. В конце концов, зачем жертвовать чем-то, что у вас уже есть, если вы не видите, как получить при помощи этого что-то большее? А это означает, что мы должны проектировать лучшее будущее не только для себя, но и для общества и для человечества в целом. Другими словами, нам нужно больше экзистенциальной надежды! Однако, как не устаю мне напоминать Мейя, футуристические образы в кино и литературе, от Франкенштейна до Терминатора, по большей части антиутопичны. Выходит, мы как общество планируем себе столь же мрачное будущее, как гипотетическая студентка MIT. И поэтому нам надо больше разумного оптимизма. И поэтому на протяжении всей книги я убеждал вас прежде всего подумать о том, какого именно будущего вы бы *хотели*, а не о том, какое будущее вас

пугает, потому что только так мы можем найти общие цели, ради которых стоит планировать и трудиться.

На протяжении всей книги мы видели, что искусственный интеллект, наверное, дает нам грандиозные возможности и создает нам грандиозные проблемы. Стратегия, которая с наибольшей вероятностью поможет разобраться практически со всеми этими проблемами, заключается в том, чтобы объединиться в совместном действии и улучшить наше человеческое общество *до того, как* искусственный интеллект по-настоящему стартует. Нам надо думать, как улучшать образование молодых людей, чтобы создаваемые ими технологии были надежны и дружелюбны, до того, как эти технологии станут по-настоящему могучими. Нам надо думать, как модернизировать свою законодательную и судебную системы, – до того, как рост технологий сделает их безнадежно устаревшими. Нам надо думать, как разрешить международные конфликты, – до того, как их эскалация приведет к гонке в сфере автономных систем оружия. Нам надо думать, как создать экономику, обеспечивающую процветание для всех, – до того, как искусственный интеллект сможет раздуть существующее неравенство. Нам надо позаботиться, чтобы не забывали об AI-безопасности и подумали о том, как применить результаты проводимых исследований уже сейчас. Мы смотрим вперед, и в предчувствии проблем, которые возникнут в связи с универсальным искусственным интеллектом, нам надо думать, как достичь согласия хотя бы по базовым этическим стандартам, прежде чем мы начнем прививать эти стандарты разумным машинам. В поляризованном и хаотическом мире люди, обладающие властью и возможностью использовать искусственный интеллект для дурных целей, окажутся и более мотивированными, и более свободными, а в состязании с другими группами, стремящимися к созданию универсального искусственного интеллекта, будут скорее стремиться закрывать глаза на решение проблем безопасности, чем работать над ними вместе с другими. Одним словом, если мы преуспеем в создании гармоничного человеческого общества, характеризующегося высоким уровнем сотрудничества в достижении общих целей, то наши общие шансы на благополучное завершение начавшейся AI-революции только вырастут.

А это значит, что улучшить будущее жизни можно, только улучшив наше завтра. У вас есть много способов сделать это. Конечно, вы можете делать это у ваших избирательных урн, втолковывая политикам, за кого голосуете, что думаете об образовании, конфиденциальности, летальном автономном оружии, технологической безработице и обо всем остальном. Но вы можете делать это и каждый день, выбирая, что покупать, какие потреблять новости, чем делиться с другими и какую ролевую модель выбрать для себя. Хотите ли вы быть одним из тех, кто не задумываясь прерывает беседу, чтобы проверить свой мессенджер в смартфоне, или одним из тех, кто чувствует себя в силах пользоваться новыми технологиями планомерно и по своему выбору? Хотите ли вы располагать технологиями или хотите, чтобы технологии располагали вами? Что, по-вашему, это должно значить: быть человеком в эпоху искусственного интеллекта? Пожалуйста, обсудите это с теми, кто вас окружает – это не только важный, но и увлекательный разговор!

Мы все – стражи будущего жизни сейчас, когда мы создаем эпоху искусственного интеллекта. Хотя я и расплакался в Лондоне, сейчас я чувствую, что в будущем нет ничего неизбежного, и я знаю, что изменить его намного проще, чем я думал. Наше будущее не высечено в камне, и оно не свалится на нас, дождавшись своего часа, – наше дело его создавать. Давайте же вместе создадим его вдохновляющим!

Сноски

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.

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

Имеется в виду одно из современных направлений так называемой DIY-субкультуры (от английского “do it yourself”, то есть “сделай сам”), но с расчетом на более развитую технологическую базу. С 2004 года у них есть свой публичный печатный орган – журнал *MAKE*. – *Прим. перев.*

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.

В англоязычном мире имя этого человека принято транслитерировать иначе: “Моравек”, но мы будем придерживаться исходного, чешского варианта. – *Прим. перев.*

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

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.*

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

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.

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.

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 странно называть слабым AI Deep Blue, Watson или AlphaGo.

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.

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.*

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.

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 $\sigma(x)$ ко всем новым вычисленным элементам.

Широко используемая аббревиатура от Good Old-Fashioned Artificial Intelligence, что означает “старый добрый искусственный интеллект”. – *Прим. перев.*

Цинь (琴) – общее название различных китайских музыкальных инструментов, из которых наиболее популярны семиструнная цитра (七絃琴) и ее более ранний вариант гуцинь (古琴). – *Прим. перев.*

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.*

Исходная фраза из английского текста книги была переведена на мандарин, потом на испанский и после этого на русский. Таким образом, описываемая автором путаница на этот раз возникла без насилия со стороны переводчика. – *Прим. перев.*

Если вы хотите более подробную карту исследовательского ландшафта AI-безопасности, вы можете найти ее в представлении одного из знаменосцев Института будущего жизни Ричарда Маллаха, пройдя по ссылке <https://futureoflife.org/landscape/>

Lyft – базирующаяся в Сан-Франциско частная компания, обеспечивающая с помощью мобильного приложения поиск сотрудничающих с ней частных водителей, которые готовы предоставлять транспортные услуги. На начало 2017 года компания работала более чем в 300 американских городах, а в 2018-м получила большое распространение и в Канаде. Успешно конкурирует в этих странах с Убером. – *Прим. перев.*

Даже при включении этого ДТП в статистику Tesla-автопилот сокращает их общее количество на 40 %: <http://tinyurl.com/teslasafety>.

Интернет вещей – это бурно развивающаяся в последние полтора десятилетия концепция, согласно которой в непосредственное информационное взаимодействие вовлекаются самые разнообразные предметы, от домашней утвари до промышленных объектов. – *Прим. перев.*

Крейсерами класса Aegis (Иджис) называют крейсера, на которых установлена корабельная многофункциональная боевая информационно-управляющая система Aegis. Phalanx – это система оружия ближнего боя, часто входящая в состав Aegis. – *Прим. перев.*

Третий ракетный крейсер типа Ticonderoga (“Тикондерога”). Крейсера класса Ticonderoga были первыми, на которых стояла система Aegis, то есть крейсерами класса Aegis. – *Прим. перев.*

Слова капитана второго ранга В. П. Орлова цитируются в русском оригинале по: Астафьева Е. Советский супергерой // *Дилетант*. 2016. Октябрь (доступно онлайн: <http://diletant.media/articles/31508321/>) и по: Язов Д. *Карибский кризис: 50 лет спустя*. М.: Центрполиграф, 2015. С. 344 (фразы “Казалось, будто сидишь в железной бочке, по которой колотят кувалдой” в книге Язова нет). – *Прим. перев.*

Цит. по: Язов Д. *Карибский кризис: 50 лет спустя*. Там же. – Прим. перев.

Кандид, или Оптимизм, гл. 30. Пер. с фр. Ф. Сологуба. – *Прим. перев.*

Я напомним здесь, что флопс – это количество операций над числами с плавающей запятой (аббревиатура английского *floating-point operations per second*), выполняемых компьютером в одну секунду, то есть это число, показывающее, например, сколько 19-значных чисел компьютер перемножит одно на другое в течение секунды.

Как объяснял Бострём, возможность симулировать ведущих разработчиков интеллектуальных систем при затратах, существенно меньших их суточных заработков, позволит AI-компаниям резко увеличить свою производительность и, концентрируя колоссальное богатство, обеспечить ускоряющийся прогресс в создании новых компьютеров и, в итоге, все более сильного интеллекта.

Сама по себе идея восходит еще к Блаженному Августину, который писал: “Ибо если вещь такова, что через разделение ее с другими нисколько не оскудевает; то если нам дают ее, а мы напротив имеем, и не даем ее другим, – мы пользуемся этой вещью не надлежащим образом” (De doctrina Christiana. Lib. 1, cap. 1. Рус. пер. цит. по: Блаженный Августин, епископ Иппонийский. *Христианская наука, или Основания св. герменевтики и церковного красноречия*. Киев: Типография Троицко-Печерской лавры, 1835. С. 14. – Прим. перев.)

Эту идею впервые сформулировал и исследовал мой друг Энтони Агирре.

Аристотель. *Политика* 1253b. Рус. пер. С. А. Жебелева. Цит. по: Аристотель. *Сочинения*: В 4 т. М.: Мысль, 1975–1984. Т. 4. С. 381. – *Прим. перев.*

Политика 1254b. Цит. соч. С. 383. – Прим. перев.

Знаменитый космолог Фред Хойл развивал похожую идею, но с оригинальным сюжетным поворотом в сериале *A for Andromeda* на британском телевидении.

Выброс углерода в атмосферу влечет два вида климатических последствий: потепление от двуокиси углерода и похолодание от дыма и золы. Не только первое иногда объявляется лишенным научных обоснований: мне иногда говорят, что ядерная зима давно развенчана как фактически неосуществимая. В ответ я всегда требую показать мне ссылки на публикации в реферируемых журналах, где делаются такие сильные заявления, и до сих пор никто не смог мне их предоставить. Хотя здесь остается значительная неопределенность, требующая дальнейших исследований, связанная в особенности с такими параметрами, как интенсивность выбросов и достигаемая ими высота, но, на мой взгляд, нет никаких оснований отвергать в настоящее время риск ядерной зимы.

См. рус. пер.: Стэплдон О. *Создатель звезд*. М.; Киев: Ваклер, Рефл-бук, 2017. – *Прим. перев.*

Если вы работаете в энергетике, то вам привычнее другое определение эффективности, при котором за меру эффективности принимается отношение полезной энергии ко всему полученному ее количеству.

См.: Хокинг С. *Краткая история времени: от Большого взрыва до чёрных дыр* / пер. с англ. Н. Я. Смородинская. СПб.: "Амфора", 2001. – Прим. перев.

Если поблизости в космическом пространстве не окажется никакой подходящей черной дыры, созданной самой природой, ее можно создать самим, засунув очень много материи в какое-нибудь достаточно маленькое место.

Здесь у нас опять некоторое чрезмерное упрощение, поскольку излучение Хокинга включает в себя и некоторые частицы, которые трудно пустить в дело. Большая черная дыра только 90 % своей массы излучает в доступном для использования виде, а остальные 10 % – в виде гравитонов. Это исключительно деликатные частицы: их почти невозможно обнаружить, не говоря уж о том, чтобы как-то использовать. По мере того как черная дыра все больше излучает и все больше сжимается, в ее излучении появляется все больше нейтрино и других массивных частиц.

Если тут случится оказаться поклонникам Дугласа Адамса, то специально для них я уточню, что это элегантный вопрос, дающий ответ на другой вопрос – о жизни, Вселенной и вообще обо всем. Значение 42 % – приближенное, на самом деле $1 - 1/\sqrt{3} \approx 42\%$.

Если вы топите свою черную дыру газовым облаком, медленно вращающимся в ту же сторону, что и дыра, то газ, съедаясь, будет раскручиваться все быстрее, раскручивая и черную дыру, вроде как фигуристка, прижимающая к груди руки во время вращения. Таким образом ее можно раскрутить до максимальной скорости, из-за чего ей придется отдать 42 % энергии газа, а потом еще и 29 % своей собственной, и тогда суммарная эффективность окажется равной $42 \% + (1 - 42 \%) \times 29 \% = 59 \%$.

Это нужно для того, чтобы достичь температур, при которых происходит объединение электромагнитных сил и сил слабого ядерного взаимодействия, а это возможно, когда частицы разгоняются в ускорителе напряжением в 200 миллиардов вольт.

До сих пор мы говорили только о веществе, построенном из атомов. Но кроме такого, есть еще и так называемая темная материя, которой в шесть раз больше. Правда, она практически неуловима: она легко проходит сквозь Землю, ни за что не зацепляясь внутри. Остается только ждать, чтобы увидеть, сможет ли будущая жизнь как-то ее захватить и для чего-то использовать.

Космическая математика на этот счет предельно ясна: если цивилизация расширяется в расширяющемся пространстве не со скоростью света c , а с какой-то меньшей скоростью v , то число галактик, где она сможет расселиться, уменьшается в пропорции $(v/c)^3$. Таким образом, цивилизации-копуши будут строго наказаны! Той, которая расширяется в 10 раз медленнее, достанется в 1000 раз меньше галактик.

Имеется в виду поговорка “Когда следишь за чайником, он не закипает”. – *Прим. перев.*

Пер. на рус. яз. В. Л. Топорова. – *Прим. перев.*

Однако и Джон Гриббин приходит к аналогичному заключению в своей книге 2011 года *Alone in the Universe*. Я советую также прочитать обсуждение целого спектра возможностей по этому поводу в книге Пола Дэвиса 2011 года *The Eerie Silence*.

См.: Тегмарк М. *Наша математическая Вселенная. В поисках фундаментальной природы реальности* / пер. с англ. А. Сергеев. М.: Corpus, 2016. – Прим. перев.

См.: Шрёдингер Э. *Что такое жизнь с точки зрения физики?* Лекции, читанные в Тринити-колледж в Дублине в феврале 1943 г./ пер. с англ. и послесл. А. А. Малиновский. М.: Гос. изд-во иностр. лит., 1947. – *Прим. перев.*

Проверенное правило, которым пользуется большинство насекомых во время полета по прямой, заключается в том, чтобы считать яркий свет Солнцем и лететь под определенным углом к нему. Если источником света оказывается расположенный рядом огонь, этот прием, к сожалению, заставляет жука лететь по сужающейся смертельной спирали.

Я говорю об улучшение “софта” в самом широком из возможных смыслах, включающем не только оптимизацию алгоритмов, но и развитие более рационального процесса принятия решений, который поможет ему все лучше и лучше достигать своих целей.

См.: Вильчек Ф. *Красота физики. Постигая устройство природы* / пер. с англ. В. Краснянская, М. Томс. М.: Альпина-нон-фикшн, 2016. – *Прим. перев.*

Пер. Е. Шесткова.

Есть и альтернативная, дуалистическая точка зрения, в соответствии с которой живые существа отличаются от неодушевленных тем, что содержат какую-то нефизическую субстанцию, которую называют “анимой”, “жизненной энергией” или “душой”. Число сторонников этой дуалистической точки зрения среди ученых все время сокращается. Чтобы понять, почему это так, вспомните, что ваше тело – это примерно 10^{29} кварков и электронов, которые двигаются в соответствии с простыми, насколько мы можем судить, физическими законами. Представьте себе будущую технологию, которая позволит отслеживать движение каждой из этих частиц. Если окажется, что все они подчиняются законам физики, то душа, следовательно, никак не влияет на частицы, из которых вы состоите, а значит, ваше сознание, ваш разум и их способность контролировать ваши движения ничего общего с душой не имеют. Если, напротив, окажется, что частицы, из которых вы состоите, не подчиняются известным законам физики, потому что этому мешает воздействие на них вашей души, то эта новая сущность должна быть физической по определению, и тогда мы сможем ее изучать, как изучали поля и частицы в прошлом.

Я использую термин “квалиа” в соответствии с определением его в словарях: он означает индивидуальный пример субъективного переживания, – то есть субъективный опыт сам по себе, а не сопутствующую субстанцию, его вызывающую. Имейте в виду, что некоторые люди используют слово в другом значении.

RHP я изначально назвал “очень трудная проблема” (very hard problem), но когда я отослал эту главу Дэвиду Чалмерсу, он мне ответил, что было бы точнее назвать ее “по-настоящему трудной”: “поскольку первые две проблемы (по крайней мере, в таком виде) не были частью трудной проблемы, над которой я думал, тогда как третья из них и есть та самая; поэтому, возможно, сказать про нее ‘по-настоящему трудная’, а не ‘очень трудная’ более соответствовало бы моему замыслу”.

Если наша физическая реальность всецело математическая (или, допуская определенную речевую вольность, ее можно назвать информационной), как я предположил в своей книге *Our Mathematical Universe*, то никакой аспект реальности – даже сознание – не может оказаться за пределами науки. Действительно, по-настоящему трудная проблема сознания (RHP) с этой точки зрения в точности совпадает с другой – пониманием, как нечто математическое может воспринимать физическое: если какая-то часть математической структуры обладает сознанием, то она будет переживать остальную часть этой структуры как внешний физический мир.

Функциональная магнитно-резонансная томография, электроэнцефалография, магнитная энцефалография, электрокортикография, электрофизиология соответственно. – *Прим. перев.*

Хотя ранее я использовал слово *персептрониум* как синоним для слова *сентрониум*, подразумеваемый смысл первого слишком ограничен, поскольку *perceptum* подразумевает лишь те субъективные ощущения, источником которых служат наши органы чувств, что исключает, например, мечты или мысли, возникшие в ходе внутреннего диалога.

Это утверждение в потенциальном напряжении с идеей о независимости сознания от субстрата, поскольку, хотя обработка информации на нижних уровнях может различаться, на верхних уровнях, определяющих поведение, она по определению тождественна.

Этот опыт также заставил меня пересмотреть свое отношение к новостям. Хотя мне и раньше было понятно, что большинство из них продаются через торговые точки со своей политической повесткой у каждой, теперь я стал видеть, что их подача всегда тенденциозна, и всякая новость, даже не политическая, предстает в максимальном удалении от своего центра тяжести.

Комментарии

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

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

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

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

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>

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.).

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

DeerMind алгоритм глубокого машинного обучения с подкреплением позволил довольно быстро научиться играть в Breakout: <https://tinyurl.com/atari.ai>

См. статью, в которой описывается искусственный интеллект DeepMind, совершенствующийся в играх на платформе Atari: <http://tinyurl.com/ataripaper>

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

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

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

GoogleTranslate доступен по ссылке: <https://translate.google.com>

Winograd Schema Challenge см.: <http://tinyurl.com/winogradchallenge>

Взрыв ракеты “Ариан-5”: <https://www.youtube.com/watch?v=qnHn8W1Em6E>

Доклад комиссии по расследованию аварийного пуска “Ариан-5”:
<http://tinyurl.com/arianeflop>

Доклад по результатам первой фазы расследования неудачного пуска миссии NASA по изучению марсианского климата: <http://tinyurl.com/marsflop>

Наиболее подробный и связный отчет о том, что авария на старте миссии NASA “Маринер-1” к Венере была вызвана одной-единственной ошибкой – математический символ, написанный рукой человека, был без верхнего подчеркивания: <http://tinyurl.com/marinerflop>

Подробное описание причин неудачи с советским зондом “Фобос-1”, направлявшимся к Марсу, можно найти на с. 308 этой книги: Wesley Huntress & Mikhail Marov. *Soviet Robots in the Solar System*, Praxis Publishing, 2011.

Как неverified софт обошелся Knight Capital в 440 миллионов долларов всего за 45 минут см.: <http://tinyurl.com/knightflop1> и <http://tinyurl.com/knightflop2>

Доклад правительства США о причинах резкого скачка индекса Доу-Джонса 6 мая 2010 года ("Flash Crash"): <http://tinyurl.com/flashcrashreport>

3D-печать зданий (<https://www.youtube.com/watch?v=SObzNdyRTBs>),
микромеханических устройств (<http://tinyurl.com/tinyprinter>) и много чего еще (<https://www.youtube.com/watch?v=xVU4FLrsPXs>)

Глобальная карта fab-лабораторий: <https://www.fablabs.io/labs/map>

Новостная заметка об убийстве Роберта Уильямса промышленным роботом:
<http://tinyurl.com/williamsaccident>

Новостная заметка об убийстве Кендзи Урада промышленным роботом:
<http://tinyurl.com/uradaaccident>

Новостная заметка об убийстве рабочего с завода Фольксвагена промышленным роботом: <http://tinyurl.com/baunatalaccident>

Отчет правительства США о несчастных случаях на производстве со смертельным исходом: https://www.bls.gov/iif/foi/cfoi_revised14.htm

Статистика ДТП со смертельным исходом: <http://fortune.com/2017/02/15/traffic-deadliest-year/>

Новостная заметка о первом ДТП со смертельным исходом с участием беспилотного автомобиля Tesla: <http://tinyurl.com/teslacrashstory>

Доклад правительства США: <http://tinyurl.com/teslacrashreport>

О крушении парома “Herald of Free Enterprize” см. кн.: R. B. Whittingham. *The Blame Machine: Why Human Error Causes Accidents*. Elsevier, 2004.

Документальный фильм о крушении французского самолета, выполнявшего рейс Air France 447: <https://www.youtube.com/watch?v=So9iFv0NoZw>

Официальное сообщение о крушении: <http://tinyurl.com/af447report>

Независимый анализ: <http://tinyurl.com/thomsonarticle>

Официальное сообщение о веерном отключении энергоснабжения в США и Канаде в 2003 году: <http://tinyurl.com/uscanadablackout>

Итоговый отчет президентской комиссии об инциденте в Три-Майл-Айленде:
<http://large.stanford.edu/courses/2017/ph241/blackwood1/docs/kemeny.pdf>

Голландское исследование, показывающее, что искусственный интеллект не уступает человеческому в радиологической диагностике рака простаты с использованием магнитно-резонансной томографии: <http://tinyurl.com/prostate-ai>

Исследование, проведенное в Стэнфорде, показывающее, что искусственный интеллект позволяет лучше проводить диагностику рака легких, чем анализы врачей:
<http://tinyurl.com/lungcancer-ai>

Результаты расследования несчастного случая лучевой терапии с использованием аппарата

Therac-25:

http://www.cse.msu.edu/~cse470/Public/Handouts/Therac/Therac_1.html

Доклад о повлекшей смерть ошибке в определении дозы облучения из-за вмешательства человека при лечении больного раком в Панаме:
<http://tinyurl.com/cobalt60accident>

Исследование роботизированных хирургических операций с неблагоприятным исходом: <https://arxiv.org/abs/1507.03518>

Статья о смертельных исходах, вызванных ненадлежащим содержанием в больницах:
<http://tinyurl.com/medaccidents>

Yahoo задал новый эталон “большого взлома”, сообщив о краже трех миллиардов (!) учетных записей своих пользователей:
<https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html>

Статья в *New York Times* о том, как убийца из Ку-клукс-клана был сначала оправдан, а потом осужден – много лет спустя: <http://tinyurl.com/kkkacquittal>

Исследование Данцигера с коллегами 2011 года (<http://www.pnas.org/content/108/17/6889.full>), доказывающее, что голодные судьи выносят более жестокие приговоры, было оспорено Вейншел-Марджел и Джоном Шапардом (<http://www.pnas.org/content/108/42/E833.full>), но Данцигер с коллегами привел новые свидетельства своей правоты: <http://www.pnas.org/content/108/42/E834.full>

Сообщение в *Pro Publica* о расовых предубеждениях программного обеспечения, предсказывающего криминальные рецидивы: <http://tinyurl.com/robojudge>

Использование функциональной магнитно-резонансной томографии и других методов сканирования мозга в качестве доказательства в судах крайне сомнительно из-за низкой достоверности самих этих методов, хотя многие исследовательские группы утверждают, что их точность превышает 90 %:

<http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00709/full>

Американская телекомпания PBS сняла фильм *The Man Who Saved the World* (Человек, который спас мир) об инциденте, в котором Василий Архипов предотвратил советский ядерный удар по Америке: <https://www.youtube.com/watch?v=4VPY2SgyG5w>

Рассказ о том, как Станислав Петров не поверил поступившим предупреждениям о готовящейся ядерной атаке Соединенных Штатов на СССР, лег в основу фильма *The Man Who Saved the World* (*Человек, который спас мир* – этот фильм не следует путать с другим, упомянутым в предыдущей сноске, с таким же названием). В 2008 году Петров был приглашен в штаб-квартиру ООН, где получил награду Ассоциации граждан мира: <https://www.youtube.com/watch?v=IncSjwWQHMo>

Открытое письмо от исследователей в области искусственного интеллекта и робототехники о запрете автономного оружия: <http://futureoflife.org/open-letter-autonomous-weapons/>

В американском правительстве, очевидно, есть люди, стремящиеся развязать гонку AI-вооружений: <http://tinyurl.com/workquote>

Исследование неравномерности распределения богатства в США с 1913 года:
<http://gabriel-zucman.eu/files/SaezZucman2015.pdf>

Доклад Oxfam о глобальном неравенстве: <http://tinyurl.com/oxfam2017>

Данные взяты из: Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman, *The World Wealth and Income Database* (<http://www.wid.world>, 31/10/2016).

Прекрасное обоснование гипотезы технологически обусловленного неравенства дано в: Erik Brynjolfsson & Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Norton, 2014.

Статья в *The Atlantic* о падающих зарплатах у представителей наименее образованных слоев общества: <http://tinyurl.com/wagedrop>

Презентация Джеймса Манийки, показывающая перераспределение доходов от труда к капиталу: http://futureoflife.org/data/PDF/james_manyika.pdf

Прогноз относительно автоматизации рабочих мест Оксфордского университета (<http://tinyurl.com/automationoxford>) и Мак-Кинзи (<http://tinyurl.com/automationmckinsey>)

Робот-повар: <https://www.youtube.com/watch?v=fE6i2OO6Y6s>

Марин Солачич исследовала эти возможности во время семинара 2016 года “Компьютеры пошли вразнос: прямое и косвенное воздействие достижений в развитии искусственного интеллекта на общество”: <http://futureoflife.org/2016/05/06/computers-gone-wild/>

Предложения Эндрю Макафи по созданию большего количества хороших рабочих мест: http://futureoflife.org/data/PDF/andrew_mcafee.pdf

В дополнение ко многим академическим статьям, доказывающим, что “на этот раз все будет по-другому”, в этом ролике *Люди на работу не требуются* в сжатой форме проводится та же мысль: <https://www.youtube.com/watch?v=7Pq-S557XQU>

Бюро по трудовой статистике США: <http://www.bls.gov/cps/cpsaat11.htm>

Аргумент “на этот раз все будет по-другому” в пользу технологической безработицы:
Federico Pistono. *Robots Will Steal Your Job, but That's OK*, 2012,
<http://robotswillstealyourjob.com>

Колебания численности популяции лошадей в США: <http://tinyurl.com/horsedecline>

Метаанализ, показывающий, как безработица отражается на благосостоянии: Maïke Luhmann et al. Subjective well-being and adaptation to life events: a meta-analysis. *Journal of personality and social psychology* 102.3 (2012): 592, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289759>

Исследования причин, определяющих субъективное благосостояние:

Angela Duckworth, Tracy Steen, and Martin Seligman 2005, *Positive Psychology in Clinical Practice* (<http://tinyurl.com/wellbeingduckworth>); Weiting Ng & Ed Diener, What matters to the rich and the poor? Subjective well-being, financial satisfaction, and postmaterialist needs across the world. *Journal of personality and social psychology* 107.2 (2014): 326 (<http://psycnet.apa.org/journals/psp/107/2/326>); Kirsten Weir. *More than job satisfaction* (<http://www.apa.org/monitor/2013/12/job-satisfaction.aspx>)

Перемножение примерно 10^{11} нейронов, примерно 10^4 синапсов на нейрон и одного (10^0) возбуждения нейрона в секунду дает примерно 10^{15} FLOPS (1 petaFLOPS) – такого быстродействия достаточно для симуляции человеческого мозга, но здесь есть много плохо понимаемых осложнений, в том числе закон, по которому происходят возбуждения, и вопрос, надо ли отдельно симулировать небольшие части нейронов и синапсов. Специалист IBM Дхармендра Модха оценил необходимую производительность в 38 petaFLOPS (<http://tinyurl.com/javln43>), а нейробиолог Генри Маркрам дал другую оценку – около 1 000 petaFLOPS (<http://tinyurl.com/6rpohqv>). Катя Грейс и Пол Кристиано утверждают, что наиболее дорогой аспект симуляции мозга не в вычислениях, а в коммуникациях, но и это под силу современному суперкомпьютеру: <http://aiimpacts.org/about>

Эта статья содержит интересные оценки вычислительной мощности человеческого мозга: Hans Moravec. When will computer hardware match the human brain. *Journal of Evolution and Technology*. 1998. Vol. 1.

Первая механическая птица: https://www.ted.com/talks/a_robot_that_flies_like_a_bird

Цит. по: Ray Kurzweil, *The Singularity is Near*, Viking Press, 2005.

См.: сценарий Бена Гёрцля: https://wiki.lesswrong.com/wiki/Nanny_AI

Статья об отношениях людей и машин (можно ли считать машины нашими рабами):
<http://tinyurl.com/aislaves>

Мысленные преступления обсуждаются Бострёмом в его книге *Superintelligence* и, с несколько большим количеством технических деталей, в статье: Nick Bostrom, Allan Dafoe & Carrick Flynn (2016) *Policy Desiderata in the Development Of Machine Superintelligence*, <http://www.nickbostrom.com/papers/aipolicy.pdf>

Воспоминания бывшего подполковника восточногерманских спецслужб см.:
<http://www.mcclatchydc.com/news/nation-world/national/article24750439.html>

Наводящие на размышления суждения о том, что может подталкивать людей к совершению поступков с ни для кого не желательными последствиями, можно найти в: Scott Alexander, *Meditations on Molloch* (<http://slatestarcodex.com/2014/07/30/meditations-on-moloch>)

Интерактивная хроника настораживающих событий, которые доказывают, что ядерная война может начаться случайно: <http://tinyurl.com/nukeoops>

О компенсационных выплатах жертвам ядерных испытаний в США см.:
https://www.justice.gov/sites/default/files/civil/legacy/2014/10/20/Tre_SysClaimsToDateSummary.pdf

Доклад комиссии США об угрозе электромагнитного импульса, возникающего в результате ядерного взрыва: http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf

Независимые исследования американских и советских ученых, предупреждавших Рейгана и Горбачева об опасности ядерной зимы:

– Crutzen, P. J. & Birks, J. W. (1982) The atmosphere after a nuclear war: Twilight at noon, *Ambio*, 11;

– Turco, R. P., Toon, O. B., Ackerman, T. P., Pollack, J. B. & Sagan, C. (1983) Nuclear winter: Global consequences of multiple nuclear explosions, *Science*, 222, 1283–1292;

– Aleksandrov, V. V. & Stenchikov, G. L. (1983) On the modeling of the climatic consequences of the nuclear war, *Proceeding on Applied Mathematics*, 21: Computing Centre of the USSR Academy of Sciences, Moscow (см. рус. вариант статьи: Александров В. В., Стенчиков Г. Л. Об одном вычислительном эксперименте, моделирующем климатические последствия ядерной войны // *Ж. вычисл. матем. и матем. физ.* 2014. 24:1, 140–144);

– Robock, A. (1984) Snow and ice feedbacks prolong effects of nuclearwinter, *Nature*, 310, 667–670.

Расчеты климатических изменений после ядерной войны см. в: Robock A., Oman, L. & Stenchikov, L. (2007). Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences, *J. Geophys. Res.*, 12, D13107.

Подборка ресурсов, посвященных сфере Дайсона, подготовленная Андерсом Зандбергом: <http://tinyurl.com/dysonsph>

Ключевая статья Фримена Дайсона о сферах, носящих его имя: Freeman Dyson (1959) Search for Artificial Stellar Sources of Infrared Radiation, *Science*, 131: 1667–1668.

Крейн и Вестморланд объясняют принцип работы двигателя на черной дыре в:
<http://arxiv.org/pdf/0908.1803.pdf>

Иконографика от CERN об элементарных частицах: <http://tinyurl.com/cernparticles>

В этом ролике показан безъядерный прототип, приводимый в движение взрывами бомб и разработанный в рамках проекта “Орион”. Он иллюстрирует идею ядерных взрывов в качестве ракетного двигателя: <https://www.youtube.com/watch?v=E3Lxx2VAYi8>

Учебное введение в принцип действия лазерного паруса:
<http://www.lunarsail.com/LightSail/rit-1.pdf>

Джей Олсон анализирует возможности космической экспансии цивилизаций:
<http://arxiv.org/abs/1411.4359>

Первый подробный анализ нашего далекого будущего: Dyson, Freeman J. (1979) Time without end: Physics and biology in an open universe, *Reviews of Modern Physics*, 51.3, 447, <http://tinyurl.com/dysonfuture> (см. рус. пер. в: *Далекое будущее Вселенной. Эсхатология в космической перспективе* / под ред. Дж. Эллис. М.: ББИ, 2012).

Упомянутая выше формула Сета Ллойда говорит нам, что выполнение вычислительной операции в течение времени τ требует расхода энергии в $E \geq h/4\tau$ где h – постоянная Планка. Если вы хотите выполнить одну за другой N операций, то $\tau = T/N$, откуда следует, что, расходуя энергию E за время T , можно выполнить $N \leq 2\sqrt{ET}/h$ операций. Таким образом, и время, и энергия – это ресурсы, которых хотелось бы иметь побольше. Если вы распределите свою энергию E между n параллельными вычислениями, то они будут осуществляться медленнее и эффективность вырастет. Ник Бострём оценивает количество операций, требуемых на симуляцию 100 лет человеческой жизни, в 10^{27} .

Если вы хотите получить веский аргумент, почему возникновение жизни требует довольно редкой флуктуации, из-за чего наши ближайшие соседи оказываются за 10^{1000} метров от нас, я советую посмотреть видео физика и астробиолога из Принстона Эдвина Тёрнера: <https://www.youtube.com/watch?v=Bt6n6Tu1beg>

Эссе Мартина Риса о поиске внеземных цивилизаций см.:
<https://www.edge.org/annual-question/2016/response/26665>

Популярное изложение статьи Джереми Ингланда о диссипативно направленной адаптации см. в: <https://www.scientificamerican.com/article/a-new-physics-theory-of-life/>

Основные идеи взяты из книги: Ilya Prigogine & Isabelle Stengers (1984) *Order Out of Chaos: Man's New Dialogue with Nature*, Bantam (см. рус. пер.: Пригожин И., Стенгерс И. *Порядок из хаоса: Новый диалог с природой* / пер. с англ. Ю. А. Данилова. М.: Прогресс, 1986. – Прим. перев.).

О чувствах и их физиологических причинах см.:

- William James (1890) *Principles of Psychology*, Henry Holt & Co.;
- Robert Ornstein (1992) *Evolution of Consciousness: The Origins of the Way We Think*, Simon & Schuster;
- António Damásio (2005) *Descartes' Error: Emotion, Reason, and the Human Brain*, Penguin;
- António Damásio (2012) *Self Comes to Mind: Constructing the Conscious Brain*, Vintage.

Элиезер Юдковски обсуждал возможность приведения в соответствие с целями дружественного искусственного интеллекта не просто наших целей, а наших целей (“волений”), гладко экстраполированных в будущее (coherent extrapolated volition, CEV). Образно говоря, это то, чего хотела бы идеализированная версия нас самих, если бы мы знали больше, думали быстрее и были бы в большей степени людьми, чем сами бы хотели. Юдковски начал критиковать идею CEV вскоре после предания ее гласности в 2004 году: (<http://intelligence.org/files/CEV.pdf>) и потому, что ее сложно применить, и потому, что неясны возможности преобразования ее во что-нибудь хорошо определенное.

В подходе обратного обучения с подкреплением ключевая идея состоит в том, что искусственный интеллект стремится к максимальному удовлетворению не своих целей, а целей его владельца-человека. Поэтому он должен проявлять осторожность в тех случаях, когда непонятно, чего в точности хочет владелец, и надо сначала с этим разобраться. Тогда проблемы с исключением не возникает: оно указывает на возникшее непонимание истинных намерений владельца.

См. статью Стива Омохундро: <http://tinyurl.com/omohundro2008>

Глубокая и содержательная книга Ханны Аренд о том, что бывает, когда разум слепо подчиняется приказам, не обращаясь к вопросам этики: Hanna Arendt (1963), *Eichmann in Jerusalem: A Report on the Banality of Evil*, Penguin (есть русский перевод: Арендт Х. *Банальность зла. Эйхман в Иерусалиме* / пер. с англ. С. Кастальский, Н. Рудницкая. М.: Европа, 2008. – Прим. перев.). Близкая дилемма возникает в работе Эрика Дрекслера (<http://www.fhi.ox.ac.uk/reports/2015-3.pdf>): чтобы удерживать сверхразум под своим контролем, его расчленяют на меньшие части, каждая из которых не в состоянии понять всю картину целиком. Если такое получится, то опять же возникает мощное орудие без внутренних моральных ограничителей, способное безропотно выполнять волю своего владельца. Это напоминает расчлененную бюрократию в условиях тиранической антиутопии: одна часть контролирует создание оружия, не зная, как оно будет использоваться, другая – казнит осужденных, не зная, в чем их вина, и т. п.

Современный вариант этого “золотого правила” предложил Джон Ролз: гипотетическая ситуация морально приемлема, если никто не хочет ее изменить, не зная заранее, какое место в ней займет.

Например, IQ многих людей в ближайшем окружении Гитлера был весьма высок:
<http://tinyurl.com/nurembergiq>

Словарная статья о сознании Стюарта Сазерленда довольно любопытна. См.: *Macmillan Dictionary of Psychology*, Macmillan, 1989.

Эрвин Шрёдингер, один из создателей квантовой механики, сделал это замечание в книге *Mind and Matter* (см.: Шрёдингер Э. *Разум и материя* / пер. с англ. А. В. Монаков. Ижевск: НИЦ "Регулярная и хаотическая динамика", 2000. – Прим. перев.), обсуждая *прошлое*: что бы произошло, если бы сознательная жизнь никогда не возникла и не заняла главенствующего места? С другой стороны, возникновение искусственного интеллекта ставит логическую проблему, связанную с тем, что теперь мы сами можем начать разыгрывать пьесу перед пустыми креслами в будущем.

101 Стэнфордская философская энциклопедия приводит подробный обзор различных способов определения и использования слова "сознание". См.: <http://tinyurl.com/stanfordconsciousness>

Yuval Noah Harari (2017) *Homo Deus*, p. 116.

Прекрасное введение в теорию “Системы 1” и “Системы 2” см.: Daniel Kahneman (2011) *Thinking, Fast and Slow*, Farrar, Straus & Giroux.

См.: Christof Koch (2004) *The Quest for Consciousness: A Neurobiological Approach*, W. H. Freeman.

Мы осознаем лишь незначительную часть информации (скажем, около 10–50 бит), ежесекундно проходящей через наш мозг. См.: Küpfmüller, K. (1962) Nachrichtenverarbeitung im Menschen, in: Steinbuch, K. (Ed.) *Taschenbuch der Nachrichtenverarbeitung*, 1481–1502.

Nørretranders, T. (1991) *The User Illusion: Cutting Consciousness Down to Size*, Viking.

- Michio Kaku (2014) *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind*, Doubleday
- Jeff Hawkins & Sandra Blakeslee (2007) *On Intelligence*, Times Books;
 - Stanislas Dehaene, Michel Kerszberg & Jean-Pierre Changeux (1998) A neuronal model of a global workspace in effortful cognitive tasks, *Proceedings of the National Academy of Sciences*, 95, 14529–14534.

Знаменитый эксперимент Пенфилда “Я чувствую запах подгоревшего тоста”, см.: <https://www.youtube.com/watch?v=pUOG2g4hj8s>

Детали о сенсоромоторных отделах коры головного мозга см.: Elaine Marieb & Katja Hoehn (2008) *Anatomy & Physiology*, 3rd Ed., Pearson, p. 391–395.

Изучение нейронных коррелятов сознания (NCCs) стало занимать в последние годы центральное место среди интересов нейробиологического сообщества. См., например:

Geraint Rees, Gabriel Kreiman & Christof Koch (2002) Neural correlates of consciousness in humans, *Nature Reviews Neuroscience*, 3, 261–270, и Thomas Metzinger (2000) *Neural correlates of consciousness: Empirical and conceptual questions*, MIT press.

О том, как работает непрерывное подавление вспышек, см.:

- Christof Koch (2004) *The Quest for Consciousness: A Neurobiological Approach*, W. H. Freeman;
- Christof Koch & Naotsugu Tsuchiya (2005) Continuous flash suppression reduces negative afterimages, *Nature Neuroscience*, 8, 1096–1101.

См.: Christof Koch, Marcello Massimini, Melanie Boly & Giulio Tononi (2016) Neural correlates of consciousness: progress and problems, *Nature Reviews Neuroscience*, 17, 307.

См. с. 250 в: Christof Koch (2004) *The Quest for Consciousness: A Neurobiological Approach*, Op. cit, а также дальнейшее обсуждение вопроса в Стэнфордской философской энциклопедии (<http://tinyurl.com/consciousnessdelay>).

О синхронизации осознанного восприятия см.:

- David Eagleman (2015) *The Brain: The Story of You*, Pantheon;
- *Stanford Encyclopedia of Philosophy*, <http://tinyurl.com/consciousnesssync>

См.: Benjamin Libet (2004) *Mind Time – The Temporal Factor in Consciousness*, Harvard University Press, Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, John-Dylan Haynes (2008) Unconscious determinants of free decisions in the human brain, *Nature Neuroscience*, 11, 543–545, (<http://www.nature.com/neuro/journal/v11/n5/full/nn.2112.html>)

См. примеры теоретических подходов к исследованию сознания в:

- Daniel Dennett (1992) *Consciousness explained*, Back Bay Books;
- Bernard Baars (2001) *In the Theater of Consciousness: The Workspace of the Mind*, Oxford Univ. Press;
- Christof Koch (2004) *The Quest for Consciousness: A Neurobiological Approach*, Roberts;
- Gerald Edelman & Giulio Tononi (2008) *A Universe Of Consciousness How Matter Becomes Imagination: How Matter Becomes Imagination*, Hachette;
- António Damásio (2012) *Self Comes to Mind: Constructing the Conscious Brain*, Vintage;
- Stanislas Dehaene (2014) *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*, Viking;
- Stanislas Dehaene, Michel Kerszberg & Jean-Pierre Changeux (1998) A neuronal model of a global workspace in effortful cognitive tasks, *Proceedings of the National Academy of Sciences*, 95, 14529–14534;
- Stanislas Dehaene, Lucie Charles, Jean-Rémi King & Sébastien Marti (2014) Toward a computational theory of conscious processing, *Current opinion in neurobiology*, 25, 760–84.

Основательное обсуждение использования термина “emergence” в физике и философии см. в статье Дэвида Чалмерса: <http://cse3521.artifice.cc/Chalmers-Emergence.pdf>

Мои работы, где я доказываю, что сознание – это то, что происходит с информацией при ее обработке определенным сложным образом:

– <https://arxiv.org/abs/physics/0510188>

– <https://arxiv.org/abs/0704.0646>

Max Tegmark (2014) *Our Mathematical Universe*, Random House.

Дэвид Чалмерс формулирует близкую точку зрения в своей книге 1996 года *The Conscious Mind*: “Переживание – это информация изнутри, физика – это информация снаружи”.

См.: Adenauer Casali et al. (2013) A theoretically based index of consciousness independent of sensory processing and behavior, *Science translational medicine*, 5, 198ra105, <http://tinyurl.com/zapzip>

Теория интегральной информации (ИИ) не работает в непрерывных системах. См.:

- <https://arxiv.org/abs/1401.1219>
- <http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00063/full>
- <https://arxiv.org/abs/1601.02626>

Интервью с Клайвом Уэрингом, кратковременная память которого всего 30 секунд,
см.: <https://www.youtube.com/watch?v=WmzU47i2xgw>

Критика ИТ Скоттом Аарансоном, см.: <http://www.scottaaronson.com/blog/?p=1799>

Критика ИТ Черрулло, который утверждает, что интеграция – недостаточное условие для сознания, см.: <http://tinyurl.com/cerrullocritique>

Предсказание на основе ИТ, что симулированные люди будут зомби, см.:
<http://rstb.royalsocietypublishing.org/content/370/1668/20140167>

Критика ИТ Шанаханом, см.: <http://arxiv.org/ftp/arxiv/papers/1504/1504.05696.pdf>

О слепом пятне см.: <http://tinyurl.com/blindsight-paper>

См. сн. 105.

“Сознание без доступа”, за и против, см.: Victor Lamme (2010), How neuroscience will change our view on consciousness, *Cognitive Neuroscience*, 204–20 (<http://www.tandfonline.com/doi/abs/10.1080/17588921003731586>).

См.: <https://www.youtube.com/watch?v=vJG698U2Mvo>

См. сн. 126.

Подробное обсуждение этих вопросов см. в книге Дэниела Деннета *Consciousness explained*.

См. сн. 103.

См. по поводу свободы воли в *Стэнфордской философской энциклопедии*:
<https://plato.stanford.edu/entries/freewill>

Видео, в котором Сет Ллойд объясняет, почему искусственный интеллект должен ощущать себя обладающим свободной волей, см.: <https://www.youtube.com/watch?v=Eрj3DF8jDWk>

См. *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature* (см. рус. пер.: Вайнберг С. *Мечты об окончательной теории: физика в поисках самых фундаментальных законов природы* / пер. с англ. А. В. Берков. М.: УРСС, 2004. – Прим. перев.).

См. сн. 91.

За открытым письмом (<http://futureoflife.org/ai-open-letter/>), написанным по итогам конференции в Пуэрто-Рико, о том, что проведение исследований, направленных на достижение надежности и дружелюбности искусственного интеллекта, необходимо и своевременно, появился документ о приоритетах, где перечислены существующие уже сегодня конкретные исследовательские направления, которые позволяют двигаться к этой цели: http://futureoflife.org/data/documents/research_priorities.pdf

Мое видеоинтервью с Илоном Маском о AI-безопасности:
<https://www.youtube.com/watch?v=rBw0eoZTY-g>

Видеокомпиляция практически всех попыток посадить ракету SpaceX, включая успешную, см.:

<https://www.youtube.com/watch?v=AllaFzIPaG4>

Сообщение о нашем конкурсе по AI-безопасности Маск опубликовал в своем Twitter:
<https://twitter.com/elonmusk/status/555743387056226304>

Твитт Маска о нашем открытом письме, призывающем к исследованиям по AI-безопасности: <https://twitter.com/elonmusk/status/554320532133650432>

Статья, подшучивающая над новостной бурей в прессе по поводу нашего открытого письма, см.: <http://www.popsoci.com/open-letter-everyone-tricked-fearing-ai>

Твитт Маска о его большом пожертвовании в пользу нашего института
и исследований по AI-безопасности:
<https://twitter.com/elonmusk/status/555743387056226304>

Сайт Партнерства по созданию искусственного интеллекта на благо людей и общества (Partnership on AI to benefit people and society), см.: <https://www.partnershiponai.org>

Примеры недавних отчетов по искусственному интеллекту, см.:

- 100-летняя программа исследований искусственного интеллекта в Стэнфорде:
<http://tinyurl.com/stanfordai>
- Отчет Белого дома о будущем искусственного интеллекта:
<http://tinyurl.com/obamaAIreport>
- Отчет Белого дома о воздействии искусственного интеллекта на рынок труда:
<http://tinyurl.com/AIjobsreport>
- Отчет IEEE об искусственном интеллекте и человеческом благосостоянии:
http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf
- Дорожная карта US Robotics: <http://tinyurl.com/roboticsmap>

Среди принципов, не попавших в итоговый документ, один из самых мной любимых был такой: “Пока нет общего согласия, мы должны избегать сильных допущений относительно возможности искусственного интеллекта обладать сознанием и чувствами или требовать их для себя”. Он прошел много итераций, и на последней из них вызывавшее споры слово “сознание” было заменено на “субъективные переживания”, и все равно этот принцип набрал только 88 % голосов, лишь немного недотянув до требуемых 90 %.

Дискуссионная панель с Илоном Маском и другими великими умами, см.:
<http://tinyurl.com/asilomarAI>