Рациональные и убедительные доводы Ника Бострома об опасности создания искусственного интеллекта заставят задуматься даже заядлых скептиков.

Евгений Касперский,
генеральный директор «Лаборатории Касперского»

Ник Бостром

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

ЭТАПЫ. УГРОЗЫ. СТРАТЕГИИ

Nick Bostrom

Superintelligence

Paths, Dangers, Strategies

Ник Бостром

Искусственный интеллект

Этапы. Угрозы. Стратегии

Бостром, Ник
Искусственный интеллект. Этапы. Угрозы. Стратегии / Ник Бостром ; пер. с англ. С. Филина.

Что случится, если машины превзойдут людей в интеллекте? Они будут помогать нам или уничтожат человеческую расу? Можем ли мы сегодня игнорировать проблему развития искусственного интеллекта и чувствовать себя в полной безопасности?
В своей книге Ник Бостром пытается осознать проблему, встающую перед человечеством в связи с перспективой появления сверхразума, и проанализировать его ответную реакцию.

Оглавление

Предисловие

> …У меня есть один знакомый, — сказал Эдик. — Он утверждает, будто человек — промежуточное звено, необходимое природе для создания венца творения: рюмки коньяка с ломтиком лимона.
>
> *Аркадий и Борис Стругацкие. Понедельник начинается в субботу*

Компьютеры, а точнее алгоритмы, опирающиеся на непрерывно растущие вычислительные мощности, лучше людей играют в шахматы, шашки и нарды. Они очень неплохо водят самолеты. Они смогли пройти тест Тьюринга, убедив судей в своей «человечности». Однажды таксист в Дублине — городе, где расположены европейские штаб-квартиры многих глобальных IT-компаний, — сказал мне, что приветствует бурное развитие технологического сектора своей страны, но потом с сожалением добавил: «Одна беда — из-за этих умных ребят довольно скоро таксисты будут не нужны». Автомобили без водителей, управляемые компьютерами, уже проходят испытания на обычных дорогах в нескольких странах. По мнению философа Ника Бострома, чью книгу вы держите в руках, — все это звенья одной цепи и довольно скоро из-за развития компьютерных технологий нам всем, человеческому роду, может прийти конец.

Автор считает, что смертельная угроза связана с возможностью создания искусственного интеллекта, превосходящего человеческий разум. Катастрофа может разразиться как в конце XXI века, так и в ближайшие десятилетия. Вся история человечества показывает: когда происходит столкновение представителя нашего вида, человека разумного, и любого другого, населяющего нашу планету, побеждает тот, кто умнее. До сих пор умнейшими были мы, но у нас нет гарантий, что так будет длиться вечно.

Ник Бостром пишет, что если умные компьютерные алгоритмы научатся самостоятельно делать еще более умные алгоритмы, а те, в свою очередь, еще более умные, случится взрывной рост искусственного интеллекта, по сравнению с которым люди будут выглядеть приблизительно как сейчас муравьи рядом с людьми, в интеллектуальном смысле, конечно. В мире появится новый, хотя и искусственный, но сверхразумный вид. Неважно, что ему «придет в голову», попытка сделать всех людей счастливыми или решение остановить антропогенное загрязнение мирового океана наиболее эффективным путем, то есть уничтожив человечество, — все равно сопротивляться этому у людей возможности не будет. Никаких шансов на противостояние в духе кинофильма про Терминатора, никаких перестрелок с железными киборгами. Нас ждет шах и мат — как в поединке шахматного компьютера «Дип Блю» с первоклассником.

За последнюю сотню-другую лет достижения науки у одних пробуждали надежду на решение всех проблем человечества, у других вызывали и вызывают безудержный страх. При этом, надо сказать, обе точки зрения выглядят вполне оправданными. Благодаря науке побеждены страшные болезни, человечество способно сегодня прокормить невиданное прежде количество людей, а из одной точки земного шара можно попасть в противоположную меньше чем за сутки. Однако по милости той же науки люди, используя новейшие военные технологии, уничтожают друг друга с чудовищной скоростью и эффективностью.

Подобную тенденцию — когда быстрое развитие технологий не только приводит к образованию новых возможностей, но и формирует небывалые угрозы, — мы наблюдаем и в области информационной безопасности. Вся наша отрасль возникла и существует исключительно потому, что создание и массовое распространение таких замечательных вещей, как компьютеры и интернет, породило проблемы, которые было бы невозможно вообразить в докомпьютерную эру. В результате появления информационных технологий произошла революция в человеческих коммуникациях. В том числе ею воспользовались разного рода киберпреступники. И только сейчас человечество начинает постепенно осознавать новые риски: все больше объектов физического мира управляются с помощью компьютеров и программного обеспечения, часто несовершенного, дырявого и уязвимого; все большее число таких объектов имеют связь с интернетом, и угрозы

кибермира быстро становятся проблемами физической безопасности, а потенциально — жизни и смерти.

Именно поэтому книга Ника Бострома кажется такой интересной. Первый шаг для предотвращения кошмарных сценариев (для отдельной компьютерной сети или всего человечества) — понять, в чем они могут состоять. Бостром делает очень много оговорок, что создание искусственного интеллекта, сравнимого с человеческим разумом или превосходящего его, — искусственного интеллекта, способного уничтожить человечество, — это лишь вероятный сценарий, который может и не реализоваться. Конечно, вариантов много, и развитие компьютерных технологий, возможно, не уничтожит человечество, а даст нам ответ на «главный вопрос жизни, Вселенной и всего такого» (возможно, это и впрямь окажется число 42, как в романе «Автостопом по Галактике»). Надежда есть, но опасность очень серьезная — предупреждает нас Бостром. На мой взгляд, если вероятность такой экзистенциальной угрозы человечеству существует, то отнестись к ней надо соответственно и, чтобы предотвратить ее и защититься от нее, следует предпринять совместные усилия в общемировом масштабе.

Завершить свое вступление хочется цитатой из книги Михаила Веллера «Человек в системе»:

> Когда фантастика, то бишь оформленная в образы и сюжеты мысль человеческая, долго и детально что-то повторяет — ну так дыма без огня не бывает. Банальные голливудские боевики о войнах людей с цивилизацией роботов несут в себе под шелухой коммерческого смотрива горькое зернышко истины.
>
> Когда в роботы будет встроена передаваемая программа инстинктов, и удовлетворение этих инстинктов будет встроено как безусловная и базовая потребность, и это пойдет на уровень самовоспроизводства — вот тогда, ребята, кончай бороться с курением и алкоголем, потому что будет самое время выпить и закурить перед ханой всем нам.

Евгений Касперский,
генеральный директор «Лаборатории Касперского»

Неоконченная история о воробьях

Однажды, в самый разгар гнездования, утомленные многодневным тяжким трудом воробьи присели передохнуть на заходе солнца и пощебетать о том о сем.

— Мы такие маленькие, такие слабые. Представьте, насколько проще было бы жить, держи мы в помощниках сову! — мечтательно прочирикал один воробей. — Она могла бы вить нам гнезда…

— Ага! — согласился другой. — А еще присматривать за нашими стариками и птенцами…

— И наставлять нас, и защищать от соседской кошки, — добавил третий.

Тогда Пастус, самый старший воробей, предложил:

— Пусть разведчики полетят в разные стороны на поиски выпавшего из гнезда совенка. Впрочем, подойдет и совиное яйцо, и вороненок, и даже детеныш ласки. Эта находка обернется для нашей стаи самой большой удачей! Вроде той, когда мы обнаружили на заднем дворе неоскудевающий источник зерна.

Возбудившиеся не на шутку воробьи расчирикались что было мочи.

И только одноглазый Скронфинкл, въедчивый, с тяжелым нравом воробей, похоже, сомневался в целесообразности данного предприятия.

— Мы избрали гибельный путь, — убежденно промолвил он. — Разве не следует сначала серьезно проработать вопросы укрощения и одомашнивания сов, прежде чем впускать в свою среду такое опасное существо?

— Сдается мне, — возразил ему Пастус, — искусство приручения сов — задача не из простых. Найти совиное яйцо — и то чертовски сложно.

Так что давайте начнем с поиска. Вот сумеем вывести совенка, тогда и задумаемся о проблемах воспитания.

— Порочный план! — нервно чирикнул Скронфинкл.

Но его уже никто не слушал. По указанию Пастуса воробьиная стая поднялась в воздух и отправилась в путь.

На месте остались лишь воробьи, решившие все-таки выяснить, как приручать сов. Довольно быстро они поняли правоту Пастуса: задача оказалась неимоверно сложной, особенно в отсутствие самой совы, на которой следовало бы практиковаться. Однако птицы старательно продолжали изучать проблему, поскольку опасались, что стая вернется с совиным яйцом прежде, чем им удастся открыть секрет, каким образом можно контролировать поведение совы.

Автору неизвестно, чем закончилась эта история,
но он посвящает свою книгу Скронфинклу и всем его последователям.

Введение

Внутри нашего черепа располагается некая субстанция, благодаря которой мы можем, например, читать. Указанная субстанция — человеческий мозг — наделена возможностями, отсутствующими у других млекопитающих. Собственно, своим доминирующим положением на планете люди обязаны именно этим характерным особенностям. Некоторых животных отличает мощнейшая мускулатура и острейшие клыки, но ни одно живое существо, кроме человека, не одарено настолько совершенным умом. В силу более высокого интеллектуального уровня нам удалось создать такие инструменты, как язык, технология и сложная социальная организация. С течением времени наше преимущество лишь укреплялось и расширялось, поскольку каждое новое поколение, опираясь на достижения предшественников, шло вперед.

Если когда-нибудь разработают искусственный разум, превосходящий общий уровень развития человеческого разума, то в мире появится сверхмощный интеллект. И тогда судьба нашего вида окажется в прямой зависимости от действий этих разумных технических систем — подобно тому, как сегодняшняя участь горилл в большей степени определяется не самими приматами, а людскими намерениями.

Однако человечество действительно обладает неоспоримым преимуществом, поскольку оно и создает разумные технические системы. В принципе, кто мешает придумать такой сверхразум, который возьмет под свою защиту общечеловеческие ценности? Безусловно, у нас имеются весьма веские основания, чтобы обезопасить себя. В практическом плане нам придется справиться с труднейшим вопросом контроля — как управлять замыслами и действиями сверхразума. Причем люди смогут использовать один-единственный шанс. Как только недружественный искусственный интеллект (ИИ) появится на свет, он сразу начнет препятствовать нашим усилиям избавиться от него

или хотя бы откорректировать его установки. И тогда судьба человечества будет предрешена.

В своей книге я пытаюсь осознать проблему, встающую перед людьми в связи с перспективой появления сверхразума, и проанализировать их ответную реакцию. Пожалуй, нас ожидает самая серьезная и пугающая повестка, которую когда-либо получало человечество. И независимо от того, победим мы или проиграем, — не исключено, что этот вызов станет для нас последним. Я не привожу здесь никаких доводов в пользу той или иной версии: стоим ли мы на пороге великого прорыва в создании искусственного интеллекта; возможно ли с определенной точностью прогнозировать, когда свершится некое революционное событие. Вероятнее всего — в нынешнем столетии. Вряд ли кто-то назовет более конкретный срок.

В первых двух главах я рассмотрю разные научные направления и слегка затрону такую тему, как темпы экономического развития. Однако в основном книга посвящена тому, что произойдет после появления сверхразума. Нам предстоит обсудить следующие вопросы: динамику взрывного развития искусственного интеллекта; его формы и потенциал; варианты стратегического выбора, которыми он будет наделен и вследствие которых получит решающее преимущество. После этого мы проанализируем проблему контроля и попытаемся решить важнейшую задачу: возможно ли смоделировать такие исходные условия, которые позволят нам сохранить собственное превосходство и в итоге выжить. В последних главах мы отойдем от частностей и посмотрим на проблему шире, чтобы охватить в целом ситуацию, сложившуюся в результате нашего изучения. Я предложу вашему вниманию некоторые рекомендации, что следует предпринять уже сегодня, дабы в будущем избежать катастрофы, угрожающей существованию человечества.

Писать эту книгу было нелегко. Надеюсь, что пройденный мною путь пойдет на пользу другим исследователям. Они без лишних препятствий достигнут новых рубежей и полные сил смогут быстрее включиться в работу, благодаря которой люди полностью осознают всю сложность стоящей перед ними проблемы. (Если все-таки дорога изучения покажется будущим аналитикам несколько извилистой и местами изрытой ухабами, надеюсь, они оценят, насколько непроходимым был ландшафт *прежде*.)

Невзирая на сложности, связанные с работой над книгой, я старался излагать материал доступным языком; правда, сейчас вижу, что не вполне с этим справился. Естественно, пока я писал, то мысленно обращался

к потенциальному читателю и почему-то всегда в данной роли представлял себя, только несколько моложе настоящего, — получается, я делал книгу, которая могла бы вызвать интерес прежде всего у меня самого, но не обремененного прожитыми годами. Возможно, именно это определит в дальнейшем малочисленность читательской аудитории. Тем не менее, на мой взгляд, содержание книги будет доступно многим людям. Надо лишь приложить некоторые умственные усилия, перестать с ходу отвергать новые идеи и воздерживаться от искушения подменять все непонятное удобными стереотипами, которые мы все легко выуживаем из своих культурных запасов. Читателям, не обладающим специальными знаниями, не стоит пасовать перед встречающимися местами математическими выкладками и незнакомыми терминами, поскольку контекст всегда позволяет понять основную мысль. (Читатели, желающие, напротив, узнать больше подробностей, найдут много интересного в примечаниях[1].)

Вероятно, многое в книге изложено некорректно[2]. Возможно, я упустил из виду какие-то важные соображения, в результате чего некоторые мои заключения — а может быть, и все — окажутся ошибочными. Чтобы не пропустить мельчайший нюанс и обозначить степень неопределенности, с которой мы имеем дело, мне пришлось обратиться к специфическим маркерам — поэтому мой текст перегружен такими уродливыми словесными кляксами, как «возможно», «могло бы», «может быть», «похоже», «вероятно», «с большой долей вероятности», «почти наверняка». Однако я всякий раз прибегаю к помощи вводных слов крайне осторожно и весьма продуманно. Впрочем, для обозначения общей ограниченности гносеологических допущений одного такого стилистического приема явно недостаточно; автор должен выработать системный подход, чтобы рассуждать в условиях неопределенности и прямо указывать на возможность ошибки. Речь ни в коей мере не идет о ложной скромности. Искренне признаю, что в моей книге могут быть и серьезные заблуждения, и неверные выводы, но при этом я убежден: альтернативные точки зрения, представленные в литературе, — еще хуже. Причем это касается и общепринятой «нулевой гипотезы», согласно которой на сегодняшний день мы можем с абсолютным основанием игнорировать проблему появления сверхразума и чувствовать себя в полной безопасности.

Глава первая

Прошлые достижения и сегодняшние возможности

Начнем с обращения к далекому прошлому. В общих чертах история представляет собой последовательность различных моделей роста, причем процесс носит прогрессивно ускоряющийся характер. Эта закономерность дает нам право предполагать, что возможен следующий — еще более быстрый — период роста. Однако вряд ли стоит придавать слишком большое значение подобному соображению, поскольку тема нашей книги — не «технологическое ускорение», не «экспоненциальный рост» и даже не те явления, которые обычно подаются под понятием «сингулярность». Далее мы обсудим историю вопроса: как развивались исследования по искусственному интеллекту. Затем перейдем к текущей ситуации: что сегодня происходит в этой области. И наконец, остановимся на некоторых последних оценках специалистов и поговорим о нашей неспособности прогнозировать сроки дальнейшего развития событий.

Модели роста и история человечества

Всего несколько миллионов лет назад предки людей еще жили в кронах африканских деревьев, перепрыгивая с ветки на ветку. Появление *Homo sapiens*, или человека разумного, отделившегося от наших общих с человекообразными обезьянами предков, с геологической и даже эволюционной точки зрения происходило очень плавно. Древние люди принимали вертикальное положение, а большие пальцы на их кистях стали заметно отстоять от остальных. Однако самое главное — происходили относительно незначительные изменения в объеме мозга и организации нервной системы, что в конце концов привело к гигантскому рывку в умственном развитии человека. Как следствие, у людей появилась способность к абстрактному

мышлению. Они начали не только стройно излагать сложные мысли, но и создавать информационную культуру, то есть накапливать сведения и знания и передавать их от поколения к поколению. Надо сказать, человек научился делать это значительно лучше любых других живых существ на планете.

Древнее человечество, используя появившиеся у него способности, разрабатывало все более и более рациональные способы производства, благодаря чему смогло мигрировать далеко за пределы джунглей и саванн. Сразу после возникновения земледелия стремительно начали расти величина населения и его плотность. Больше народа — больше идей, причем высокая плотность способствовала не только быстрому распространению новых веяний, но и появлению разных специалистов, а это означало, что в среде людей шло постоянное совершенствование профессиональных навыков. Данные факторы повысили *темпы экономического развития*, сделали возможным рост производительности и формирование технического потенциала. В дальнейшем такой же по значимости прогресс, приведший к промышленной революции, вызвал второй исторический скачок в ускорении темпа роста.

Такая динамика темпа роста имела важные последствия. Например, на заре человечества, когда Землю населяли прародители современных людей, или гоминиды*, экономическое развитие происходило слишком медленно, и потребовалось порядка миллиона лет для прироста производственных мощностей, чтобы население планеты позволило себе увеличиться на миллион человек, причем существовавших на грани выживания. А после неолитической революции, к 5000 году до н. э., когда человечество перешло от охотничье-собирательского общества к сельскохозяйственной экономической модели, темпы роста выросли настолько, что для такого же прироста населения хватило двухсот лет. Сегодня, после промышленной революции, мировая экономика растет примерно на ту же величину каждые полтора часа[1].

Существующий темп роста — даже если он законсервируется на относительно продолжительное время — приведет к впечатляющим результатам. Допустим, мировая экономика продолжит расти со средним темпом, харак-

---

* *Гоминиды* (лат. *Hominidae*) — высокоорганизованное семейство человекообразных обезьян; гоминид, человек ископаемый, представляет собой промежуточное звено между приматом и человеком разумным. *Здесь и далее: прим. ред.*

терным для последних пятидесяти лет, все равно население планеты в будущем станет богаче, чем сегодня: к 2050 году — в 4,8 раза, а к 2100 году — в 34 раза[2].

Однако перспективы стабильного экспоненциального роста меркнут в сравнении с тем, что может произойти, когда в мире свершится следующее скачкообразное изменение, темп развития которого по значимости и последствиям будет сравним с неолитической и промышленной революциями. По оценкам экономиста Робина Хэнсона, основанным на исторических данных о хозяйственной деятельности и численности населения, время удвоения экономик охотничье-собирательского общества эпохи плейстоцена составляло 224 тысячи лет, аграрного общества — 909 лет, индустриального общества — 6,3 года[3]. (В соответствии с парадигмой Хэнсона современная экономическая модель, имеющая смешанную аграрно-индустриальную структуру, еще не развивается в удвоенном темпе каждые 6,3 года.) Если в мировом развитии уже случился бы такой скачок, сопоставимый по своему революционному значению с двумя предыдущими, то экономика вышла бы на новый уровень и удваивала бы темпы роста примерно каждые две недели.

С точки зрения сегодняшнего дня подобные темпы развития кажутся фантастическими. Но и свидетели минувших эпох тоже вряд ли могли предположить, что темпы роста мировой экономики когда-нибудь будут удваиваться несколько раз на протяжении жизни одного поколения. То, что для них представлялось совершенно немыслимым, нами воспринимается как норма.

Идея приближения момента технологической сингулярности стала чрезвычайно популярной после появления новаторских работ Вернона Винджа, Рэя Курцвейла и других исследователей[4]. Впрочем, понятие «сингулярность», которое используется в самых разных значениях, уже приобрело устойчивый смысл в духе технологического утопизма и даже обзавелось ореолом чего-то устрашающего и в тоже время вполне величественного[5]. Поскольку большинство определений слова *сингулярность* не имеют отношения к предмету нашей книги, мы достигнем большей ясности, если избавимся от него в пользу более точных терминов.

Интересующая нас идея, связанная с понятием сингулярности, — это потенциальное *взрывоподобное развитие интеллекта*, особенно в перспективе создания искусственного сверхразума. Возможно, представленные на рис. 1 кривые роста убедят кого-то из вас, что мы стоим на пороге нового

интенсивного скачка в темпе развития — скачка, сопоставимого с неолитической и промышленной революциями. Скорее всего, людям, доверяющим диаграммам, даже трудно вообразить сценарий, в котором время удвоения мировой экономики сокращается до недель без участия сверхмощного разума, во много раз превосходящего по скорости и эффективности своей работы наш привычный биологический ум. Однако не обязательно упражняться в рисовании кривых роста и экстраполяции исторических темпов экономического развития, чтобы начать ответственно относиться к революционному появлению искусственного интеллекта. Эта проблема настолько серьезна, что не нуждается в аргументации подобного рода. Как мы увидим, есть гораздо более веские причины проявлять осмотрительность.

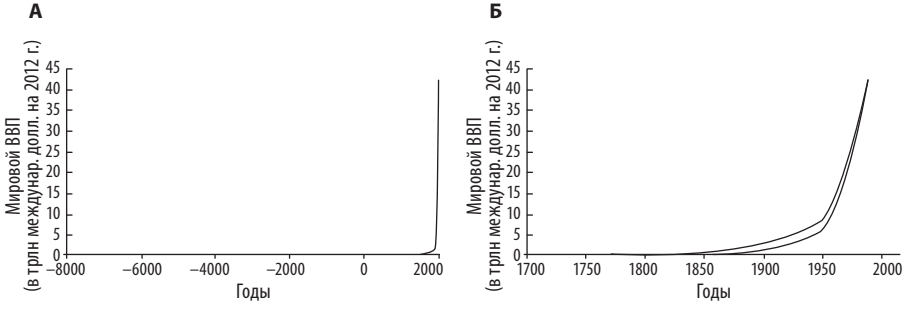

Рис. 1. Динамика мирового ВВП за длительный исторический период. На линейной шкале история мировой экономики отображена как линия, сначала почти сливающаяся с горизонтальной осью, а затем резко устремляющаяся вертикально вверх. **А.** Даже расширив временные границы до десяти тысяч лет в прошлое, мы видим, что линия делает рывок вверх из определенной точки почти под девяносто градусов. **Б.** Линия заметно отрывается от горизонтальной оси только на уровне приблизительно последних ста лет. (Разность кривых на диаграммах объясняется разным набором данных, поэтому и показатели несколько отличаются друг от друга[6].)

Завышенные ожидания

С момента изобретения в 1940-х годах первых электронно-вычислительных машин люди не перестают прогнозировать появление компьютера, уровень интеллекта которого будет сравним с человеческим. Имеется в виду разумная техническая система, наделенная здравым смыслом, обладающая способностью к обучению и размышлению, умеющая планировать и комплексно обрабатывать информацию, собранную по самым разным источникам —

реальным и теоретическим. В те времена многие ожидали, что такие машины станут реальностью уже лет через двадцать[7]. С тех пор сроки сдвигаются со скоростью одного года в год, то есть сегодня футурологи, убежденные в вероятности создания искусственного интеллекта, продолжают верить, что «умные машины» появятся через пару десятков лет[8].

Срок в двадцать лет любим всеми предсказателями коренных перемен. С одной стороны, это не слишком долго — и потому предмет обсуждения привлекает к себе широкое внимание; с другой стороны, это не так быстро, что дает возможность помечтать о целом ряде важнейших научных открытий — правда, представления о них на момент прогнозирования весьма расплывчаты, но их реализация практически не вызывает сомнения. Сопоставим это с более короткими прогностическими сроками, установленными для разных технологий, которым суждено оказать значительное влияние на мир: от пяти до десяти лет — на момент прогнозирования большинство технических решений уже частично применяются; пятнадцать лет — на момент прогнозирования эти технологии уже существуют в виде лабораторных версий. Кроме того, двадцатилетний срок чаще всего близок к средней продолжительности оставшейся профессиональной деятельности прогнозиста, что уменьшает репутационный риск, связанный с его дерзким предсказанием.

Впрочем, из-за слишком завышенных и несбывшихся ожиданий прошлых лет не следует сразу делать вывод, что создание искусственного интеллекта невозможно в принципе и что никто никогда не будет его разрабатывать[9]. Основная причина, почему прогресс шел медленнее, чем предполагалось, связана с техническими проблемами, возникавшими при разработке разумных машин. Первопроходцы не предусмотрели всех трудностей, с которыми им пришлось столкнуться. Причем вопросы: велика ли степень серьезности этих препятствий и насколько мы далеки от их преодоления — до сих пор остаются открытыми. Порою задачи, первоначально кажущиеся безнадежно сложными, имеют удивительно простое решение (хотя чаще, пожалуй, бывает наоборот).

Мы рассмотрим пути, которые могут привести к появлению искусственного интеллекта, не уступающего человеческому, в следующей главе. Но уже сейчас хотелось бы обратить ваше внимание на один важный аспект. Нас ожидает много остановок между нынешним отправным пунктом и тем будущим, когда появится искусственный интеллект, но этот момент — отнюдь не конечная станция назначения. Довольно близкой от нее следующей оста-

новкой будет станция «Сверхразум» — осуществление искусственного интеллекта такого уровня, который не просто равен человеческому уму, а значительно превосходит его. После последней остановки наш поезд разгонится до такой степени, что у станции «Человек» не сможет не только остановиться, но даже замедлить ход. Скорее всего, он со свистом промчится мимо. Британский математик Ирвинг Джон Гуд, работавший во времена Второй мировой войны шифровальщиком в команде Алана Тьюринга, скорее всего, был первым, кто изложил важнейшие подробности этого сценария. В своей часто цитируемой статье 1965 года о первых сверхразумных машинах он писал:

> Давайте определим сверхразумную машину как машину, которая в значительной степени превосходит интеллектуальные возможности любого умнейшего человека. Поскольку создание таких машин является результатом умственной деятельности человека, то машина, наделенная сверхразумом, будет способна разрабатывать еще более совершенные машины; вследствие этого, бесспорно, случится такой «интеллектуальный взрыв», что человеческий разум окажется отброшенным далеко назад. Таким образом, первая сверхразумная машина станет последним достижением человеческого ума — правда, лишь в том случае, если она не обнаружит достаточную сговорчивость и сама не объяснит нам, как держать ее под контролем[10].

Взрывное развитие искусственного интеллекта может повлечь за собой один из главных экзистенциальных рисков* — в наши дни такое положение вещей воспринимается как тривиальное; следовательно, перспективы подобного роста должны оцениваться с крайней серьезностью, даже если было бы заведомо известно (но это не так), что вероятность угрозы относительно низка. Однако пионеры в области искусственного интеллекта, несмотря на всю убежденность в неминуемом появлении искусственного интеллекта, не уступающего человеческому, в массе своей отрицали возможность появления сверхразума, превосходящего человеческий ум. Создается впечатление, что их воображение — в попытках постичь предельную возможность будущих машин, сравнимых по своим мыслительным способностям с человеком, — просто иссякло, и они легко прошли мимо неизбежного вывода: дальнейшим шагом станет рождение сверхразумных машин.

---

* Согласно Центру по изучению экзистенциальных рисков (Кембридж), таковыми считаются потенциальные угрозы для человечества: искусственный интеллект, изменение климата, ядерное оружие, биотехнологии.

Большинство первопроходцев не поддерживали зарождавшееся в обществе беспокойство, считая полной ерундой, будто их проекты несут в себе определенный риск для человечества[11]. Никто из них ни на словах, ни на деле — ни одного серьезного исследования на эту тему — не пытался осмыслить ни тревогу по поводу безопасности, ни этические сомнения, связанные с созданием искусственного интеллекта и потенциального доминирования компьютеров; данный факт вызывает удивление даже на фоне характерных для той эпохи не слишком высоких стандартов оценки новых технологий[12]. Остается только надеяться, что ко времени, когда их смелый замысел в итоге воплотится в жизнь, мы не только сумеем достичь достойного научно-технического опыта, чтобы нейтрализовать взрывное развитие искусственного интеллекта, но и поднимемся на высочайший уровень профессионализма, которое совсем не помешает, если человечество хочет пережить пришествие сверхразумных машин в свой мир.

Но прежде чем обратить свой взор в будущее, было бы полезно коротко напомнить историю создания машинного интеллекта.

Путь надежды и отчаяния

Летом 1956 года в Дартмутском колледже собрались на двухмесячный семинар десять ученых, объединенных общим интересом к нейронным сетям, теории автоматов и исследованию интеллекта. Время проведения Дартмутского семинара обычно считают точкой отсчета новой области науки — изучения искусственного интеллекта. Большинство его участников позднее будут признаны основоположниками этого направления. Насколько оптимистично ученые глядели в будущее, говорит текст их обращения в Фонд Рокфеллера, собиравшийся финансировать мероприятие:

> Нами предполагается провести семинар по исследованию искусственного интеллекта, который продлится два месяца и в котором примут участие десять ученых… Изучение вопроса будет опираться на предположение, что на сегодняшний день существует принципиальная возможность моделирования интеллекта, поскольку теперь мы в состоянии точно описать каждый аспект обучения машины и любые отличительные признаки умственной деятельности. Будет предпринята попытка определить пути, как разработать машину, способную использовать язык, формировать абстракции и концепции, решать задачи, сейчас доступные лишь человеческому уму, и саморазвиваться. Считаем, что добьемся существенного прогресса в решении отдельных указанных проблем, если тщательно отобранная группа специалистов получит возможность трудиться сообща в течение лета.

После эпохального события, отмеченного столь энергичным прологом, прошло шестьдесят лет, за которые исследования в области искусственного интеллекта преодолели нелегкий путь: от громогласного ажиотажа до падения интереса, от завышенных ожиданий к обманутым надеждам.

Первый период всеобщего воодушевления начался с Дартмутского семинара. Позднее его главный организатор Джон Маккарти описал это время как эпоху вполне успешного освоения в духе детского «смотри, мам, без рук могу!». В те далекие годы ученые выстраивали системы, целью которых было опровергнуть довольно часто звучавшие утверждения скептиков, будто машины «ни на что не способны». Чтобы парировать удар, исследователи искусственного интеллекта разрабатывали небольшие программы, которые выполняли действие *X* в условном микромире (четко определенной ограниченной области, предназначенной для демонстрации упрощенной версии требуемого поведения), тем самым доказывая правильность концепции и показывая принципиальную возможность выполнения действия *X* машинами. Одна из таких ранних систем, названная «Логик-теоретик» (Logical Theorist), смогла доказать большую часть теорем из второго тома «Оснований математики» (Principia Mathematica) Альфреда Уайтхеда и Бертрана Рассела*; причем одно из доказательств оказалось изящнее оригинального. Тем самым ученые, продемонстрировав способность машины к дедукции и созданию логических построений, сумели развеять миф, будто она «мыслит лишь цифрами»[13]. За «Логик-теоретик» последовала программа «Универсальный решатель задач» (General Problem Solver, GPS), предназначенная решать, в принципе, любую формально определенную задачу[14]. Были созданы системы, которые справлялись с такими проблемами, как: математические задачи университетских курсов первого года обучения; визуальные головоломки по выявлению геометрических аналогий, применяемые при проверке показателя интеллекта; простые вербальные задачи по алгебре[15]. Робот «Трясучка» (Shakey) — названный так из-за вибрации во время работы — показал, что машина может продумывать и контролировать свою двигательную активность, когда логическое мышление совмещено с восприятием окружающей действительности[16]. Программа ELIZA прекрасно имитировала поведение психотерапевта[17]. В середине 1970-х годов

* *Альфред Уайтхед, Бертран Рассел*. Основания математики. В 3 т. / Под ред. Г. П. Ярового, Ю. Н. Радаева. Самара: Самарский университет, 2005–2006.

программа SHRDLU продемонстрировала, как смоделированный робот в смоделированном мире спокойно манипулирует объемными геометрическими фигурами, не только выполняя инструкции пользователя, но и отвечая на его вопросы[18]. В последующие десятилетия были созданы программы, способные сочинять классическую музыку разных жанров, решать проблемы клинической диагностики быстрее и увереннее врачей-стажеров, самостоятельно управлять автомобилями и делать патентоспособные изобретения[19]. Появилась даже интеллектуальная система, выдававшая оригинальные шутки[20] (не сказать, чтобы уровень был высок, но дети, как говорят, находили их забавными).

Однако методы, хорошо зарекомендовавшие себя при разработке тех первых, практически демонстрационных, образцов интеллектуальных систем, не удавалось применить в тех случаях, когда речь заходила о широком спектре проблем и более трудных задачах. Одна из причин заключалась в комбинаторном взрыве, то есть скачкообразном росте количества возможных вариантов, которые приходилось изучать с помощью средств, основанных на простейшем методе перебора. Этот метод хорошо себя проявил на примере несложных задач, но не подходил для чуть более трудных. Например, для решения теоремы с доказательством длиной в пять строк системе логического вывода с одним правилом и пятью аксиомами требовалось просто пронумеровать все 3125 возможных комбинаций и проверить, какая из них приведет к нужному заключению. Исчерпывающий поиск также работал для доказательств длиной в шесть или семь строк. Но поиск методом полного перебора возможных вариантов начинал пробуксовывать, когда проблема усложнялась. Время для решения теоремы с доказательством не в пять, а пятьдесят строк будет отнюдь не в десять раз больше: если использовать полный перебор, то потребуется проверить $5^{50} \approx 8{,}9 \times 10^{34}$ возможных последовательностей — вычислительно немыслимая задача даже для самого сверхмощного компьютера.

Чтобы справиться с комбинаторным взрывом, нужны алгоритмы, способные анализировать структуру целевой области и использовать преимущества накопленного знания за счет эвристического поиска, долгосрочного планирования и свободных абстрактных представлений, — однако в первых интеллектуальных системах все перечисленные возможности были разработаны довольно плохо. Кроме того, из-за ряда обстоятельств — неудовлетворительные методы обработки неопределенности, использование нечетких и произвольных символических записей, скудость данных, серьезные

технические ограничения по объему памяти и скорости процессора — страдала общая производительность этих систем. Осознание проблем пришло к середине 1970-х годов. Осмысление того, что многие проекты никогда не оправдают возложенных на них ожиданий, обусловило приход первой «зимы искусственного интеллекта»: наступил период регресса, в течение которого сократилось финансирование и вырос скептицизм, а сама идея искусственного интеллекта перестала быть модной.

Весна вернулась в начале 1980-х годов, когда в Японии решили приступить к созданию компьютера пятого поколения. Страна собиралась совершить мощный бросок в будущее и сразу выйти на сверхсовременный уровень технологического развития, разработав архитектуру параллельных вычислительных систем для сверхмощных компьютеров с функциями искусственного интеллекта. Это была хорошо финансируемая правительственная программа с привлечением крупных частных компаний. Появление проекта совпало со временем, когда японское послевоенное чудо приковывало к себе внимание всего западного мира: политические и деловые круги с восхищением и тревогой следили за успехами Японии, стремясь разгадать секретную формулу ее экономического взлета и надеясь воспроизвести ее у себя дома. Как только Япония решила инвестировать огромные средства в изучение искусственного интеллекта, ее примеру последовали многие высокоразвитые страны.

В последующие годы широкое распространение получили *экспертные системы*, призванные заменить специалистов-экспертов при разрешении проблемных ситуаций. Они представляли собой автоматизированные компьютерные системы, программы которых базировались на наборе правил, позволяющих распознавать ситуации и делать простые логические умозаключения, выводя их из баз знаний, составленных специалистами в соответствующих предметных областях и переведенных на формальный машинный язык. Были разработаны сотни таких экспертных систем. Однако выяснилось, что от небольших систем толку мало, а более мощные оказались слишком громоздкими в применении и дорогостоящими в разработке, апробации и постоянном обновлении. Специалисты пришли к выводу, что непрактично использовать отдельный компьютер для выполнения всего одной программы. Таким образом, уже к концу 1980-х годов этот период подъема тоже выдохся.

Японский проект, связанный с появлением компьютера пятого поколения, в принципе, провалился, как и аналогичные разработки в США и Европе. Наступила вторая зима искусственного интеллекта. Теперь маститый критик

мог вполне обоснованно посетовать, что, мол, «вся история исследований искусственного интеллекта вплоть до сегодняшнего дня складывается из череды отдельных эпизодов, когда, как правило, очень умеренная удача на исключительно узком участке работы довольно скоро оборачивается полной несостоятельностью на более широком поле, к исследованию которого, казалось бы, поощрял первоначальный успех»[21]. Частные инвесторы старались держаться на почтительном расстоянии от любых начинаний, имевших малейшее отношение к проблеме искусственного интеллекта. Даже в среде ученых и финансировавших их организаций сам этот термин стал нежелательным[22].

Однако технический прогресс не стоял на месте, и к 1990-м годам вторая зима искусственного интеллекта сменилась оттепелью. Всплеску оптимизма способствовало появление новых методов, которые, казалось, придут на смену привычному логическому программированию — обычно его именуют или «старый добрый искусственный интеллект, или «классический искусственный интеллект» (КИИ). Эта традиционная парадигма программирования была основана на высокоуровневой манипуляции символами и достигла своего расцвета в 1980-е годы, в период увлечения экспертными системами. Набиравшие популярность интеллектуальные методы, например, такие как нейронные сети и генетические алгоритмы, подавали надежду, что все-таки удастся преодолеть присущие КИИ недостатки, в частности, его «уязвимость» (машина обычно выдавала полную бессмыслицу, если программист делал хотя бы одно ошибочное предположение). Новые методы отличались лучшей производительностью, поскольку больше опирались на естественный интеллект. Например, нейронные сети обладали таким замечательным свойством, как отказоустойчивость: небольшое нарушение приводило лишь к незначительному снижению работоспособности, а не полной аварии. Еще важнее, что нейронные сети представляли собой самообучающиеся интеллектуальные системы, то есть накапливали опыт, умели делать выводы из обобщенных примеров и находить скрытые статистические образы во вводимых данных[23]. Это делало сети хорошим инструментом для решения задач классификации и распознавания образов. Например, создав определенный набор сигнальных данных, можно было обучить нейронную сеть воспринимать и распознавать акустические особенности подводных лодок, мин и морских обитателей с большей точностью, чем это могли делать специалисты, — причем система справлялась без

всяких предварительных выяснений, какие нужно задать параметры, чтобы учитывать и сопоставлять те или иные характеристики.

Хотя простые модели нейронных сетей были известны с конца 1950-х годов, ренессанс в этой области начался после создания метода обратного распространения ошибки, который позволил обучать многослойные нейронные сети[24]. Такие многослойные сети, в которых имелся как минимум один промежуточный («скрытый») слой нейронов между слоями ввода и вывода, могут обучиться выполнению гораздо большего количества функций по сравнению с их более простыми предшественниками[25]. В сочетании с последним поколением компьютеров, ставших к тому времени намного мощнее и доступнее, эти усовершенствования алгоритма обучения позволили инженерам строить нейронные сети, достаточно успешно решающие практические задачи во многих областях применения.

По своим свойствам и функциональному сходству с биологическим мозгом нейронные сети выгодно отличались от жестко заданной логики и уязвимости традиционных, основанных на определенных правилах систем КИИ. Контраст оказался настолько сильным, что даже возникла очередная концепция коннективистской модели; сам термин *коннективизм** особенно подчеркивал важность массово-параллельной обработки субсимвольной информации. С тех пор об искусственных нейронных сетях написано более ста пятидесяти тысяч научных работ, а сами сети продолжают оставаться важным методом в области машинного обучения.

Еще одним фактором, приблизившим приход очередной весны искусственного интеллекта, стали генетический алгоритм и генетическое программирование. Эти разновидности методов эволюционных вычислений получили довольно широкую известность, хотя, возможно, с научной точки зрения не приобрели столь большого значения, как нейронные сети. В эволюционных моделях в первую очередь создаются начальные популяции тех или иных решений (могут быть либо структуры данных, либо программы обработки данных), затем — в результате случайной мутации и размножения («скрещивания») имеющихся популяций — генерируются новые популяции.

---

* *Коннективизм*, или коннекционизм (connectionism), — моделирует в сетях мыслительные и поведенческие явления из взаимосвязанных простых элементов; на самом деле понятие коннективизма возникло намного раньше самих искусственных нейронных систем; как подход он применяется не только в области искусственного интеллекта, но и в философии сознания, психологии, когнитивистике.

Периодически вследствие применения критерия отбора (по наличию целевой функции, или функции пригодности) количество популяций сокращается, что позволяет войти в новое поколение лишь лучшим решениям-кандидатам. В ходе тысяч итераций среднее качество решений в популяции постепенно повышается. С помощью подобных алгоритмов генерируются самые продуктивные программы, способные ориентироваться в весьма широком круге вопросов; причем отобранные решения иногда на удивление получаются новаторскими и неожиданными, чаще напоминающими естественную систему, нежели смоделированную человеком структуру. Весь процесс может происходить, по сути, без участия человека, за исключением случаев, когда необходимо назначить целевую функцию, которая, в принципе, определяется очень просто. Однако на практике, чтобы эволюционные методы работали хорошо, требуются и профессиональные знания, и талант, особенно при создании понятного формата представления данных. Без эффективного метода кодирования решений-кандидатов (генетического языка, адекватного латентной структуре целевой области) эволюционный процесс, как правило, или бесконечно блуждает в открытом поисковом пространстве, или застревает в локальном оптимуме. Но даже когда найден правильный формат представления, эволюционные вычисления требуют огромных вычислительных мощностей и часто становятся жертвой комбинаторного взрыва.

Такие примеры новых методов, как нейронные сети и генетические алгоритмы, сумели стать альтернативой закосневшей парадигме КИИ и потому вызвали в 1990-е годы новую волну интереса к интеллектуальным системам. Но у меня нет намерений ни воздавать им хвалу, ни возносить на пьедестал в ущерб другим методам машинного обучения. По существу, одним из главных теоретических достижений последних двадцати лет стало ясное понимание, что внешне несходные методы могут считаться особыми случаями в рамках общей математической модели. Скажем, многие типы искусственных нейронных сетей могут рассматриваться как классификаторы, выполняющие определенные статистические вычисления (оценка по максимуму правдоподобия)[26]. Такая точка зрения позволяет сравнивать нейронные сети с более широким классом алгоритмов для обучения классификаторов по примерам — деревья принятия решений, модели логистической регрессии, методы опорных векторов, наивные байесовские классификаторы, методы ближайшего соседа[27]. Точно так же можно считать, что генетические алгоритмы выполняют локальный стохастический поиск с восхождением

к вершине, который, в свою очередь, является подмножеством более широкого класса алгоритмов оптимизации. Каждый из этих алгоритмов построения классификаторов или поиска в пространстве решений имеет свой собственный набор сильных и слабых сторон, которые могут быть изучены математически. Алгоритмы различаются требованиями ко времени вычислений и объему памяти, предполагаемыми областями применения, легкостью, с которой в них может быть включен созданный вовне контент, а также тем, насколько прозрачен для специалистов механизм их работы.

За суматохой машинного обучения и творческого решения задач скрывается набор хорошо понятных математических компромиссов. Вершиной является идеальный байесовский наблюдатель, то есть тот, кто использует доступную ему информацию оптимальным с вероятностной точки зрения способом. Однако эта вершина недостижима, поскольку требует слишком больших вычислительных ресурсов при реализации на реальном компьютере (см. врезку 1). Таким образом, можно смотреть на искусственный интеллект как на поиск коротких путей, то есть как на способ приблизиться к байесовскому идеалу на приемлемое расстояние, пожертвовав некоторой оптимальностью или универсальностью, но при этом сохранив довольно высокий уровень производительности в интересующей исследователя области.

Отражение этой картины можно увидеть в работах, выполненных в последние двадцать лет на графовых вероятностных моделях, таких как байесовские сети. Байесовские сети являются способом сжатого представления вероятностных и условно независимых отношений, характерных для определенной области. (Использование таких независимых отношений критически важно для решения проблемы комбинаторного взрыва, столь же важной в случае вероятностного вывода, как и при логической дедукции.) Кроме того, они стали значимым инструментом для понимания концепции причинности[28].

ВРЕЗКА 1. ОПТИМАЛЬНЫЙ БАЙЕСОВСКИЙ АГЕНТ

Идеальный байесовский агент начинается с задания «априорного распределения вероятности», то есть функции, приписывающей определенную вероятность всем «возможным мирам» — иначе говоря, результатам всех сценариев, по которым может меняться мир[29]. Априорное распределение вероятности включает в себя индуктивное смещение, то есть более простым возможным мирам присваивается более высокая вероятность. (Один из способов формально определить простоту возможного мира — использовать показатель колмогоровской сложности, основанный на длине максимально короткой компьютерной программы, генерирующей полное описание этого

мира[30].) При этом в априорном распределении вероятности учитываются любые знания, которые программисты желают передать агенту.

После того как агент получает со своих сенсоров новую информацию, он меняет распределение вероятности, «обусловливая» распределение с учетом этой новой информации в соответствии с теоремой Байеса[31]. Обусловливание — это математическая операция, которая заключается в присвоении нулевых значений вероятности тем мирам, которые не согласуются с полученной информацией, и нормализации распределения вероятности оставшихся возможных миров. Результатом становится «апостериорное распределение вероятности» (которое агент может использовать в качестве априорного на следующем шаге). По мере того как агент проводит свои наблюдения, распределение вероятности концентрируется на все сильнее сжимающемся наборе возможных миров, которые согласуются с полученными свидетельствами; и среди этих возможных миров наибольшую вероятность всегда имеют самые простые.

Образно говоря, вероятность похожа на песок, рассыпанный на большом листе бумаги. Лист разделен на области различного размера, каждая из которых соответствует одному из возможных миров, причем области большей площади эквивалентны более простым мирам. Представьте также слой песка или любого порошка, покрывающего бумагу, — это и есть наше априорное распределение вероятности. Когда проводится наблюдение, в результате которого исключаются какие-то из возможных миров, мы убираем песок из соответствующих областей и распределяем его равномерно по областям, «остающимся в игре». Таким образом, общее количество песка на листе остается неизменным, просто по мере накопления наблюдений он концентрируется во все меньшем количестве областей. Здесь представлено описание обучения в его самом чистом виде. (Чтобы рассчитать вероятность *гипотезы*, мы просто измеряем количество песка во всех областях, соответствующих возможным мирам, в которых эта гипотеза истинна.)

Итак, мы определили правило обучения. Чтобы получить агента, нам потребуется также правило принятия решений. Для этого мы наделяем агента «функцией полезности», которая присваивает каждому возможному миру определенное число. Это число представляет собой желательность соответствующего мира с точки зрения базовых предпочтений агента[32]. (Чтобы выявить действие с максимальной ожидаемой полезностью, агент мог бы составить список всех возможных действий. А затем рассчитать условное распределение вероятности с учетом каждого действия — то есть распределение вероятности, которое стало бы следствием обусловливания текущего распределения вероятности после наблюдения за результатами этого действия. И наконец, рассчитать ожидаемую ценность действия можно как сумму ценностей всех возможных миров, умноженных на условную вероятность этих миров с учетом осуществления действия[33].)

Правило обучения и правило принятия решений задают «определение оптимальности» агента. (В сущности такое же определение оптимальности широко используется в искусственном интеллекте, эпистемологии, философии науки, экономике и статистике[34].) В реальном мире такого агента получить невозможно, поскольку для проведения необходимых расчетов не хватит никаких вычислительных мощностей. Любая попытка сделать это приводит к комбинаторному взрыву вроде описанного нами при обсуждении

КИИ. Чтобы представить это, рассмотрим крошечное подмножество всех возможных миров, состоящее из единственного компьютерного монитора, висящего в бесконечном пустом пространстве. Разрешение монитора — 1000 × 1000 пикселей, каждый из которых постоянно или светится, или нет. Даже такое подмножество всех возможных миров невероятно велико: количество возможных состояний монитора, равное $2^{(1000 \times 1000)}$, превосходит объем всех вычислений, которые когда-либо будут выполнены в обозримой Вселенной. То есть мы не можем даже просто пронумеровать возможные миры в этом небольшом подмножестве всех возможных миров, не говоря уже о том, чтобы провести какие-то более сложные расчеты по каждому из них.

Но определение оптимальности может иметь теоретический интерес, даже несмотря на невозможность его физической реализации. Он представляет собой стандарт, с которым можно соотносить эвристические аппроксимации и который иногда позволяет нам судить, как именно поступил бы оптимальный агент в той или иной ситуации. С некоторыми альтернативными определениями оптимальности мы еще встретимся в двенадцатой главе.

Одно из преимуществ связи задачи обучения в определенных областях с общей задачей байесовского вывода состоит в том, что эти новые алгоритмы, делающие байесовский вывод более эффективным, немедленно приводят к прогрессу во множестве различных областей. Например, метод Монте-Карло непосредственно применяется в машинном зрении, робототехнике и вычислительной генетике. Еще одно преимущество заключается в том, что исследователям, работающим в различных областях, стало проще объединять результаты своих изысканий. Графовые модели и байесовские статистики представляют собой общий фокус исследований в таких областях, как машинное обучение, статистическая физика, биоинформатика, комбинаторная оптимизация и теория коммуникации[35]. Заметный прогресс в машинном обучении стал следствием использования формальных результатов, изначально полученных в других областях науки. (Конечно, машинное обучение значительно выиграло от появления более быстрых компьютеров и доступности больших наборов данных.)

Последние достижения

Во многих областях деятельности уровень искусственного интеллекта уже превосходит уровень человеческого. Появились системы, способные не только вести логические игры, но и одерживать победы над людьми. Приведенная в табл. 1 информация об отдельных игровых программах демонстрирует, как разнообразные виды ИИ побеждают чемпионов многих турниров[36].

Таблица 1. Игровые программы с искусственным интеллектом

Шашки	Уровень интеллекта выше человеческого	Компьютерная игра в шашки, написанная в 1952 году Артуром Самуэлем и усовершенствованная им в 1955 году (версия включала модуль машинного обучения), стала первой интеллектуальной программой, которая в будущем научится играть лучше своего создателя[37]. Программа «Чинук» (CHINOOK), созданная в 1989 году группой Джонатана Шеффера, сумела в 1994 году обыграть действующего чемпиона мира — первый случай, когда машина стала победителем в официальном чемпионате мира. Те же разработчики, использовав алгоритм поиска «альфа-бета отсечение» в базе данных для 39 трлн эндшпилей, представили в 2002 году оптимальную версию игры в шашки — это программа, всегда выбирающая лучший из ходов. Правильные ходы обеих сторон приводят к ничьей[38]
Нарды	Уровень интеллекта выше человеческого	Компьютерная игра в нарды, созданная в 1970 году Хансом Берлинером и названная им BKG, в 1979 году стала первой интеллектуальной программой, обыгравшей чемпиона мира в показательном матче — хотя впоследствии сам Берлинер приписывал эту победу удачно брошенным костям[39]. Созданная в 1991 году Джералдом Тезауро программа TD-Gammon уже в 1992 году достигла такого уровня мастерства, что могла сразиться на чемпионате мира. Ради самосовершенствования программа постоянно играла сама с собой, причем Тезауро использовал такую форму укрепляющего обучения, как метод временных различий[40]. С тех пор программы для игры в нарды по своему уровню в значительной степени превосходили лучших игроков мира[41]
«Эвриско» в космической битве Traveller TCS	Уровень интеллекта выше человеческого в сотрудничестве с самим человеком[42]	Дугласом Ленатом в 1976 году была создана программа «Эвриско» (Eurisco), представлявшая собой набор эвристических, то есть логических, правил («если — то»). В течение двух лет (1981, 1982) эта экспертная система выигрывала чемпионат США по фантастической игре Traveller TCS

		(межгалактическое сражение); организаторы даже меняли правила игры, но ничто не могло остановить победного шествия «Эвриско», в результате они приняли решение больше не допускать «Эвриско» к участию в чемпионате[43]. Для построения своего космического флота и сражения с кораблями противника «Эвриско» использовала эвристические правила, которые — в процессе самообучения — корректировала и улучшала при помощи других эвристических правил
Реверси («Отелло»)	Уровень интеллекта выше человеческого	Программа для игры в реверси Logistello выиграла в 1997 году подряд шесть партий у чемпиона мира Такэси Мураками[44]
Шахматы	Уровень интеллекта выше человеческого	Шахматный суперкомпьютер Deep Blue в 1997 году выиграл у чемпиона мира Гарри Каспарова, Каспаров, хотя и имел претензии к создателям машины, все-таки заметил в ее игре проблески истинного разума и творческого подхода[45]. С тех пор игровые шахматные программы продолжают совершенствоваться[46]
Кроссворды	Профессиональный уровень	Программа Proverb в 1999 году стала лучшей среди программ для решения кроссвордов среднего уровня[47]. Созданная в 2012 году Мэттом Гинзбергом программа Dr. Fill вошла в группу лучших участников чемпионата США по кроссвордам. (Показатели программы не были стабильными. Dr. Fill идеально справилась с кроссвордами, считавшимися наиболее сложными среди участников-людей, но оказалась бессильна перед нестандартными, в которых встречались слова, написанные задом наперед, и вопросы, расположенные по диагонали[48].)
«Скрабл» («Эрудит»)	Уровень интеллекта выше человеческого	По состоянию на 2002 год программы для игры в слова превосходят лучших игроков среди людей[49]
Бридж	Уровень интеллекта не уступает уровню лучших игроков	Программы для игры в бридж «Контракт» к 2005 году достигли уровня профессионализма лучших игроков среди людей[50]

Суперкомпьютер IBM Watson в телепередаче Jeopardy!	Уровень интеллекта выше человеческого	IBM Watson, созданный в IBM суперкомпьютер с системой ИИ, в 2010 году обыграл Кена Дженнингса и Брэда Раттера — двух рекордсменов Jeopardy![51]. Jeopardy! — телевизионная игра-викторина с простыми вопросами из области истории, литературы, спорта, географии, массовой культуры, науки и проч. Вопросы задаются в виде подсказок, при этом часто используется игра слов
Покер	Уровень разный	Игровые программы для покера на сегодняшний день несколько уступают лучшим игрокам в техасский холдем (популярная разновидность покера), но превосходят людей в некоторых других разновидностях игры[52]
Пасьянс «Свободная ячейка» («Солитер»)	Уровень интеллекта выше человеческого	Развитие эвристических алгоритмов привело к созданию программы для пасьянса «Свободная ячейка» (Free Cell), которая оказалась сильнее игроков самого высокого уровня[53]. В своей обобщенной форме эта игровая программа является NP-полной задачей.
Го	Уровень сильного игрока-любителя	По состоянию на 2012 год серия программ для игры в го «Дзен» (Zen) — использовав дерево поиска методом Монте-Карло и технологии машинного обучения — получила шестой дан (разряд) в быстрых играх[54]. Это уровень весьма сильного любителя. В последние годы игровые программы го совершенствуются со скоростью примерно один дан в год. Если этот темп развития сохранится, то, скорее всего, через десять лет они превзойдут чемпиона мира среди людей

Вряд ли сегодня данные факты смогут произвести хоть какое-то впечатление. Но это обусловлено тем, что наши представления о стандартах несколько смещены, поскольку мы уже знакомы с теми выдающимися достижениями, которые появились после описываемых событий. В прежние времена, например, профессиональное умение шахматиста считалось высшим проявлением умственной деятельности человека. Некоторые специалисты конца 1950-х годов считали: «Если когда-нибудь получится создать удачную машину для игры в шахматы, возможно, люди постигнут суть своих

интеллектуальных усилий»[55]. В наше время все выглядит иначе. Остается лишь согласиться с Джоном Маккарти, когда-то посетовавшим, что «стоит системе нормально начать работать, как ее сразу перестают называть искусственным интеллектом»[56].

Однако появление интеллектуальных шахматных систем не обернулось тем торжеством разума, на которое многие рассчитывали, — и это имело определенное объяснение. По мнению ученых того времени — мнению, наверное, небезосновательному, — компьютер станет играть в шахматы наравне с гроссмейстерами, только когда будет наделен высоким *общим уровнем* интеллектуального развития[57]. Казалось бы, великий шахматист должен соответствовать немалым требованиям: иметь крепкую теоретическую подготовку; быть способным оперировать абстрактными понятиями; стратегически мыслить и разумно действовать; заранее выстраивать хитроумные комбинации; обладать дедуктивным мышлением и даже уметь моделировать ход мысли противника. Отнюдь. Выяснилось, что достаточно разработать идеальную шахматную программу на основе алгоритма с узкоцелевым назначением[58]. Если программу поставить на быстродействующий процессор — а скоростные компьютеры стали доступны уже в конце XX века, — то она демонстрирует весьма сильную игру. Однако подобный искусственный интеллект слишком однобок. Он ничего другого не умеет, кроме как играть в шахматы[59].

В других случаях изучения и применения искусственного интеллекта выявились проблемы более *сложного порядка*, чем ожидалось, поэтому и развитие шло значительно медленнее. Профессор Дональд Кнут, крупнейший специалист в области программирования и вычислительной математики, с удивлением заметил: «Искусственный интеллект, преуспев сегодня во всем, где требуется „разум“, неспособен на те действия, которые люди и животные совершают „бездумно“, — эта задача оказалась гораздо труднее!»[60] Затруднения вызывала, например, разработка системы управления поведением роботов, а также такие их функции, как распознавание зрительных образов и анализ объектов при взаимодействии с окружающей средой. Тем не менее и сделано было немало, и продолжает поныне делаться, причем работа идет не только над развитием программного обеспечения — постоянно совершенствуются аппаратные средства.

В один ряд с исследованием инстинктивного поведения можно поставить логику здравого смысла и понимание естественных языков — явления,

которые тоже оказались не самыми легкими для систем искусственного интеллекта. Сейчас принято считать, что решение подобных проблем на уровне, сопоставимом с человеческим, является AI-полной задачей* — то есть их сложность эквивалентна трудности разработки машин, таких же умных и развитых, как люди[61]. Иными словами, если кто-то *добьется успеха* в создании ИИ, способного понимать естественный язык так же, как понимает его взрослый человек, то, скорее всего, он или уже создал ИИ, который может делать все, на что способен человеческий разум, или будет находиться в шаге от его создания[62].

Высокий уровень игры в шахматы, как оказалось, достижим с помощью исключительно простого алгоритма. Возникает соблазн считать, будто и другие способности, например общее умение осмысливать или некоторые основные навыки программирования, можно также обеспечить за счет некоего удивительно несложного алгоритма. То обстоятельство, что в определенный момент оптимальная продуктивность достигается в результате применения сложного механизма, вовсе не означает, что ни один простой механизм не способен делать ту же работу так же хорошо и даже лучше. Птолемеева система мира (в центре Вселенной находится неподвижная Земля, а вокруг нее вращаются Солнце, Луна, планеты и звезды) выражала представление науки об устройстве мироздания на протяжении тысячи лет. Чтобы лучше объяснять характер движения небесных тел, ученые от века к веку усложняли модель системы, добавляя все новые и новые эпициклы, за счет чего повышалась точность ее прогнозов. Пришло время, и на смену геоцентрической пришла гелиоцентрическая система мира; теория Коперника была намного проще, а после доработки ее Кеплером стала и прогностически более точной[63].

В современном мире методы искусственного интеллекта используют столь широко, что вряд ли целесообразно рассматривать здесь все области их применения, но некоторые стоит упомянуть, чтобы дать общее представление о масштабе распространения самой идеи. Помимо представленных в табл. 1 логических игровых программ, сегодня разрабатывают: слуховые аппараты на базе алгоритмов, отфильтровывающих фоновый шум;

---

* *AI-полная задача* (где AI — artificial intelligence («искусственный интеллект»)) — неформальный термин, который применяется в теории ИИ по аналогии с NP-полным классом задач. По существу означает задачу создания искусственного интеллекта человеческого уровня.

навигационные системы, отображающие карты и подсказывающие маршрут водителям; рекомендательные системы, предлагающие книги и музыкальные альбомы пользователям на основе анализа их предыдущих покупок и оценок; системы поддержки принятия медицинских решений, помогающие врачам, например, диагностировать рак молочной железы, подбирать варианты лечения и расшифровывать электрокардиограммы. В настоящее время, кроме промышленных роботов, которых уже больше миллиона, появились самые разные роботы-помощники: домашние питомцы; пылесосы; газонокосильщики; спасатели; хирурги[64]. Общая численность роботов в мире превысила десять миллионов[65].

Современные системы распознавания речи, основанные на статистических методах вроде скрытых марковских моделей, являются довольно точными для практического использования (с их помощью были созданы некоторые начальные фрагменты этой книги). Персональные цифровые помощники (например, Siri — приложение Apple) реагируют на голосовые команды, могут отвечать на простые вопросы и выполнять распоряжения. Повсеместно распространено оптическое распознавание рукописного и машинописного текста — на нем основаны, в частности, приложения для сортировки почты и оцифровки исторических документов[66].

До сих пор остаются несовершенными системы машинного перевода, тем не менее для определенных целей они вполне пригодны. На стадии ранних версий, в которых использовался метод КИИ и которые основывались на правилах, был создан принцип кодировки в ручном режиме для грамматик всех естественных языков — причем работа проводилась силами самых высококвалифицированных лингвистов. Новые системы основаны на статистических методах машинного обучения, которые автоматически выстраивают статистические модели на основе наблюдаемых ими закономерностей использования слов и фраз. Программы выводят параметры этих моделей, анализируя корпус текстов на двух языках. Такой подход позволяет не привлекать лингвистов, а программисты, разрабатывающие эти системы, могут даже не владеть языками, с которыми им приходится иметь дело[67].

Системы распознавания лиц за последнее время были настолько усовершенствованы, что сейчас ими успешно пользуются пограничные службы в Европе и Австралии. Автоматическая идентификационная система работает в Госдепартаменте США, с ее помощью в процессе выдачи виз обрабатывается более семидесяти пяти миллионов фотографий в год. В системах

наблюдения применяются все более совершенные методы ИИ и новейшие технологии по извлечению информации, с помощью которых проводят интеллектуальный анализ речевых, текстовых и видеоматериалов — основная часть их привлекается из общемировых коммуникационных сетей и гигантских центров сбора и обработки данных.

Автоматическое доказательство теорем и решение уравнений стало настолько общим местом, что уже не воспринимается как разработка искусственного интеллекта. Устройства для решения уравнений встроены в научные компьютерные программы, например систему Mathematica. Формальные методы проверки, в том числе системы автоматического доказательства теорем, повсеместно используются производителями микропроцессоров для проверки поведения схемы перед запуском в производство.

Американскими военными и разведывательными ведомствами широко и успешно внедряются так называемые боевые роботы — саперы для нахождения и обезвреживания бомб и мин; беспилотные летательные аппараты, предназначенные как для разведки, так и для боевых действий; другие автоматические виды вооружений. Сегодня эти устройства в основном управляются дистанционно операторами-специалистами, однако неустанно ведется работа над расширением их автономной деятельности.

Большой успех достигнут в области интеллектуального планирования и снабжения. В ходе операции «Буря в пустыне» в 1991 году была развернута система DART для обеспечения автоматизированного планирования поставок и составления графиков перевозок. Программа оказалась исключительно эффективной: по сводкам Агентства по перспективным оборонным научно-исследовательским разработкам США (Defense Advanced Research Projects Agency in the United States, DARPA), она одна окупила тридцатилетнее финансирование Министерством обороны работ в области ИИ[68]. Сложные программы календарного планирования и тарификации используются для систем бронирования авиабилетов. Компании активно применяют самые разные методы ИИ для контроля складских запасов. Автоматические системы телефонного бронирования и линии поддержки, соединенные с программами распознавания речи, способны провести несчастного потребителя через лабиринт взаимосвязанных вариантов выбора.

Технологии искусственного интеллекта лежат в основе многих интернет-сервисов. Общемировой трафик электронной почты проверяется специальным программным обеспечением — причем байесовская фильтрация спама,

несмотря на постоянные усилия спамеров приспособиться и обойти защиту, в основном справляется с задачей и держит оборону. Электронные программы, используя компоненты ИИ, обеспечивают безопасность операций по банковским картам: отвечают за их автоматическое одобрение или отклонение и постоянно отслеживают действия по счету с целью обнаружить малейшие признаки мошенничества. Системы поиска информации также активно используют машинное обучение. А поисковая система Google, без сомнения, представляет собой величайшую из когда-либо созданных систем искусственного интеллекта.

Здесь стоит подчеркнуть, что граница между искусственным интеллектом и обычным программным обеспечением определена не очень четко. Некоторые из перечисленных выше программ могли бы скорее считаться приложениями многофункциональных программных обеспечений, нежели интеллектуальными системами, — тут невольно снова вспомнишь слова Маккарти, что «стоит системе нормально начать работать, как ее сразу перестают называть искусственным интеллектом». Для наших целей важнее обратить внимание на другое различие: есть системы, у которых имеется ограниченный набор когнитивных способностей (неважно, относятся они к ИИ или нет), и есть системы, обладающие широкоприменимыми инструментами для решения общих задач. В основном все используемые сейчас системы относятся к первому типу — узкодиапазонному. Однако многие из них содержат компоненты, способные либо сыграть роль в создании будущего искусственного интеллекта, который будет отличаться развитым общим уровнем развития, либо стать его частью, — это такие компоненты, как классификаторы, алгоритмы поиска, модули планирования, решатели задач и схемы представлений.

Системы искусственного интеллекта качественно работают еще в одной области, где ставки очень высоки, а конкуренция слишком жестока, — это мировой финансовый рынок. Автоматизированные системы торговли акциями широко используются крупными инвестиционными банками. И хотя некоторые из них всего лишь дают возможность автоматизировать исполнение заказов на покупку и продажу, выставленных управляющей компанией, другие реализуют сложные торговые стратегии, способные приспосабливаться к меняющимся условиям рынка. Чтобы изучать закономерности и тенденции фондового рынка, определять зависимость динамики котировок от внешних переменных, таких как, например, ключевые позиции в сводках финансовых новостей, — для всего этого в аналитических системах

используется большой набор методик интеллектуального анализа данных и временных последовательностей. Новые потоковые котировки, выпускаемые агентствами финансовой информации, специально отформатированы под интеллектуальные автоматизированные системы. Другие системы специализируются на поиске возможностей совершать арбитражные операции либо на определенном рынке ценных бумаг, либо одновременно на нескольких рынках, либо с помощью алгоритмического высокочастотного трейдинга*, целью которого является получение прибыли на незначительных колебаниях цен в пределах нескольких милисекунд (на таких временных интервалах начинают играть роль задержки в поступлении информации даже в оптоволоконных сетях, где она распространяется со скоростью света, и преимущество получают те, чьи компьютеры находятся в непосредственной близости от биржи). На долю алгоритмических высокочастотных трейдингов приходится более половины оборота фондового рынка США[69]. Существует мнение, что ответственность за так называемый мгновенный обвал фондовых индексов 6 мая 2010 года лежит именно на алгоритмической торговле (см. врезку 2).

ВРЕЗКА 2. «МГНОВЕННЫЙ ОБВАЛ» 2010 ГОДА

К полудню 6 мая 2010 года американский фондовый рынок уже упал на 4% на беспокойстве по поводу европейского долгового кризиса. Крупный игрок (группа взаимных фондов) инициировал в 14:32 алгоритм продажи для реализации большого количества фьючерсных контрактов E-Mini S&P 500 по цене, привязанной к показателю изменения ликвидности биржевых торгов. Эти контракты, приобретенные с помощью алгоритмических высокочастотных трейдингов, были запрограммированы быстро закрывать свои временные длинные позиции путем продажи контрактов другим игрокам. Поскольку спрос со стороны инвесторов, ориентирующихся на фундаментальные показатели, снизился, игроки алгоритмического трейдинга начали продавать фьючерсы E-Mini другим игрокам алгоритмического трейдинга, которые, в свою очередь, продавали их третьим таким же игрокам, создавая, таким образом, эффект «горячей картошки», которую пытаются «скинуть» как можно быстрее, — этот эффект раздувал объемы торгов, что было интерпретировано алгоритмом продажи как показатель высокой ликвидности. Поскольку игроки начали еще быстрее сбрасывать

---

* *Алгоритмический высокочастотный трейдинг*, или алгоритмическая высокочастотная торговля (algorithmic high-frequency trading), — формализованный процесс совершения торговых операций на финансовых рынках по заданному алгоритму с использованием специализированных компьютерных систем (торговых роботов).

друг другу E-Mini, на фондовом рынке возник настоящий порочный круг. В какой-то момент игроки начали просто выводить средства, еще больше повышая ликвидность на фоне продолжающегося падения цен. Сделки по E-Mini были приостановлены в 14:45 автоматическим прерывателем — специальной программой, контролирующей неожиданное и чрезмерное движение цен акций на бирже. Буквально через пять секунд торги возобновились, при этом цены стабилизировались и вскоре отыграли большую часть падения. Но в течение этих критических минут с рынка был «смыт» триллион долларов, поскольку значительное число сделок прошло по абсурдным ценам: акция могла продаваться и за один цент, и за 100 тысяч долларов. После того как торги закончились, состоялась встреча представителей бирж и регулирующих органов, на которой было принято решение отменить все сделки, исполненные по ценам, отличающимся от докризисного уровня на 60% и более. Договаривающиеся стороны сочли эти цены «явно ошибочными», а потому — в соответствии с существующими биржевыми правилами — подлежащими отмене задним числом)[70].

Изложенный сюжет представляет собой безусловное отступление от темы нашей книги, поскольку компьютерные программы, якобы ответственные за те минуты финансового кризиса, получившего название «мгновенный обвал», не были ни особенно интеллектуальными, ни слишком изощренными. Специфика созданной ими опасности принципиально отличается от характера угрозы, которую несет в себе появление искусственного сверхразума. Тем не менее из описанных событий можно вынести несколько полезных уроков.

Первое предупреждение. Взаимодействие нескольких простых компонентов (например, алгоритмы продаж и алгоритмическая высокочастотная торговля) может приводить к сложным и непредсказуемым последствиям. Если добавлять в налаженную систему новые элементы, возникают системные риски, не слишком очевидные до момента, когда что-то пойдет не так (да и то не всегда)[71].

Второе предупреждение. Несмотря на то что специалисты в области искусственного интеллекта обучают программу на основании предположений, кажущихся здравыми и логичными (например, объем торгов является верным показателем ликвидности рынка), это может приводить к катастрофическим результатам. В непредвиденных обстоятельствах, когда исходные допущения оказываются неверными, программа с железобетонной логической стойкостью продолжает поступать в соответствии с полученными инструкциями. Алгоритм «тупо» делает свою обычную работу, которую делал всегда, и его совсем не беспокоит — если он, конечно, не принадлежит к редчайшей разновидности алгоритмов, — что мы хватаемся за голову в ужасе от абсурдности его действий. К этой теме мы еще вернемся.

Третье предупреждение. Несомненно, автоматизация процесса внесла свой вклад в возникновение инцидента, однако, без всяких сомнений, она также способствовала и разрешению проблемы. Программа контроля, отвечавшая за приостановку торгов в случае слишком большого отклонения цен от нормального уровня, сработала автоматически, поскольку ее создатели справедливо предполагали, что события, которые

приводят к такому отклонению, могут происходить на временных интервалах, слишком коротких, чтобы на них успели отреагировать люди. Налицо потребность не полагаться во всем на контроль со стороны человека, а иметь в качестве подстраховки заранее разработанные и автоматически исполняемые алгоритмы безопасности. Кстати, это наблюдение предваряет тему, крайне важную в нашем последующем обсуждении машинного сверхразума[72].

Будущее искусственного интеллекта — мнение специалистов

Успех, достигнутый на двух магистральных направлениях: во-первых, создание более прочного статистического и информационно-теоретического основания для машинного обучения; во-вторых, практическая и коммерческая эффективность различных конкретных приложений, узкоспециальных с точки зрения решаемых проблем и областей применения, — привел к тому, что пошатнувшийся было престиж исследований искусственного интеллекта удалось несколько восстановить. Но, похоже, у научного сообщества, имеющего отношение к этой теме, от прошлых неудач остался довольно горький опыт, вынуждающий многих ведущих исследователей отказываться от собственных устремлений и больших задач. Поэтому один из основателей направления Нильс Нильсон укоряет своих нынешних коллег в отсутствии той творческой дерзости, которая отличала поколение первопроходцев:

> Соображение «благопристойности», на мой взгляд, оказывает дурное влияние на некоторых исследователей, выхолащивая саму идею искусственного интеллекта. Я будто слышу, как они говорят: «ИИ критиковали за отсутствие результатов. Теперь, добившись видимого успеха, мы не хотим рисковать собственной репутацией». Подобная осмотрительность приведет к тому, что все интересы ученых будут ограничены созданием программ, предназначенных предоставлять помощь человеку в его в интеллектуальной деятельности, то есть уровнем, который мы называем «слабый ИИ». Это неизбежно отвлечет их от усилий реализовать машинный аналог человеческого разума — то есть то, что мы называем «сильный ИИ»[73].

Нильсону вторят такие патриархи, как Марвин Мински, Джон Маккарти и Патрик Уинстон[74].

В последние годы наблюдается возрождение интереса к искусственному интеллекту, который вполне может обернуться новыми попытками создать *универсальный* ИИ (по Нильсону — сильный ИИ). Эти проекты будут поддерживаться, с одной стороны, производством новейших аппаратных средств, с другой — научным прогрессом в информатике и программировании

в целом, во многих специализированных предметных сферах в частности, а также в смежных областях, например нейроинформатике. Себастиан Трун и Питер Норвиг подготовили в Стэнфордском университете на осень 2011 года бесплатный онлайновый вводный курс по искусственному интеллекту. Реакцию на объявление о нем можно рассматривать как самый убедительный показатель неудовлетворенного спроса на качественную информацию и образование — на курс записались около 160 тысяч человек со всего мира (окончили его 23 тысячи)[75].

Существует множество вариантов экспертных оценок относительно будущего, уготованного искусственному интеллекту. Разногласия касаются и времени его появления, и того вида, в каком он когда-нибудь предстанет перед миром. Как заметили авторы одного недавнего исследования, прогнозы перспектив развития ИИ «различны настолько, насколько они категоричны»[76].

Мы не в состоянии охватить полную картину всех современных положений об интересующей нас теме, однако некоторое, пусть даже поверхностное, представление дают скупые опросы специалистов и высказанные ими частные мнения. Например, не так давно мы попросили представителей нескольких экспертных сообществ ответить на вопрос, когда они ожидают появления искусственного интеллекта человеческого уровня (ИИЧУ) — причем *уровень* определялся как «способность освоить большинство профессий, по крайней мере тех, которыми мог бы владеть среднестатистический человек». Респондентов просили строить свои предположения на основании того, что «научная деятельность в этом направлении будет продолжаться без серьезных сбоев»[77]. Ответы специалистов показаны в табл. 2. По данным выборки получились следующие средние оценки:

— 2022 год — средний прогноз с 10-процентной вероятностью;
— 2040 год — средний прогноз с 50-процентной вероятностью;
— 2075 год — средний прогноз с 90-процентной вероятностью.

Поскольку размер выборки слишком мал, а с точки зрения генеральной совокупности опрошенных ее нельзя считать репрезентативной, то результаты стоит рассматривать с некоторой долей скептицизма. Однако они согласуются с результатами других опросов[78].

Данные упомянутого опроса также соответствуют мнению примерно двух десятков исследователей, интервью с которыми появились за последние несколько лет. Назову только Нильса Нильсона. Ученый, многие десятилетия

плодотворно трудившийся над фундаментальными вопросами ИИ (методы поиска, автоматическое планирование, системы представления знаний, робототехника), написавший несколько учебников, недавно завершивший самую подробную историю исследований ИИ[79], — когда его спросили о сроках появления ИИЧУ, Нильсон дал следующее заключение[80]:

— 2030 год — средний прогноз с 10-процентной вероятностью;
— 2050 год — средний прогноз с 50-процентной вероятностью;
— 2100 год — средний прогноз с 90-процентной вероятностью.

Таблица 2. Когда будет создан искусственный интеллект человеческого уровня?[81]

	10%	**50%**	**90%**
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
Топ-100	2024	2050	2070
В среднем	2022	2040	2075

Судя по опубликованным интервью, названное профессором Нильсоном распределение вероятности вполне репрезентативно — многие эксперты думали так же. Однако еще раз хочу подчеркнуть: мнения расходились очень сильно, поскольку некоторые специалисты-практики горячо верили, что ИИЧУ будет создан за период 2020–2040 годов, а некоторые ученые были убеждены, что либо этого не случится никогда, либо это произойдет, но в неопределенно далеком будущем[82]. Кроме того, одни интервьюируемые считали, что определение «человеческого уровня» по отношению к искусственному интеллекту сформулировано некорректно и может вводить в заблуждение, а другие — по каким-то своим соображениям — просто воздержались от прогнозов.

На мой взгляд, прогнозы, отодвигающие создание ИИЧУ на более поздние сроки (по средним цифрам, полученным в результате опросов), определенно пессимистичны. 10-процентная вероятность появления ИИЧУ в 2075, и тем более в 2100 году (даже при условии, что «научная деятельность в этом направлении будет продолжаться без серьезных сбоев») представляется слишком низкой.

История показывает, что исследователи не могут похвастаться способностью предсказывать ни успехи в разработках искусственного интеллекта,

ни формы его воплощения. С одной стороны, выяснилось, что некоторые задачи, скажем, игра в шахматы, могут быть решены при помощи удивительно простых программ, и скептики, заявлявшие, будто машины «никогда» не смогут делать те или иные вещи, раз за разом оказываются посрамлены. С другой — наиболее типичной ошибкой специалистов является недооценка трудностей, связанных с разработкой устойчивой интеллектуальной системы, способной справляться с задачами реальной жизни, и переоценка возможностей их собственных проектов или методов.

В ходе одного из опросов были заданы еще два вопроса, актуальные для нашего исследования. Респондентов спросили, сколько, по их мнению, потребуется времени после создания ИИЧУ, чтобы машина смогла развить сверхразум. Ответы приведены в табл. 3. Второй вопрос касался темы долговременного воздействия на человечество, которое будет оказывать ИИЧУ. Ответы суммированы на рис. 2.

Таблица 3. Сколько времени пройдет между созданием искусственного интеллекта человеческого уровня и появлением сверхразума?

	Меньше двух лет	**Меньше 30 лет**
Топ-100	5%	50%
В среднем	10%	75%

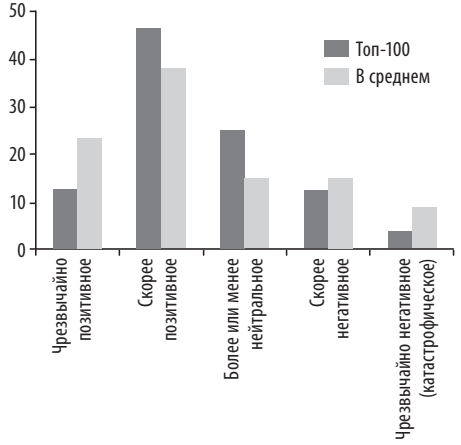

Рис. 2. Долговременное воздействие искусственного интеллекта человеческого уровня[83]

Мое мнение снова расходится с теми, которые были высказаны в ходе опроса. Я считаю гораздо более вероятным, что сверхразум появится сравнительно быстро после создания ИИЧУ. Кроме того, мой взгляд на последствия этого события также принципиально другой: вероятность чрезвычайно сильного воздействия — позитивного или негативного — на человечество гораздо более высока, чем вероятность нейтрального влияния. Причины этого вскоре станут ясны.

Не стоит полагаться всерьез ни на экспертные опросы, ни на интервью — в силу больших погрешностей данных методов. Небольшая выборка, ее возможные ошибки, а самое главное, ненадежность, изначально присущая субъективным мнениям, — все это не позволяет нам прийти к строгим умозаключениям. Однако пусть поверхностные — за неимением более достоверных аналитических данных, — но какие-то выводы мы в состоянии сделать. Во-первых, искусственный интеллект человеческого уровня имеет довольно высокую вероятность быть созданным к середине нынешнего столетия и имеет ненулевую вероятность быть созданным немного ранее или много позже. Во-вторых, после его создания, скорее всего, довольно быстро появится сверхразум. В-третьих, появление сверхразума может привести к огромным последствиям — как чрезвычайно позитивным, так и чрезвычайно негативным, вплоть до гибели человечества[84].

Полученные выводы по меньшей мере говорят нам, что тема заслуживает тщательного рассмотрения.

Глава вторая

Путь к сверхразуму

На сегодняшний день, если брать уровень общего интеллектуального развития, машины абсолютно уступают людям. Но однажды — по нашему предположению — разум машины превзойдет разум человека. Каков будет наш путь от нынешнего момента до того, который нас ожидает? В этой главе описаны несколько возможных технологических вариантов. Сначала мы рассмотрим такие темы, как искусственный интеллект, полная эмуляция головного мозга, усовершенствование когнитивных способностей человека, нейрокомпьютерный интерфейс, сети и организации. Затем оценим перечисленные аспекты с точки зрения вероятности, смогут ли они служить ступенями восхождения к сверхразуму. При нескольких вариантах пути шанс когда-нибудь достигнуть места назначения явно повышается.

Предварительно определим понятие сверхразума. Это *любой интеллект, значительно превосходящий когнитивные возможности человека фактически в любых областях*[1]. В следующей главе мы более подробно обсудим, что такое сверхразум, разложим его на составляющие и дифференцируем все возможные его воплощения. Но сейчас позволим себе ограничиться такой общей и поверхностной характеристикой. Заметьте, в данном описании не нашлось места ни претворению сверхразума в жизнь, ни его квалиа*, то есть будет ли он наделен субъективными переживаниями и опытом сознания. А ведь в определенном смысле, особенно этическом, вопрос весьма немаловажный. Однако сейчас, оставив в стороне интеллектуальную метафизику[2], мы уделим внимание двум вопросам: предпосылкам возникновения сверхразума и последствиям этого явления.

---

* *Квалиа* (от множ. числа лат. *qualia* — «свойства, качества») — философский термин, обозначающий субъективные ощущения, свойства чувственного опыта.

Согласно нашему определению, шахматная программа Deep Fritz не является сверхинтеллектуальной, поскольку «сильна» лишь в очень узкой — игра в шахматы — области. И тем не менее очень важно, чтобы сверхразум имел свои предметные специализации. Поэтому каждый раз, когда речь зайдет о том или ином сверхинтеллектуальном поведении, ограниченном предметной областью, я буду отдельно оговаривать его конкретную сферу деятельности. Например, искусственный интеллект, значительно превышающий умственные способности человека в сферах программирования и конструирования, получит название инженерного сверхинтеллекта. Но для обозначения систем, *в целом* превосходящих общий уровень человеческого интеллекта — если не указано иное, — остается термин *сверхразум*.

Как мы достигнем того времени, когда окажется возможным его появление? Какой путь выберем? Давайте рассмотрим некоторые возможные варианты.

Искусственный интеллект

Дорогой читатель, не стоит ожидать от этой главы концептуальной разработки вопроса, как создать универсальный, или сильный, искусственный интеллект. Проекта по его программированию просто не существует. Но даже будь я счастливым обладателем такого плана, то, безусловно, не стал бы обнародовать его в своей книге. (Если причины этого пока не очевидны, надеюсь, в последующих главах мне удастся недвусмысленно разъяснить собственную позицию.)

Однако уже сегодня можно распознать некоторые обязательные характеристики, присущие подобной интеллектуальной системе. Совершенно очевидно, что способность к обучению как неотъемлемое свойство ядра системы должна закладываться при проектировании, а не добавляться в качестве запоздалого соображения позднее в виде расширения. То же самое касается способности эффективно работать с неопределенной и вероятностной информациями. Скорее всего, среди основных модулей современного ИИ должны быть средства извлечения полезной информации из данных от внешних и внутренних датчиков и преобразования полученных концепций в гибкие комбинаторные представления для дальнейшего использования в мыслительных процессах, основанных на логике и интуиции.

Первые системы классического искусственного интеллекта по преимуществу не были нацелены на обучение, работу в условиях неопределенности

и формирование концепций — вероятно, из-за того, что в те времена были недостаточно развиты соответствующие методы анализа. Нельзя сказать, что все базовые идеи ИИ являются принципиально новаторскими. Например, мысль использовать обучение как средство развития простой системы и доведения ее до человеческого уровня была высказана еще Аланом Тьюрингом в 1950 году в статье «Вычислительная техника и интеллект», где он изложил свою концепцию «машина-ребенок»:

> Почему бы нам, вместо того чтобы пытаться создать программу, имитирующую ум взрослого, не попытаться создать программу, которая бы имитировала ум ребенка? Ведь если ум ребенка получает соответствующее воспитание, он становится умом взрослого человека*[3].

Тьюринг предвидел, что для создания «машины-ребенка» потребуется итеративный процесс:

> Вряд ли нам удастся получить хорошую «машину-ребенка» с первой же попытки. Надо провести эксперимент по обучению какой-либо из машин такого рода и выяснить, как она поддается научению. Затем провести тот же эксперимент с другой машиной и установить, какая из двух машин лучше. Существует очевидная связь между этим процессом и эволюцией в живой природе…
>
> Тем не менее можно надеяться, что этот процесс будет протекать быстрее, чем эволюция. Выживание наиболее приспособленных является слишком медленным способом оценки преимуществ. Экспериментатор, применяя силу интеллекта, может ускорить процесс оценки. В равной степени важно и то, что он не ограничен использованием только случайных мутаций. Если экспериментатор может проследить причину некоторого недостатка, он, вероятно, в состоянии придумать и такого рода мутацию, которая приведет к необходимому улучшению**[4].

Мы знаем, что слепые эволюционные процессы способны привести к появлению общего интеллекта человеческого уровня — по крайней мере один раз это уже случилось. Вследствие прогнозирования эволюционных процессов — то есть генетического программирования, когда алгоритмы разрабатываются и управляются разумным человеком-программистом, — мы должны получить аналогичные результаты с гораздо большей

---

* *А. Тьюринг*. Может ли машина мыслить? / Пер. с англ. Ю. А. Данилова. М.: Гос. изд-во физико-математической литературы, 1960. С. 34.

** *А. Тьюринг*. Может ли машина мыслить? С. 35.

эффективностью. Именно на это положение опираются многие ученые, среди которых философ Дэвид Чалмерс и исследователь Ханс Моравек[5], утверждающие, что ИИЧУ не только теоретически возможен, но и практически осуществим уже в XXI столетии. По их мнению, в деле создания интеллекта, оценивая относительные возможности эволюции и человеческой инженерной мысли, мы обнаружим, что последняя во многих областях значительно превосходит эволюцию и, скорее всего, довольно скоро обгонит ее в оставшихся. Таким образом, если в результате эволюционных процессов когда-то появился естественный интеллект, то из этого следует, что человеческие замыслы в области проектирования и разработок вскоре смогут привести нас к искусственному интеллекту. Например, Моравек писал еще в 1976 году:

> Существование нескольких примеров интеллекта, появившегося в условиях такого рода ограничений, должно вселять в нас уверенность, что очень скоро мы сможем достичь того же. Ситуация аналогична истории создания машин, которые могут летать, хотя они тяжелее воздуха: птицы, летучие мыши и насекомые продемонстрировали эту возможность явно задолго до того, как человек сделал летательные аппараты[6].

Впрочем, следует быть осторожнее с выводами, построенными на подобной цепочке рассуждений. Конечно, нет сомнений, что полет нечеловеческих живых существ, которые тяжелее воздуха, стал возможен в результате эволюции намного раньше того, как в этом преуспели люди — правда, преуспели при помощи механизмов. В поддержку этого можно вспомнить и другие примеры: гидролокационные системы; магнитометрические системы навигации; химические средства ведения войны; фотодатчики и прочие приспособления, обладающие механическими и кинетическими характеристиками эффективности. Однако с таким же успехом мы перечислим области, в которых результативность человеческих усилий еще очень далека от эффективности эволюционных процессов: морфогенез; механизмы самовосстановления; иммунная защита. Таким образом, аргументация Моравека все-таки не «вселяет в нас уверенность», что ИИУЧ будет создан «очень скоро». В лучшем случае верхним пределом сложности создания интеллекта может служить эволюция разумной жизни на Земле. Но этот уровень пока недосягаем для нынешних технологических возможностей человечества.

Еще один довод в пользу развития искусственного интеллекта по модели эволюционного процесса — это возможность запускать генетические алгоритмы на довольно мощных процессорах и в итоге добиться результатов, соизмеримых с теми, которые получились в ходе биологической эволюции. Таким образом, эта версия аргументации предполагает усовершенствовать ИИ посредством определенного метода.

Насколько справедливо утверждение, что довольно скоро в нашем распоряжении окажутся вычислительные ресурсы, достаточные для воспроизведения соответствующих эволюционных процессов, вследствие которых образовался человеческий интеллект? Ответ зависит от следующих условий: во-первых, будет ли в течение следующих десятилетий достигнут значимый прогресс компьютерных технологий; во-вторых, какая потребуется вычислительная мощность, чтобы механизмы запуска генетических алгоритмов были аналогичны естественному отбору, приведшему к появлению человека. Надо сказать, что выводы, к которым мы приходим по цепочке наших рассуждений, крайне неопределенны; но, несмотря на такой обескураживающий факт, все-таки представляется уместным попробовать дать хотя бы приблизительную оценку этой версии (см. врезку 3). За неимением других возможностей даже ориентировочные расчеты привлекут внимание к некоторым любопытным неизвестным величинам.

Суть в том, что вычислительная мощность, требуемая лишь для воспроизведения нужных эволюционных процессов, приведших к появлению человеческого интеллекта, практически недостижима и надолго останется таковой, даже если закон Мура будет действовать еще целое столетие (см. рис. 3 на с. 57). Однако существует вполне приемлемый выход: мы очень сильно повлияем на эффективность, когда вместо прямолинейного повторения естественных эволюционных процессов разработаем поисковый механизм, ориентированный на создание интеллекта, задействуя самые разные очевидные преимущества по сравнению с естественным отбором. Безусловно, оценить количественно полученный выигрыш в эффективности сейчас очень трудно. Мы даже не знаем, о каких порядках величины идет речь — пяти или двадцати пяти. Следовательно, если построенная на эволюционной модели аргументация не будет разработана должным образом, мы не сможем удовлетворить свои ожидания и никогда не узнаем, насколько сложны дороги к искусственному интеллекту человеческого уровня и как долго нам ожидать его появления.

ВРЕЗКА 3. ОЦЕНКА УСИЛИЙ ПО ВОСПРОИЗВЕДЕНИЮ ЭВОЛЮЦИОННОГО ПРОЦЕССА

Не все достижения антропогенеза, имеющие отношение к человеческому разуму, имеют ценность для современных специалистов, работающих над проблемой эволюционного развития искусственного интеллекта. В дело идет лишь незначительная часть того, что получилось в итоге естественного отбора на Земле. Например, проблемы, которые люди не могут не принимать во внимание, являются результатом лишь незначительных эволюционных усилий. В частности, поскольку мы можем питать наши компьютеры электричеством, у нас нет необходимости заново изобретать молекулы системы клеточной энергетической экономики для создания разумных машин — а ведь на молекулярную эволюцию метаболического механизма, вполне возможно, потребовалась значительная часть общего расхода мощности естественного отбора, находившейся в распоряжении эволюции на протяжении истории Земли[7].

Существует концепция, что ключом к созданию ИИ является структура нервной системы, появившаяся меньше миллиарда лет назад[8]. Если мы примем данное положение, количество «экспериментов», необходимых для эволюции, значительно сократится. Сегодня в мире существует приблизительно $(4–6) \times 10^{30}$ прокариотов, но лишь 10^{19} насекомых и меньше 10^{10} представителей человеческого рода (кстати, численность населения накануне неолитической революции была на порядки меньше)[9]. Согласитесь, эти цифры не столь пугающи.

Однако для эволюционных алгоритмов требуется не только разнообразие вариантов, но и оценка приспособленности каждого из вариантов — обычно наиболее затратный компонент с точки зрения вычислительных ресурсов. В случае эволюции искусственного интеллекта для оценки приспособленности требуется, по всей видимости, моделирование нейронного развития, а также способности к обучению и познанию. Поэтому лучше не смотреть на общее число организмов со сложной нервной системой, а оценить количество нейронов в биологических организмах, которые нам, возможно, придется моделировать для расчета целевой функции эволюции. Грубую оценку можно сделать, обратившись к насекомым, которые доминируют в наземной биомассе (на долю одних только муравьев приходится 15–20%)[10]. Объем головного мозга насекомых зависит от многих факторов. Чем насекомое крупнее и социальнее (то есть ведет общественный образ жизни), тем больше его мозг; например, у пчелы чуть меньше 10^6 нейронов, у дрозофилы — 10^5 нейронов, муравей со своими 250 тысячами нейронов находится между ними[11]. Мозг большинства более мелких насекомых содержит всего несколько тысяч нейронов. Предлагаю с предельной осторожностью остановиться на усредненном значении (10^5) и приравнять к дрозофилам всех насекомых (которых всего в мире — 10^{19}), тогда суммарное число их нейронов составит 10^{24}. Добавим еще порядок величины за счет ракообразных, птиц, рептилий, млекопитающих и т. д. — и получим 10^{25}. (Сравним это с тем, что до возникновения сельского хозяйства на планете было меньше 10^7 человек, причем на каждого приходилось

примерно 10^{11} нейронов — то есть в общей сложности сумма всех нейронов составляла меньше чем 10^{18}, хотя человеческий мозг содержал — и содержит — намного больше синапсов.)

Вычислительные затраты на моделирование одного нейрона зависят от необходимой степени детализации модели. Для крайне простой модели нейрона, работающей в режиме реального времени, требуется примерно 1000 операций с плавающей запятой в секунду (далее — FLOPS). Для электро- и физиологически реалистичной модели Ходжкина–Хаксли нужно 1 200 000 FLOPS. Более сложная мультикомпонентная модель нейрона добавила бы два-три порядка величины, а модель более высокого уровня, оперирующая системами нейронов, требует на два-три порядка меньше операций на один нейрон, чем простые модели[12]. Если нам нужно смоделировать 10^{25} нейронов на протяжении миллиарда лет эволюции (это больше, чем срок существования нервных систем в их нынешнем виде) и мы позволим компьютерам работать над этой задачей в течение года, то требования к их вычислительной мощности попадут в диапазон 10^{31}–10^{44} FLOPS. Для сравнения, самый сверхмощный компьютер в мире китайский Tianhe-2 (на сентябрь 2013 года) способен выдавать всего $3{,}39 \times 10^{16}$ FLOPS. В последние десятилетия обычные компьютеры увеличивали свою производительность на порядок примерно раз в 6,7 года. Даже если вычислительная мощность станет расти по закону Мура в течение целого столетия, то это окажется недостаточным, чтобы преодолеть существующий разрыв. Использование более специализированных вычислительных систем или увеличение времени вычислений способны снизить требования к мощности всего на несколько порядков.

Оценка количества нейронов носит условный характер еще по одной причине. Природа, создавая человеческий разум, вряд ли ставила перед собой какую-то определенную задачу. Иными словами, целевая функция эволюционной системы отбирала организмы не только ради развития у них интеллекта или его предшественника — «конкретного мышления»[13]. Даже если организмы с лучшими способностями к обработке информации при определенных условиях извлекали дополнительные выгоды, то это обстоятельство не являлось главным фактором отбора особи, поскольку развитое мышление могло означать (и часто означало) возникновение дополнительных издержек: затрату большего количества энергии или более медленное созревание, — что перевешивало преимущества разумного поведения. Высокая смертность также снижала ценность интеллекта — чем короче средняя продолжительность жизни, тем меньше времени для того, чтобы «окупились» повышенные способности к обучению. Сниженное давление отбора замедляло распространение инноваций, основанных на интеллекте, и, как следствие, уменьшало возможность отбора последующих инноваций. Более того, эволюция могла тормозиться в локальных оптимумах, которые исследователи в состоянии заметить и обойти за счет изменения баланса между поиском и памятью или за счет плавного повышения сложности тестов на интеллект[14]. Как уже говорилось ранее, эволюция тратит значительную часть мощности отбора на свойства, не имеющие отношения к интеллекту, — скажем,

на эволюционную конкуренцию между иммунной системой и паразитами, названную «гонка Черной королевы». Эволюция продолжает растрачивать ресурсы на заведомо обреченные мутации и неспособна принимать во внимание статистическое сходство различных мутаций. Приведенные здесь примеры не должны отпугивать специалистов, разрабатывающих эволюционные алгоритмы для создания интеллектуальных программ, так как неэффективность естественного отбора (с точки зрения развития интеллекта) довольно легко преодолима.

Вполне вероятно, что устранение такого рода неэффективности поможет сэкономить несколько порядков требуемой мощности в 10^{31}–10^{44} FLOPS, рассчитанной ранее. К сожалению, трудно сказать, сколько именно. Трудно дать даже приблизительную оценку — можно только гадать, будет ли это пять порядков, десять или двадцать пять[15].

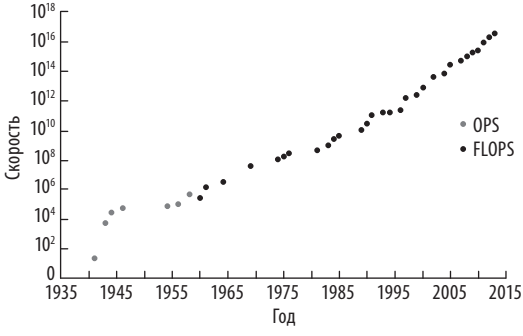

Рис. 3. Производительность сверхмощных компьютеров. В прямом смысле то, что называют «закон Мура», — это наблюдение, согласно которому количество транзисторов, размещаемых на кристалле интегральной схемы, удваивается примерно каждые два года. Однако часто закон обобщают, считая, что так же по экспоненте растут и другие показатели производительности компьютеров. На нашем графике показано изменение во времени пиковой скорости наиболее сверхмощных компьютеров в мире (по логарифмической вертикальной шкале). В последние годы скорость последовательных вычислений расти перестала, но за счет распространения параллельных вычислений общее количество операций продолжает увеличиваться с прежним темпом[16].

Есть еще одно осложнение, связанное с эволюционными факторами, выдвигаемыми в качестве последнего аргумента. Проблема заключается в том, что мы не в состоянии вычислить — даже очень приблизительно — верхнюю границу трудности получения интеллекта эволюционным путем. Да, на Земле когда-то появилась разумная жизнь, но из этого факта еще не следует, будто процессы эволюции с высокой степенью вероятности

приводят к возникновению интеллекта. Подобное заключение было бы в корне ошибочным, поскольку не учитывается так называемый эффект наблюдения при отборе, подразумевающий, что все наблюдатели находятся на планете, где зародилась разумная жизнь, независимо от того, насколько вероятно или невероятно такое событие на любой другой планете. Предположим, для появления разумной жизни, помимо систематических погрешностей естественного отбора, требуется огромное количество *удачных совпадений* — настолько большое, что разумная жизнь появилась всего лишь на одной из 10^{30} планет, где существуют простые гены-репликаторы. В таком случае исследователи, запуская генетические алгоритмы в попытке воспроизвести созданное эволюцией, могут столкнуться с тем, что понадобится сделать примерно 10^{30} итераций, прежде чем они найдут комбинацию, в которой все элементы сложатся правильно. Кажется, это вполне согласуется с нашим наблюдением, что жизнь зародилась и развивалась здесь, на Земле. Обойти данный гносеологический барьер отчасти можно путем тщательных и до некоторой степени громоздких логических ходов — анализируя случаи конвергентной эволюции характеристик, имеющих отношение к интеллекту, и принимая во внимание эффект наблюдения при отборе. Если ученые не возьмут на себя труд провести такой анализ, то в дальнейшем уже никому из них не придется оценивать максимальное значение и выяснить, насколько предполагаемая верхняя граница необходимой вычислительной мощности для воспроизведения эволюции интеллекта (см. врезку 3) может оказаться ниже тридцатого порядка (или какой-то другой столь же большой величины)[17].

Перейдем к следующему варианту достижения нашей цели: аргументом в пользу осуществимости эволюции искусственного интеллекта служит деятельность головного мозга человека, на которую ссылаются как на базовую модель для ИИ. Различные версии такого подхода отличаются лишь степенью воспроизведения — насколько точно предлагается имитировать функции биологического мозга. На одном полюсе, представляющем собой своеобразную «игру в имитацию», мы имеем концепцию *полной эмуляции мозга*, то есть полномасштабного имитационного моделирования головного мозга (к этому мы вернемся немного позже). На другом полюсе находятся технологии, в соответствии с которыми функциональность мозга служит лишь стартовой точкой, но разработка низкоуровневого моделирования не планируется. В конечном счете мы приблизимся к пониманию общей

идеи деятельности мозга, чему способствуют успехи в нейробиологии и когнитивной психологии, а также постоянное совершенствование инструментальных и аппаратных средств. Новые знания, несомненно, станут ориентиром в дальнейшей работе с ИИ. Нам уже известен пример ИИ, появившегося в результате моделирования работы мозга, — это нейронные сети. Еще одна идея, взятая из нейробиологии и перенесенная на машинное обучение, — иерархическая организация восприятия. Изучение обучения с подкреплением было обусловлено (по крайней мере частично) той важной ролью, которую эта тема играет в психологических теориях, описывающих поведение и мышление животных, а также техники обучения с подкреплением (например, TD-алгоритм). Сегодня обучение с подкреплением широко применяется в системах ИИ[18]. В будущем подобных примеров, безусловно, будет больше. Поскольку набор базовых механизмов функционирования мозга весьма ограничен — на самом деле их очень небольшое количество, — все эти механизмы рано или поздно будут открыты благодаря постоянным успехам нейробиологии. Однако возможен вариант, что еще раньше придет к финишу некий гибридный подход, сочетающий модели, разработанные, с одной стороны, на основе деятельности головного мозга человека, с другой — исключительно на основе технологий искусственного интеллекта. Совсем не обязательно, что полученная в результате система должна во всем напоминать головной мозг, даже если при ее создании и будут использованы некоторые принципы его деятельности.

Деятельность головного мозга человека в качестве базовой модели представляет собой сильный аргумент в пользу осуществимости создания и дальнейшего развития искусственного интеллекта. Однако ни один даже самый мощный довод не приблизит нас к пониманию будущих сроков, поскольку трудно предсказать, когда произойдет то или иное открытие в нейробиологии. Можно сказать только одно: чем глубже в будущее мы заглядываем, тем больше вероятность, что секреты функционирования мозга будут раскрыты достаточно полно для воплощения систем искусственного интеллекта.

Исследователи, работающие в области искусственного интеллекта, придерживаются разных точек зрения относительно того, насколько многообещающим является нейроморфный подход сравнительно с технологиями, основанными на полностью композиционных подходах. Полет птиц демонстрировал физическую возможность появления летающих механизмов

тяжелее воздуха, что в итоге привело к строительству летательных аппаратов. Однако даже первые поднявшиеся в воздух аэропланы не взмахивали крыльями. По какому пути пойдет разработка искусственного интеллекта? Вопрос остается открытым: по принципу ли закона аэродинамики, удерживающего в воздухе тяжелые железные механизмы, — то есть учась у живой природы, но не подражая ей напрямую; по принципу ли устройства двигателя внутреннего сгорания — то есть непосредственно копируя действия природных сил.

Концепция Тьюринга о разработке программы, получающей большую часть знаний за счет обучения, а не в результате задания исходных данных, применима и к созданию искусственного интеллекта — как к нейроморфному, так и композиционному подходам.

Вариацией тьюринговой концепции «машины-ребенка» стала идея зародыша ИИ[19]. Однако если «машине-ребенку», как это представлял Тьюринг, полагалось иметь относительно фиксированную архитектуру и развивать свой потенциал за счет накопления *контента*, зародыш ИИ будет более сложной системой, самосовершенствующей собственную *архитектуру*. На ранних стадиях существования зародыш ИИ развивается в основном за счет сбора информации, действуя методом проб и ошибок не без помощи программиста. «Повзрослев», он должен научиться самостоятельно *разбираться* в принципах своей работы, чтобы уметь проектировать новые алгоритмы и вычислительные структуры, повышающие его когнитивную эффективность. Требуемое понимание возможно лишь в тех случаях, когда зародыш ИИ или во многих областях достиг довольно высокого общего уровня интеллектуального развития, или в отдельных предметных областях — скажем, кибернетике и математике — преодолел некий интеллектуальный порог.

Это подводит нас к еще одной важной концепции, получившей название «рекурсивное самосовершенствование». Успешный зародыш ИИ должен быть способен к постоянному саморазвитию: первая версия создает улучшенную версию самой себя, которая намного умнее оригинальной; улучшенная версия, в свою очередь, трудится над еще более улучшенной версией и так далее[20]. При некоторых условиях процесс рекурсивного самосовершенствования может продолжаться довольно долго и в конце концов привести к взрывному развитию искусственного интеллекта. Имеется в виду событие, в ходе которого за короткий период времени общий интеллект системы

вырастает со сравнительно скромного уровня (возможно, во многих аспектах, кроме программирования и исследований в области ИИ, даже ниже человеческого) до сверхразумного, радикально превосходящего уровень человека. В четвертой главе мы вернемся к этой перспективе, весьма важной по своему значению, и подробнее проанализируем динамику развития событий.

Обратите внимание, что такая модель развития предполагает возможность сюрпризов. Попытки создать универсальный искусственный интеллект могут, с одной стороны, закончиться полной неудачей, а с другой — привести к последнему недостающему критическому элементу — после чего зародыш ИИ станет способен на устойчивое рекурсивное самосовершенствование.

Прежде чем закончить этот раздел главы, хотелось бы подчеркнуть еще одну вещь: совсем не обязательно, чтобы искусственный интеллект был уподоблен человеческому разуму. Вполне допускаю, что ИИ станет совершенно «чужим» — скорее всего, так и случится. Можно ожидать, что когнитивная архитектура ИИ будет резко отличаться от когнитивной системы человека; например, на ранних стадиях когнитивная архитектура будет иметь совсем другие сильные и слабые признаки (хотя, как мы увидим далее, ИИ удастся преодолеть исходные недостатки). Помимо всего, целеустремленные системы ИИ могут не иметь ничего общего с системой целеустремлений человечества. Нет оснований утверждать, что ИИ среднего уровня начнет руководствоваться человеческими чувствами, такими как любовь, ненависть, гордость, — для такой сложной адаптации потребуется огромный объем дорогостоящих работ, более того, к появлению подобной возможности у ИИ следует отнестись очень осмотрительно. Это одновременно и большая проблема, и большие возможности. Мы вернемся к мотивации ИИ в дальнейших главах, но эта идея настолько важна для книги, что ее стоит держать в голове постоянно.

Полная эмуляция головного мозга человека

В процессе полномасштабного имитационного моделирования головного мозга, который мы называем «полная эмуляция мозга» или «загрузка разума», искусственный интеллект создается путем сканирования и точного воспроизведения вычислительной структуры биологического мозга. Таким образом, приходится всецело черпать вдохновение у природы — крайний

случай неприкрытого плагиата. Чтобы полная эмуляция мозга прошла успешно, требуется выполнить ряд определенных шагов.

Первый этап. Делается довольно подробное сканирование человеческого мозга. Это может включать фиксацию мозга умершего человека методом витрификации, или стеклования (в результате ткани становятся твердыми, как стекло). Затем одним аппаратом с ткани делаются тонкие срезы, которые пропускают через другой аппарат для сканирования, возможно, при помощи электронных микроскопов. На этой стадии применяется окраска материала специальными красителями, чтобы выявить его структурные и химические свойства. При этом параллельно работают множество сканирующих аппаратов, одновременно обрабатывающих различные срезы ткани.

Второй этап. Исходные данные со сканеров загружают в компьютер для автоматической обработки изображений, чтобы реконструировать трехмерную нейронную сеть, отвечающую за познание в биологическом мозгу. Дабы сократить количество снимков в высоком разрешении, которые необходимо хранить в буфере, этот этап может выполняться одновременно с первым. Полученную карту комбинируют с библиотекой нейровычислительных моделей на нейронах разного типа или на различных нейронных элементах (например, могут отличаться синапсы). Некоторые результаты сканирования и обработки изображений с применением современной технологии показаны на рис. 4.

Третий этап. Нейросетевая вычислительная структура, полученная на предыдущем этапе, загружается в довольно мощный компьютер. В случае полного успеха результат станет цифровой копией исходного интеллекта с неповрежденной памятью и нетронутым типом личности. Эмуляция человеческого разума теперь существует в виде программного обеспечения на компьютере. Разум может как обитать в виртуальном пространстве, так и взаимодействовать с реальным миром при помощи роботизированных конечностей.

Работа над полной эмуляцией мозга не предполагает, что исследователи должны разбираться в процессе познания или программировании искусственного интеллекта. Им нужно лишь быть высокими профессионалами в таком вопросе, как низкоуровневые функциональные характеристики базовых вычислительных элементов мозга. Для успешно проведенной эмуляции не потребуются ни фундаментальные концепции, ни теоретические открытия.

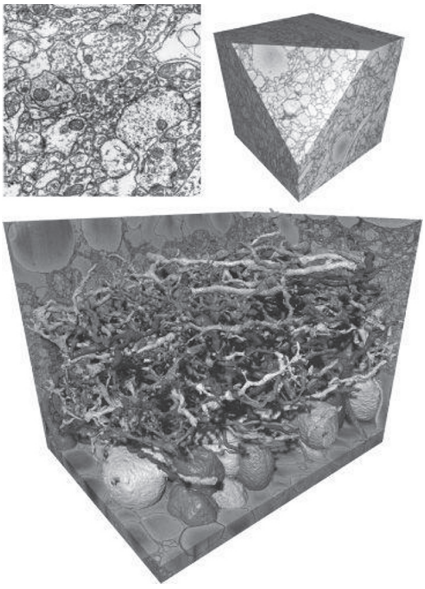

Рис. 4. Трехмерная реконструкция срезов, проведенная нейроанатомическим методом (изображения под электронным микроскопом). *Слева вверху*: дендриты и аксоны — типовой электронно-микроскопический снимок, показывающий поперечное сечение нейронов. *Справа вверху*: объемная реконструкция среза сетчатки глаза кролика, полученная по снимкам сканирующего электронного микроскопа[21]. Отдельные двумерные снимки «складываются» в куб со стороной примерно 11 мкм. *Внизу*: реконструкция подмножества нейронных проекций, составляющих нейропиль, созданная с помощью алгоритма автоматической сегментации[22].

Без применения самых передовых технологий полная эмуляция головного мозга практически неосуществима. Прежде всего нужно, имея в наличии необходимое оборудование и соблюдая все условия, провести три главные манипуляции:

1) *сканирование* — высокопроизводительные микроскопы с хорошим разрешением, дающие возможность обнаружить нужные свойства;
2) *трансляция* — автоматизированный анализ изображений для перевода исходных данных сканирования в связанную трехмерную модель из релевантных вычислительных элементов;
3) *моделирование* — компьютер, достаточно мощный для обработки полученной оцифрованной структуры.

По сравнению с перечисленными этапами, связанными с довольно напряженным и высокоточным трудом (см. табл. 4), покажется относительно незамысловатым делом разработать базовую виртуальную реальность или роботизированную внешность с аудиовизуальным каналом для ввода данных и каким-нибудь простым каналом для их вывода. Отвечающие минимальным требованиям простые системы ввода и вывода, похоже, можно получить, даже с помощью имеющихся под рукой технологий и оборудования[23].

Таблица 4. Технологии, необходимые для полной эмуляции мозга человека

Сканирование	Предварительная обработка/фиксация		Подготовка тканей, сохранение их микроструктуры и состояния
	Физические манипуляции		Методы манипуляции тканями мозга и их срезами до, во время и после сканирования
	Получение изображения	Объем	Возможность сканировать весь объем мозга в приемлемые сроки и с приемлемыми издержками
		Разрешение	Разрешение должно быть соответствующим для реконструкции
		Функциональная информация	Возможность обнаружить все функционально значимые свойства ткани
Трансляция	Обработка изображений	Коррекция пропорций	Устранение диспропорций из-за несовершенства сканирования
		Интерполяция данных	Восполнение утерянных данных
		Устранение помех	Повышение качества изображения
		Трассировка данных	Определение структуры и ее обработка для получения цельной трехмерной модели ткани
	Интерпретация изображения	Идентификация типа клетки	Идентификация типа клетки

		Идентификация синапса	Идентификация синапсов и видов синаптических контактов
		Оценка параметров	Оценка функционально значимых параметров клеток, синапсов и прочих объектов
		Создание базы данных	Создание эффективного хранилища полученной информации
	Модель нейронной системы	Математическая модель	Модель совокупности объектов и их поведения
		Эффективная реализация	Реализация модели
Моделирование	Хранение		Хранение оригинальной модели и ее текущего состояния
	Пропускная способность		Эффективная межпроцессорная связь
	Микропроцессор		Мощность процессора, достаточная для запуска модели
	Моделирование организма		Моделирование организма, способного существовать в виртуальной среде и взаимодействовать с внешним миром посредством роботизированных механизмов
	Моделирование взаимодействия со средой		Виртуальная среда для виртуального организма

Мы неслучайно надеемся, что необходимые инновационные технологии пусть не в ближайшем будущем, но когда-нибудь станут достижимыми. У нас уже существуют более или менее точные компьютерные модели многих типов биологических нейронов и нейронных процессов. Разработано программное обеспечение для распознавания образов, способное отследить аксоны и дендриты в стопке двумерных изображений (хотя их точность еще предстоит повысить). Имеются средства съемки с нужным разрешением — с помощью сканирующего туннельного микроскопа можно «увидеть» отдельные атомы, причем с разрешением гораздо выше необходимого. Полная эмуляция головного мозга человека потребует весьма мощного технологи-

ческого прорыва — и это отлично понимают все исследователи; однако имеющийся на сегодняшний день багаж знаний и возможностей дает все основания полагать, что нет никакого непреодолимого барьера для появления нужных технологий[24]. Например, необязательно иметь микроскопы с высочайшим разрешением — должны быть просто очень мощные микроскопы. Слишком затратно по времени и стоимости использовать для съемки исследуемого материала туннельные сканирующие микроскопы с атомным разрешением. Наверное, более оправданным стало бы применение электронных микроскопов с меньшим разрешением, что, естественно, потребует лучшего обеспечения видимости важных элементов, скажем, синаптической микроструктуры, — в свою очередь, это повлечет разработку новых методов подготовки и окраски кортикальной ткани. Уже пора задуматься над такими вопросами, как расширение нейровычислительных библиотек, усовершенствование автоматизированной обработки образов и интерпретации результатов сканирования.

Осуществимость полной эмуляции головного мозга человека не зависит от теоретических разработок так сильно, как, скажем, создание искусственного интеллекта; загрузка разума в основном возлагается на технологические возможности. Требования к технологиям определяются лишь уровнем абстракции, на котором происходит эмулирование. В этом смысле придется искать баланс между теорией и технологией. С одной стороны, чем слабее наше сканирующее оборудование, чем менее производительны компьютеры, тем меньше мы можем рассчитывать на низкоуровневое имитационное моделирование физико-химических процессов головного мозга и тем больше потребуется теоретического понимания вычислительной архитектуры, которую мы стремимся моделировать, чтобы создать более абстрактные представления значимой функциональности[25]. С другой стороны, при достаточном количестве и качестве передовой сканирующей техники и сверхмощных вычислительных средств нам вряд ли понадобятся сильная теоретическая подготовка и профессиональные знания о происходящих в мозгу процессах — ведь при условии «технологического изобилия» мы сможем решить задачу моделирования методом простого перебора — то есть элементарно «в лоб». Правда, есть третий вариант: давайте проведем эмуляцию мозга на уровне элементарных частиц, то есть отнесемся к мозгу как к квантово-механической системе и решим нашу проблему с помощью уравнения Шрёдингера. В этом случае — совсем крайнем по своей маловероятности — нам

придется вовсе абстрагироваться от биологической модели и опираться исключительно на существующие знания физики. Но все размышления на тему элементарных частиц умозрительны, поскольку сразу возникает вопрос о требованиях, которые будут предъявляться к вычислительной мощности и обработке данных, — условия для нас совершенно невыполнимые. Гораздо более правдоподобным вариантом могла бы стать эмуляция отдельных нейронов и их матрицы смежности с построением структуры их дендритных деревьев и, возможно, каких-то статических переменных, описывающих индивидуальные синапсы; при этом, не трогая по отдельности молекулы-нейротрансмиттеры, возможно будет моделировать изменение их концентрации в виде грубой структуры.

Чтобы оценить осуществимость полной эмуляции головного мозга человека, необходимо определить критерии удачного завершения процесса. Вряд ли ученые стремятся создать детальную и точную модель мозга, которую можно было бы подвергать воздействию последовательности определенных стимулов, и на основании результатов точно предсказывать «поведение» биологического мозга. Безусловно, нет. Они хотят всего лишь воспроизвести вычислительные функциональные свойства мозга в таком объеме, чтобы использовать полученный эмулятор для выполнения интеллектуальных задач. При подобном целеполагании можно не принимать во внимание многие компоненты биологического мозга с его довольно сложными и запутанными структурами.

Чтобы понять, насколько удачно проведена эмуляция, следует оценить, в какой степени удалось сохранить церебральные функции обработки информации в полученном эмуляторе. Для этого придется провести более глубокий анализ. Например, можно выделить следующие виды эмуляторов мозга:

1) *высокоточная модель* — сохраняет всю совокупность знаний, навыков, возможностей и ценностей биологического мозга;
2) *искаженная модель* — с выраженными нечеловеческими задатками, но в основном способна выполнять ту же умственную работу, что и биологический мозг;
3) *обобщенная модель* (может быть искаженной) — в некотором смысле ребенок, лишенный навыков и воспоминаний, приобретенных биологическим мозгом взрослого человека, но способный научиться большинству вещей, которым обучается среднестатистический человек[26].

Если мы пойдем по этому пути, то естественно предположить, что высокоточная модель мозга в конце концов будет осуществлена, но *первая* модель, которую нам удастся создать, по всей видимости, окажется совсем простой. Прежде чем получить нечто, идеально работающее, вероятно, придется сделать нечто, работающее пока еще несовершенно. Более того, не исключено, что выбор такого пути, как эмуляция мозга, приведет нас к созданию нейроморфного ИИ, который будет основан на обнаруженных в процессе эмулирования принципах нейровычислительной системы и композиционных подходах. Скорее всего, этот промежуточный вариант появится раньше завершения полномасштабного имитационного моделирования головного мозга. Допустимость появления побочного результата вроде нейроморфного ИИ заставляет серьезно задуматься, стоит ли ускорять развитие технологии эмулирования, — более подробно мы рассмотрим этот вопрос в одной из следующих глав.

Насколько далеки мы сегодня от построения полной компьютерной модели мозга? В одной из недавно опубликованных работ приводится технический сценарий и делается вывод — правда, с широким интервалом неопределенности*, — что все необходимые условия будут созданы к середине нашего века[27]. Основные контрольные точки этого пути показаны на рис. 5. Однако кажущаяся простота сценария может быть обманчива: нам не следует недооценивать объем предстоящей работы. Пока еще не удалось разработать ни одну компьютерную модель биологического мозга. Возьмем *Caenorhabditis elegans* — прозрачный круглый червь (нематода), длиной около одного миллиметра и имеющий всего 302 нейрона. Это скромное неприметное создание служит для науки модельным организмом. Его коннектом, то есть полное описание структуры связей его нейронов, известен с середины 1980-х годов. Матрица связей была составлена после тщательно проделанной работы по подготовке материала: разделение ткани на слои, исследование под электронным микроскопом и нанесение маркировки на образцы вручную[28]. Но нам недостаточно просто видеть, как нейроны связаны друг с другом. Чтобы сконструировать компьютерную модель мозга, нужно знать, какие синапсы — возбуждающие, а какие — тормозящие; понимать устойчивость связей, а также различные динамические

---

* *Интервал неопределенности* (uncertainty interval) — количественная мера, характеризующая степень неопределенности результата оценки.

свойства аксонов, синапсов и дендритов. Полной информации пока нет даже в описании такой простой нервной системы, как у *C. elegans* (хотя сейчас это вполне доступно сделать в рамках целевого исследовательского проекта среднего размера)[29]. Успех, достигнутый при эмуляции крошечного мозга червя C. elegans, поможет лучше оценить тот путь, который нам предстоит пройти, чтобы провести эмуляцию мозга более крупных размеров.

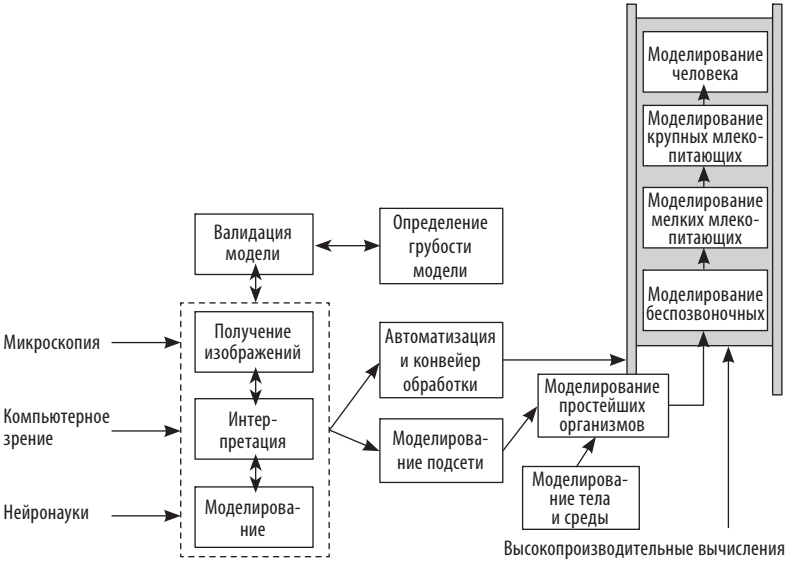

Рис. 5. Сценарий полной эмуляции мозга. Схематично изображены входные параметры, задачи и контрольные точки[30].

На определенной стадии развития технологий, когда появятся методы автоматического эмулирования небольших фрагментов тканей мозга, останется решить проблему масштабирования. Обратите внимание, что символическая «лестница», изображенная на рис. 5, передает последовательность шагов, которые должны быть пройдены для достижения цели. Каждая ступень лестницы соответствует определенной стадии полной эмуляции мозга — стадии, становящейся все более сложной с точки зрения строения нервной системы организмов, например: *нематода* → *пчела* → *мышь* → *макака-резус* → *человек*. Поскольку разрыв между ступенями — по крайней

мере после первого шага — носит характер скорее количественный, чем качественный, и определяется (в большей степени, хотя и не полностью) разницей в размерах моделируемого мозга, его можно преодолеть за счет относительно прямолинейного увеличения возможностей сканирования и моделирования[31].

Когда мы дойдем до последней ступени лестницы, перспектива осуществления компьютерной модели человеческого мозга станет более ясной[32]. Как только вдруг мы услышим о высокопроизводительном сканирующем оборудовании и сверхмощных компьютерах, то есть таком аппаратном обеспечении, которого не хватало, чтобы приступить к последнему этапу моделирования в реальном масштабе, — его появление послужит для нас предупредительным сигналом, что мы готовы вступить на путь полной эмуляции головного мозга и через какое-то время, возможно, окажемся в зоне действия искусственного интеллекта человеческого уровня. Но если последним недостающим звеном станет нейромоделирование, переход от невзрачных прототипов к работающей модели головного человеческого мозга может быть более резким. Тогда представим сценарий, по которому неожиданно обнаруживается — при всем изобилии программ для сканирования и быстродействующих компьютеров, — что модели биологических нейронов довольно трудно заставить работать правильно. Конечно, мы в конце концов справимся с этим глюком, и наши абсолютно беспомощные нейромодели, до той поры будто бы пребывавшие в обморочном состоянии после большого эпилептического припадка, бодро и согласованно возьмутся за дело. В этом случае успеху не будет предшествовать долгая череда эмуляций мозга живых организмов, чьи размеры последовательно возрастают — от червя до обезьяны; а научные изыскания не будет сопровождать вереница газетных статей, шрифт которых соответственно увеличивается от петита до крупного кегля. Тем из нас, кто внимательно следит за ходом событий, будет довольно трудно заранее определить, сколько еще остается недееспособных нейромоделей и как долго придется ждать устранения дефектов; вряд ли мы сразу поймем, что этот этап уже пройден и ученые стоят на пороге решающего достижения. (Как только удастся осуществить эмуляцию головного мозга человека, последуют и другие потенциально взрывоопасные открытия, но мы отложим их обсуждение до четвертой главы.)

Итак, мы видим, что допустим самый неожиданный сценарий, даже если все исследования будут проводиться совершенно открыто. Тем не менее

прогресс, связанный с полной эмуляцией мозга, вероятнее всего будет иметь вполне определенные и понятные признаки, поскольку изучение этой проблемы зависит от конкретных и доступных нашему пониманию технологий, тогда как характер развития искусственного интеллекта, от классического до универсального, базируется на теоретических разработках. Кроме того, в отличие от развития ИИ, в случае с загрузкой разума мы можем с большей уверенностью говорить, что компьютерная модель мозга вряд ли будет осуществима в ближайшем будущем (скажем, в течение пятнадцати лет) — из-за отсутствия на сегодняшний день даже исходных версий некоторых необходимых сложнейших технологических разработок. Напротив, в отношении искусственного интеллекта совсем не исключено, что, *в принципе,* некий человек *может* просто взять и написать программу зародыша ИИ для обычного современного компьютера, и соответственно не исключено — хотя и маловероятно, — что где-нибудь кто-нибудь разберется, как это сделать уже в ближайшее время.

Усовершенствование когнитивных способностей человека

Третий путь создания интеллекта, превосходящего человеческий, это улучшение функционирования биологического мозга. В принципе, этого можно было бы достичь без применения технологий, а за счет селекции. Однако любая попытка запустить классическую программу евгеники столкнется и с политическими, и этическими препятствиями. Кроме того, для получения сколько-нибудь значимых результатов — если только отбор не будет чрезмерно строгим — потребуется множество поколений. Задолго до того как такая программа принесет плоды, человечество в результате развития биотехнологий получит прямой контроль над генетикой и нейробиологией, что сделает ненужными проекты по селекции людей. Поэтому мы обращаем внимание на методы, которые приведут к результату намного быстрее: на протяжении жизни нескольких поколений и даже одного, — для этого есть потенциальные возможности.

Мы научились повышать свои индивидуальные познавательные, или когнитивные, способности разными способами, в том числе не пренебрегая и традиционными, например обучением и тренировкой. Развитие нервной системы можно ускорить за счет таких низкотехнологичных методов, как оптимизация внутриутробного и младенческого питания, устранение из окружающей среды свинца и других нейротоксичных загрязнений, уни-

чтожение паразитов; обеспечение полноценного сна и физической нагрузки; профилактика заболеваний, влияющих на умственную деятельность[33]. Безусловно, каждое средство из перечисленных помогает развитию когнитивных функций, хотя, скорее всего, значимость успеха слишком незначительна, особенно в обществах, где дети уже получают вполне качественное питание и образование. Действуя лишь таким образом, мы вряд ли сумеем развить в ком-нибудь мощный сверхинтеллект, но свою посильную лепту эти методы все-таки вносят, в частности, с их помощью мы улучшаем положение малоимущих и расширяем в мировом масштабе возможности для появления одаренных людей. (Распространенной проблемой многих бедных стран, не имеющих выхода к морю, остается снижающийся на протяжении жизни уровень интеллектуальных способностей из-за дефицита йода. Ситуация абсолютно неприемлемая, поскольку в наше время вопрос решается крайне просто — это обогащенная йодом столовая соль, которая по стоимости дороже обычной всего на несколько центов на одного человека в год[34].)

Большего эффекта удается добиться с помощью медико-биологических средств. В настоящее время появляются лекарственные препараты, способные, как утверждают, улучшать память, концентрацию внимания и умственные силы — по крайней мере, некоторым они помогают[35]. (Работая над этой книгой, я заправлялся кофе и поддерживал силы никотиновой жвачкой.) Действенность сегодняшних лекарств, стимулирующих умственные способности, весьма нестабильна, относительна и многими отрицается; вполне вероятно, будущие ноотропные средства, или нейрометаболические стимуляторы, начнут оказывать больше помощи и обладать меньшими побочными эффектами[36]. Однако вряд ли стоит уповать, что когда-нибудь придумают химический препарат, который, будучи введен в здоровый мозг, обеспечит резкий скачок интеллектуальных способностей человека, — как с неврологической, так и эволюционной точек зрения это кажется неправдоподобным[37]. Когнитивная деятельность мозга человека зависит от хрупкой гармонии многих факторов, особенно на критически важной стадии эмбрионального развития, поэтому для улучшения такой самоорганизующейся системы, как функциональность мозга, потребуется скорее не примитивная подкормка неким зельем,

а обеспечение ее бережным балансом, тонкой настройкой и тщательной культивацией.

Более мощные инструменты мы получим с помощью генетических манипуляций, не полагаясь на действие психотропных лекарственных средств. Вернемся снова к идее генетического отбора. Вместо попыток внедрять евгенические программы, корректируя схемы скрещивания, можно использовать механизмы клеточного отбора на уровне эмбрионов и гамет[38]. В ходе процедуры экстракорпорального оплодотворения (ЭКО) перед имплантацией эмбриона уже проводится генетическая диагностика, чтобы выявить моногенные нарушения вроде болезни Хантингтона и предрасположенность к некоторым развивающимся в более поздние периоды жизни человека заболеваниям, таким как рак молочной железы. Генетическую диагностику используют для определения пола будущего ребенка, а также для сравнения типа человеческих лейкоцитарных антигенов с данными его родных брата или сестры, по отношению к которым будущий новорожденный может выступить донором стволовых клеток, если они больны[39]. За следующие десять или двадцать лет заметно вырастет количество факторов, которые можно будет использовать в качестве критериев отбора — как позитивного, так и негативного. Значимым фактором прогресса генетики поведения являются быстро снижающиеся затраты на генотипирование и секвенирование генома. У нас на глазах появляется возможность проводить комплексный анализ характеристик всего генома, что значительно обогатит наши знания элементов генетической архитектуры, отвечающей за мыслительную и поведенческую деятельность человека[40]. Это даст возможность использовать в качестве критерия отбора любую черту личности, не относящуюся к наследственным, в том числе когнитивные способности[41]. Для отбора эмбрионов не требуется глубокого понимания причинно-следственных связей, которые в результате сложного взаимодействия между генами и окружающей средой приводят к тому или иному фенотипу, — необходимо лишь иметь генетические данные (правда, много), коррелирующие с интересующими исследователей признаками.

Можно сделать некоторые ориентировочные оценки коэффициентов максимального прироста, которые получаются при различных сценариях отбора[42].

Таблица 5. Максимальный прирост коэффициента умственного развития (IQ) в результате выбора из разного количества эмбрионов[43]

Отбор	**Дополнительные баллы IQ**
1 из 2	4,2
1 из 10	11,5
1 из 100	18,8
1 из 1000	24,3
5 поколений по 1 из 10	< 65 (каждое следующее поколение дает меньший прирост)
10 поколений по 1 из 1	< 130 (каждое следующее поколение дает меньший прирост)
Суммарный предел (с учетом сложения всех вариантов, оптимизированных с точки зрения когнитивных способностей)	100 + [< 300 (каждое следующее поколение дает меньший прирост)]

В табл. 5 показан ожидаемый рост интеллектуальных способностей в зависимости от размера популяции, в которой производится отбор, исходя из предположения, что доступна вся информация об общем количестве аддитивных генетических вариантов, лежащих в основе наследуемости интеллекта. (Неполная информация снизит эффективность селекции, хотя и не в той степени, как может показаться непосвященным[44].) Неудивительно, что отбор из большего числа эмбрионов дает лучшие результаты, хотя и не прямо пропорционально: выбор из ста эмбрионов не в пятьдесят раз предпочтительнее выбора из двух[45].

Интересно, что снижение прироста коэффициента умственного развития значительно меньше, когда результаты отбора отражаются на следующем поколении. Таким образом, гораздо лучший результат получается, если последовательно отбирать 1 из 10 на протяжении десяти поколений (когда каждое следующее поколение состоит из отобранных на предыдущем этапе), чем если один раз выбрать 1 из 100. Естественно, главная проблема последовательного отбора в том, что на него требуется больше времени. Если на каждый этап нужно двадцать–тридцать лет, тогда даже проект из пяти последовательных поколений закончится в середине XXII столетия. Скорее всего, к этому времени человечество достигнет успеха с помощью более прямых и мощных методов генной инженерии (не говоря уже об искусственном интеллекте).

Правда, появилась новая идея, которая сможет значительно увеличить благотворную роль генетического скрининга перед имплантацией, если будет исследована настолько, что будет применена к человеку, — это получение жизнеспособных сперматозоидов и яйцеклеток из стволовых клеток эмбриона[46]. С помощью этого метода уже было получено фертильное потомство мышей и человеческие гаметоподобные клетки. По сути, впереди еще много нерешенных научных проблем, и как минимум предстоит повторить полученные на мышах результаты, но уже на людях, избежав при этом эпигенетических отклонений в полученных линиях стволовых клеток. По мнению исследователя Кацухико Хаяси, решить эти задачи для клеток человека удастся «может быть, лет через десять, а может быть, через пятьдесят»[47].

С гаметами, полученными из стволовых клеток, у любой супружеской пары окажется гораздо больше возможностей для выбора. Сейчас при проведении ЭКО обычно создают меньше десяти эмбрионов. В случае получения гамет из стволовых клеток всего несколько клеток донора могут быть превращены в практически неограниченное число гамет, эмбрионы из которых будут подвергнуты генотипированию и секвенированию, чтобы выбрать наиболее многообещающие для имплантации. В зависимости от стоимости подготовки и скрининга одного эмбриона эта технология способна ощутимо увеличить селективные возможности, оказывающиеся в распоряжении родителей, которые выбрали процедуру ЭКО.

Но гораздо важнее другое: метод получения гамет из стволовых клеток позволит потратить на отбор из нескольких поколений гораздо меньше времени, чем требуется для созревания человека, поскольку предполагает использовать *итеративную селекцию эмбрионов*. Эта процедура состоит из определенных этапов[48].

1. Генотипирование и отбор эмбрионов, обладающих наилучшими необходимыми генетическими характеристиками.
2. Извлечение из этих эмбрионов стволовых клеток и превращение их в сперматозоид и яйцеклетку, созревающую в течение шести месяцев или даже менее[49].
3. Оплодотворение яйцеклетки сперматозоидом и получение новых эмбрионов.
4. Повторение этого цикла до накопления заметных генетических изменений.

Таким способом можно осуществить отбор из десяти и более поколений всего за несколько лет. (Это долгая и дорогая процедура, однако ее достаточно провести лишь один раз, а не повторять для каждого ребенка. Итоговую совокупность клеток можно будет использовать для получения очень большого количества улучшенных эмбрионов.)

Как видно из табл. 5, *средний* уровень интеллекта людей, родившихся в результате такого отбора, может быть очень высоким, возможно, равным или даже превосходящим уровень самых гениальных представителей человеческого рода. Мир, значительная часть населения которого состояла бы из людей такого интеллектуального развития, мог бы — при наличии соответствующей культуры, образования, коммуникационной инфраструктуры — представлять собой коллективный сверхразум.

Воздействие селекции эмбрионов на будущее все-таки может быть и ослаблено, и отсрочено. Существует биологически неизбежный временной лаг, связанный с развитием человека: пройдет как минимум двадцать лет, пока отобранные эмбрионы превратятся в людей, достигших производительного возраста; еще больше времени понадобится, чтобы они стали заметной частью своей социальной среды. Более того, даже если технология будет доведена до совершенства, готовность общества принять таких людей может быть очень низкой. В некоторых странах они вообще окажутся вне закона — на основании этических соображений или религиозных традиций[50]. Даже при возможности выбора многие пары все равно предпочтут естественный способ зачатия. Но приятие ЭКО постепенно начнет возрастать, если станут понятны преимущества этой процедуры, и прежде всего основная — фактическая гарантия, что ребенок окажется очень одаренным и лишенным генетической предрасположенности к болезням. В поддержку генетического отбора будут говорить невысокие затраты на медицинские манипуляции и ожидаемые в дальнейшей жизни высокие доходы. По мере того как процедура начнет пользоваться все большей популярностью, особенно среди элитарных слоев общества, может произойти культурный сдвиг в нормах воспитания — в результате выбор в пользу ЭКО станет свидетельством ответственного отношения людей к своим родительским обязанностям. В конечном счете даже скептики поддадутся моде, чтобы их дети не оказались в проигрышном положении по сравнению с «улучшенными» чадами их друзей и коллег. Некоторые страны могут ввести материальное стимулирование с целью побудить своих граждан пользоваться процедурой генетического отбора для повышения качества

«человеческого капитала» или укрепления долгосрочной социальной стабильности, используя в качестве критериев отбора такие черты личности, как покорность, готовность подчиняться, смирение, конформизм, несклонность к риску и малодушие, — естественно, культивирование таких популяций будет происходить за пределами правящих кланов.

Усиление интеллектуальных способностей человека будет также зависеть от степени отбора именно когнитивных признаков (см. табл. 6). Тем, кто решит воспользоваться процедурой отбора эмбрионов в той или иной форме, придется решать, как распределить имеющийся в их руках потенциал, поскольку интеллект в некоторой степени вступит в конкуренцию с другими не менее желанными свойствами: здоровьем, красотой, особой индивидуальностью и физической силой. Прийти к разумному компромиссу позволит итеративный характер отбора эмбрионов, с которым связаны значительные селективные возможности и благодаря которому будет осуществляться последовательный строгий отбор на основании нескольких критериев. Однако эта процедура приведет к разрушению прямой генетической связи между родителями и детьми, что может негативно сказаться на востребованности ЭКО во многих цивилизационных культурах[51].

Таблица 6. Возможное влияние генетического отбора при различных сценариях[52]

Принятие технологии	**«ЭКО+». Выбор 1 из 2 эмбрионов (4 балла)**	**«Агрессивное ЭКО». Выбор 1 из 10 эмбрионов (12 баллов)**	**«Яйцо из пробирки». Выбор 1 из 100 эмбрионов (19 баллов)**	**«Итерационный отбор эмбрионов» (100+ баллов)**
«В исключительно редком случае» Принятие — ~0,25%	Социально ничтожно, поскольку результат виден на уровне одного поколения. Неоднозначное общественное мнение берет верх над непредвзятой оценкой прямых последствий	Социально ничтожно, поскольку результат виден на уровне одного поколения. Неоднозначное общественное мнение берет верх над непредвзятой оценкой прямых последствий	«Усовершенствованные» люди формируют заметное меньшинство на позициях с высокими требованиями к когнитивным способностям	Люди из отобранных эмбрионов занимают доминирующие позиции в среде ученых, адвокатов, врачей, инженеров. Интеллектуальный ренессанс?

Принятие технологии	**«ЭКО+». Выбор 1 из 2 эмбрионов (4 балла)**	**«Агрессивное ЭКО». Выбор 1 из 10 эмбрионов (12 баллов)**	**«Яйцо из пробирки». Выбор 1 из 100 эмбрионов (19 баллов)**	**«Итерационный отбор эмбрионов» (100+ баллов)**
«Преимущественно элитарные слои» Принятие — 10%	Незначительное влияние. Небольшой когнитивный сдвиг в первом поколении в сочетании с отбором по признакам, не имеющим отношения к интеллекту, для формирования ощутимых преимуществ у меньшинства населения	Большая доля «улучшенных» студентов Гарварда. Второе поколение доминирует в профессиях с высокими требованиями к когнитивным способностям	Люди из отобранных эмбрионов занимают доминирующие позиции в среде ученых, адвокатов, врачей, инженеров уже в первом поколении	«Постчеловечество»[53]
«Новая норма» Принятие — > 90%	Неспособность к обучению встречается у детей гораздо реже. Во втором поколении количество людей с IQ выше среднего увеличилось в два раза	Значительный рост научных успехов, высокие доходы. Во втором поколении многократное увеличение числа людей с высоким IQ	В первом поколении IQ как у выдающихся ученых встречается в десятки раз чаще. Во втором поколении — в тысячи раз чаще	«Постчеловечество»

С дальнейшим развитием геномных технологий может появиться возможность синтезировать геномы в соответствии с заданной спецификацией, и тогда надобность в больших запасах эмбрионов отпадет. Сегодня еще невозможно синтезировать геном человека целиком и использовать его в репродуктивных целях — не в последнюю очередь из-за пока неразрешенных трудностей с правильным течением эпигенетических

процессов[54], — хотя синтез ДНК уже стал обычным направлением биотехнологий и почти полностью автоматизирован. Когда геномная технология достигнет высокого уровня, можно будет конструировать эмбрион с идеально точным соблюдением нужного сочетания генетических исходных данных обоих родителей. Появится возможность также добавить гены, отсутствующие у них, в том числе аллели, достаточно редко встречающиеся в популяции, но способные оказать заметный эффект на когнитивные способности ребенка[55].

После успешного синтеза человеческого генома одной из доступных операций станет генетическая диагностика эмбриона. (Приблизиться к этому способен также итерационный отбор эмбрионов.) В каждом из нас идут мутации, возможно, сотни мутаций, снижающих эффективность различных клеточных процессов[56]. Эффектом каждой отдельной мутации можно было бы пренебречь (и поэтому она так медленно удаляется из пула генов), но все вместе они могут серьезно влиять на нашу жизнеспособность[57]. Индивидуальные различия в интеллектуальных способностях могут быть в значительной степени следствием разницы в количестве и природе таких лишь слегка опасных аллелей, которые несет каждый из нас. В ходе синтеза гена мы можем взять геном эмбриона и сконструировать такую его версию, которая будет лишена генетического «шума» накопленных мутаций. Наверное, это прозвучит провокационно, но люди, созданные из таких проверенных геномов, могут оказаться более «настоящими», чем все живущие на планете сейчас, поскольку будут представлять собой менее искаженную версию человека. Не все они будут точными копиями друг друга, поскольку люди сильно отличаются генетически, даже если не брать в расчет вредоносные мутации. Но фенотипическим отражением освобожденного от нежелательных мутаций генома может быть исключительное физическое и психическое состояние человека, его превосходство в таких полигенных областях, как интеллект, состояние здоровья, смелость и внешность[58]. В качестве отдаленной аналогии приведу обобщенные портреты людей — так называемые усредненные лица, при составлении которых усредняются дефекты множества наложенных друг на друга лиц (см. рис. 6).

Рис. 6. Обобщенные портреты людей как метафора отредактированного генома. И женское, и мужское усредненное лицо получены путем наложения шестнадцати фотографий разных людей (жители Тель-Авива). Считается, что обобщенный потрет красивее любого из тех конкретных лиц, из которых он составлен, поскольку в нем усредняются характерные для его составляющих отклонения. По аналогии с этим в случае удаления индивидуальных мутаций в результате использования генетически диагностированных, то есть отредактированных, геномов могут появляться люди, близкие к идеалу Платона. При этом они не обязательно должны быть генетически идентичными, поскольку многие гены имеют целый набор в одинаковой мере функциональных аллелей. А проверка устранит лишь отклонения, возникшие в результате вредных мутаций[59].

Может оказаться востребованным такой метод биотехнологии, как клонирование. Когда-нибудь станет реальностью клонирование человека — почему бы тогда не использовать клоны для воспроизведения генома исключительно талантливых людей? Внедрение такого рода манипуляций окажется ограниченным из-за нежелания большинства потенциальных родителей терять генетическую связь с будущими детьми. Но, в принципе, не стоит пренебрегать этим средством, имеющим свои положительные стороны: во-первых, даже относительно небольшая тенденция к увеличению числа исключительно талантливых людей будет иметь довольно сильное влияние; во-вторых, вполне вероятно, что найдется страна, которая начнет осуществлять широкомасштабную евгеническую программу суррогатного материнства — разумеется, на платной основе. Со временем человек обратится к таким серьезным методам генной инженерии, как создание новых синтетических генов или включение в геном промоторов и других элементов с целью контроля экспрессии генов. Не исключено, что появятся совсем экзотические варианты: большой резервуар, наполненный сложно структурированной искусственно культивированной мозговой тканью; некие «преображенные» трансгенные существа (что-то вроде млекопитающих

с крупным мозгом, например киты или слоны, но наделенные человеческими генами). Конечно — вымысел в чистом виде, но кто может зарекаться?

До сих пор мы обсуждали вмешательства лишь на уровне зародышевой линии. Теоретически мы можем прийти к нужному результату гораздо быстрее: способом генной модификации соматических клеток — что позволит обойти цикл созревания поколения. С практической точки зрения такой путь намного сложнее, ведь потребуется вводить модифицированные гены в большое количество клеток живого организма, а если нашей целью является улучшение когнитивных функций мозга, то значит придется делать прямые инъекции в мозг. Тогда как при отборе имеющихся в нашем распоряжении половых клеток и эмбрионов генные инъекции не нужны. Даже такие методы генной терапии на уровне зародышевой линии, которые включают необходимость модификации генома (например, коррекция или соединение редких аллелей), гораздо легче задействовать на эмбриональной стадии, когда имеешь дело с небольшим количеством клеток. Кроме того, вмешательство на уровне эмбриона, возможно, приведет к лучшим результатам, поскольку влияние на мозг происходит на ранней стадии его формирования, в то время как при соматическом воздействии на взрослых особей придется ограничиться лишь корректировкой существующей структуры. (В некоторых случаях соматическая генная терапия вполне заменима медикаментозным лечением.)

Исходя из сказанного выше, нужно помнить, что при выборе такого метода, как вмешательство на уровне зародышевой линии, всегда следует учитывать временной фактор: годы, необходимые для взросления, неизбежно отодвигают значимость воздействия прихода в мир новой генерации[60]. Даже имей мы уже сегодня в своем распоряжении самую совершенную технологию, отвечающую требованиям исследователей, все равно потребовалось бы больше двух десятилетий, чтобы генетически модифицированное потомство достигло зрелости. Помимо всего, когда речь идет о новых методах, которые опробуют на людях, то между экспериментальной проверкой концепции в лабораторных условиях и началом применения метода в медицинской практике обычно проходит лет десять, в течение которых проводятся бесконечные исследования для подтверждения безопасности и масштабные клинические испытания. При простейших формах генетической селекции подобные проверки, скорее всего, не потребуются, поскольку используются стандартные методы лечения бесплодия и генетическая инфор-

мация для сознательного отбора эмбрионов, которые иначе были бы выбраны случайно.

Очевидно, в основе отсрочек могут лежать и внутренние обстоятельства, связанные не столько с боязнью ошибиться и навредить (вот откуда требования многочисленных проверок на безопасность), сколько со страхом перед успехом — страхом, вызванным опасением по поводу этической допустимости генетической селекции и ее широких социальных последствий (вот откуда потребность в разработке мер регулирования). В каждой развитой стране — в силу ее культурных, исторических и религиозных особенностей — это беспокойство выражается по-своему. После Второй мировой войны в Германии предпочитают избегать любых репродуктивных методов, хотя бы в отдаленной степени напоминающих попытку улучшения человеческой природы, — позиция более чем понятная, если учитывать мрачную историю преступлений, совершенных нацистами во имя евгеники. В остальных западных странах, вероятно, будут смотреть на вещи шире. Некоторые государства — скорее всего, Китай или Сингапур, где уже действует долгосрочная демографическая политика, — ради повышения интеллектуального уровня своего населения могут не только разрешить, но и активно продвигать использование генетической селекции и генной инженерии, когда развитие технологий сделает это возможным.

Как только будет создан прецедент и станут видны реальные результаты, сразу у всех, кто хотел, но откладывал решение проблемы, появится мощный стимул последовать примеру первопроходцев. Страны, предпочитающие держаться в стороне, обязательно столкнутся с перспективой навсегда застрять в интеллектуальном болоте, утратить экономические, научные и военные позиции и навсегда уступить свое влияние в мире государствам, не побоявшимся новых технологий совершенствования человеческих возможностей. Население начнет задумываться, почему в престижных учебных заведениях учатся только генетически отобранные дети (которые в среднем будут еще отличаться и внешней привлекательностью, и здоровьем, и усидчивостью); естественно, граждане пожелают, чтобы их будущие отпрыски тоже могли пользоваться такими же преимуществами. Есть вероятность, что после того как заработает генная инженерия и будут подтверждены ее первые результаты, в течение сравнительно короткого времени — может быть, десятилетия — произойдет серьезный поведенческий сдвиг. Проведенные в США опросы показывают значительные изменения в общественном

мнении по отношению к процедуре ЭКО с момента появления в 1978 году Луизы Браун — первого «младенца из пробирки». За несколько лет до этого всего 18 процентов американцев согласились бы сделать ЭКО в случае бесплодия; вскоре после рождения Луизы Браун согласных насчитывалось уже 53 процента, и их число продолжает расти[61]. (Для сравнения: в проведенном в 2004 году опросе 28 процентов американцев одобрили селекцию эмбрионов по критерию «сила и интеллект», 58 процентов — по критерию «избежать риска развития рака во взрослом возрасте», 68 процентов — по критерию «избежать риска неизлечимых детских болезней»[62].)

Давайте еще раз перечислим случаи, вызывающие отсрочку результатов: сбор информации, необходимой для успешной селекции из набора эмбрионов, полученных в результате процедуры ЭКО, — *от пяти до десяти лет* (возможно, потребуется значительно больше времени, чтобы гаметы из стволовых клеток стали доступны для использования в процессе репродукции человека); формирование социально значимого спроса и внедрение самой услуги — *десять лет*; время, которое потребуется «улучшенному» поколению, чтобы достичь производительного возраста, — *от двадцати до двадцати пяти лет*. Суммируя все сроки, мы увидим, что технологии по улучшению человеческих свойств на уровне зародышевой линии вряд ли начнут оказывать существенное влияние на социальную среду в первой половине текущего столетия. Однако вследствие применения генетических методов уже с середины столетия в довольно большом сегменте общества будет отмечен показательный подъем интеллектуальных способностей взрослого населения. После того как в ряды трудоспособного населения вольются когорты людей, чье зачатие было осуществлено по ультрасовременным высоким генетическим технологиям — как, например, применение эмбриональных стволовых клеток и итеративной селекции эмбрионов, — темпы интеллектуального роста намного повысятся.

Когда описанные выше генетические технологии достигнут своего полного развития (оставим пока за скобками экзотические варианты вроде интеллекта в искусственно культивированной ткани мозга), мир убедится, что представители новых поколений в среднем окажутся несравненно умнее людей из прошлого — даже обладателей наивысших коэффициентов интеллекта. Потенциал биологического совершенствования в перспективе так высок, что, возможно, его вполне хватит для появления человека сверхразумного — по крайней мере в его начальной стадии. В этом нет ничего

удивительного. В конечном счете именно так возник человек разумный: когда у определенного вида человекообразных резко повысились, по сравнению с прародителями-гоминидами, интеллектуальные способности — причем их развитие произошло в результате такого слепого и неконтролируемого метода, как эволюционный процесс. Поэтому нет оснований предполагать, будто *Homo sapiens*, дойдя якобы до вершины разумной деятельности, является максимальным достижением биологической системы. Мы далеки от того, чтобы представлять собой самый умный биологический вид, возможно, задуманный природой. Вероятно, нас лучше рассматривать как самый глупый биологический вид из умников, возможно, задуманных природой, но вид, способный создать и привести в действие технологическую цивилизацию — ту нишу, которую мы заняли вовсе не из-за своей, как принято считать, оптимальной адаптивности, а лишь потому, что добрались до нее первыми.

Прогресс на пути биологического развития вполне реален. Но из-за неизбежной отсрочки на время взросления целого поколения он не получится столь же внезапным, как в сценариях с созданием искусственного интеллекта. (Временной фактор вряд ли будет играть столь существенную роль как в случае применения генной терапии соматических клеток, так и при медикаментозном подходе, но эти методы с меньшей вероятностью способны вызвать заметные изменения.) *Максимальный потенциал* искусственного интеллекта, безусловно, намного выше природного интеллекта, присущего человеку. (Величину разрыва можно оценить, сравнив разницу в быстродействии между электронными компонентами и нервными клетками: сегодняшние транзисторы работают в десять миллионов раз быстрее, чем биологические нейроны.) Однако даже сравнительно незначительные улучшения биологического интеллекта могли бы иметь серьезные последствия. В частности, это форсировало бы научно-технологическое развитие, что, в свою очередь, способствовало бы успехам на пути освоения более действенных методов как совершенствования биологических умственных способностей, так и разработки искусственного интеллекта. Задумайтесь, какими темпами мы продвигались бы к созданию искусственного разума, если бы мир населяли миллионы людей, превосходящих по своему интеллектуальному уровню любых выдающихся мыслителей прошлого, а самый заурядный парень на земле ни в чем бы не уступал Алану Тьюрингу вместе с Джоном фон Нейманом[63].

На какое-то время отойдем от обсуждения стратегических последствий развития когнитивных способностей и постараемся подвести итоги сказанному, отметив три важных момента:

1) при помощи биотехнологических методов мы способны прийти к существованию сверхразума, по крайней мере к его начальной стадии;
2) появление усовершенствованных интеллектуально людей увеличивает возможность осуществить когда-нибудь развитие искусственного интеллекта до высокоразвитых форм, поскольку сама задача создания ИИ будет абсолютно доступна и проста для усовершенствованных людей нового поколения — при условии, конечно, что мы окажемся принципиально неспособными справиться с нею собственными силами (хотя предполагать подобное пока нет никаких причин);
3) мы рассматривали сценарии, обещающие завершиться не ранее чем во второй половине нынешнего столетия, а может быть, и позже; однако, уносясь мыслью в такую даль, нам следует учитывать, что вполне допустимо появление поколения генетически усовершенствованных групп людей: избирателей, изобретателей, ученых, причем показатели улучшения их когнитивных функций будут увеличиваться от десятилетия к десятилетию.

Нейрокомпьютерный интерфейс

Периодически выдвигаются предложения использовать прямой нейрокомпьютерный интерфейс, в частности, имплантаты, что позволит человеку использовать всю мощь электронных вычислений: идеальное хранение информации, быстрые и точные арифметические расчеты, широкополосную передачу данных — в результате такая гибридная система будет принципиально превосходить по всем характеристикам деятельность головного мозга[64]. Возможность прямого подключения компьютера к биологическому мозгу была не раз доказана, но, несмотря на это, кажется маловероятным, что прямые нейронные интерфейсы получат в обозримом будущем широкое распространение[65].

Прежде всего заметим, что в результате имплантации электрода в мозг возникает значительный риск медицинских осложнений — инфекции, смещение электрода, кровоизлияния, ухудшение умственных способностей. На сегодняшний день лечение пациентов с болезнью Паркинсона является

едва ли не самой яркой демонстрацией той пользы, которую приносит стимуляция мозга. В этом случае используется довольно простой имплантат, на самом деле не соединенный непосредственно с мозгом, а всего лишь создающий электрический разряд, воздействующий на субталамическое ядро, или ядро Льюиса. На демонстрационном видеоролике показан сидящий в кресле полностью обездвиженный болезнью человек, который после подключения электрода мгновенно возвращается к жизни: он начинает двигать руками, встает и идет по комнате, поворачивается на месте и даже делает пируэт. Но у этой совершенно простой и на удивление успешной процедуры тоже есть негативные стороны. В одном исследовании у экспериментальных пациентов с болезнью Паркинсона, по сравнению с контрольной группой, при имплантации электрода в мозг отмечены ухудшения следующих функций: беглой речи, избирательного внимания, цветовой и словесной памяти. Испытуемые пациенты часто жаловались на снижение умственных способностей[66]. Если речь идет о людях с тяжелыми заболеваниями, то можно мириться и с рисками, и с побочными эффектами. Совсем другой вопрос — здоровые граждане, соглашающиеся на нейрохирургические манипуляции. В таких случаях любое вмешательство должно приводить к существенному улучшению функций головного мозга.

Пожалуй, такое усовершенствование когнитивных способностей обернется более сложным делом, чем генная терапия, — это тоже дает право сомневаться, что путь киборгизации приведет нас к сверхразуму. Пациенты, страдающие параличом, могут получить пользу от имплантата, который заменит их пораженные нервы или активирует спинномозговые центры, отвечающие за двигательную функцию[67]. Пациенты, испытывающие проблемы со зрением или слухом, безусловно, выигрывают от имплантации искусственной улитки или сетчатки глаза[68]. Пациенты с болезнью Паркинсона или хронической мышечной болью, без сомнения, испытывают облегчение от глубокой стимуляции мозга, возбуждающей или подавляющей активность в отдельных его областях[69]. Гораздо более трудная задача — обеспечить непосредственное широкополосное взаимодействие между мозгом и компьютером для заметного повышения интеллектуальных способностей, которого невозможно добиться иными, более доступными средствами. Большинство потенциальных преимуществ, которые появятся в распоряжении здоровых людей в результате имплантации электродов, возможно получить с меньшим риском, затратами и неудобствами, просто используя

обычные органы движения и чувств при взаимодействии с компьютерами, находящимися вне пределов нашего тела. Чтобы выйти в интернет, нам не нужно подключать к себе оптоволоконный кабель. Человек не только наделен сетчаткой глаза, способной передавать данные с впечатляющей скоростью около десяти миллионов бит в секунду, но и обладает «предустановленным программным обеспечением» в виде зрительной коры головного мозга, которая отлично приспособлена для извлечения значения из этих массивов информации и взаимодействия с другими областями мозга для ее дальнейшей обработки[70]. Даже если появился бы относительно простой способ закачивать в наш мозг больше информации, эти дополнительные данные ненамного повысили бы скорость, с которой мы думаем и учимся, если только «апгрейду» не подвергнется весь нейронный механизм их обработки. А поскольку он включает в себя практически весь мозг, в действительности потребовалось бы «протезирование» мозга целиком — иначе говоря, создание универсального искусственного интеллекта. Впрочем, существуй искусственный интеллект человеческого уровня — зачем тогда понадобилась бы нейрохирургия? Ведь компьютер может быть помещен не только в костяную коробку, но и в металлический корпус. Таким образом, если мы вновь обращаемся к искусственному интеллекту, то непременно свернем на путь, уже рассмотренный нами ранее.

Ученые предлагают использовать нейрокомпьютерный интерфейс для считывания информации из головного мозга человека для коммуникации его с другими людьми или компьютерами[71]. Система, позволяющая передвигать курсор на экране с помощью мысли, помогла бы пациентам с синдромом «запертого человека»* устанавливать связь с внешним миром[72]. Ширина полосы передачи данных в таких экспериментах пока очень мала: пациент мучительно долго набирает букву за буквой со скоростью несколько слов в минуту. Можно легко представить усовершенствованную версию, по всей вероятности, с имплантами следующего поколения, которые — для трансляции внутренней речи — будут вживлять в центр Брока

* *Синдром «запертого человека»* (синдром изоляции, синдром деэфферентации) — синдром, который характеризуется отсутствием адекватной реакции больного на внешние, в том числе и словесные, стимулы из-за тетраплегии и паралича бульбарной, мимической и жевательной мускулатуры. Причинами развития синдрома могут быть инфаркт основания мозга, кровоизлияние в мозг, полиомиелит, некоторые яды.

(участок коры головного мозга, находящийся в задненижней части третьей лобной извилины, отвечающий за моторную, фонологическую и синтаксическую организацию речи)[73]. Сегодня системы обратной связи интересны скорее с точки зрения оказания помощи пациентам с мышечной атрофией и людям, перенесшим инсульт. Эта новейшая технология пока мало применима к здоровому человеку, хотя, по сути, повторяет тот же набор функций, который обеспечивается простым наличием микрофона и программой распознавания речи, то есть продуктом, уже присутствующим на нашем рынке и отличающимся в лучшую сторону такими своими характеристиками, как неболезненное и удобное применение, дешевизна и отсутствие риска, связанного с нейрохирургическим вмешательством (а также не порождающим фантазий в духе Оруэлла на тему подслушивающего устройства внутри черепной коробки). Кроме того, когда наше тело и компьютер никак не связаны физически, то последний удобнее ремонтировать и оснащать новым ПО.

Но как быть с неизбывной человеческой мечтой, чтобы люди вступали в общение не на вербальном уровне, а напрямую — через мозговую деятельность, как бы «загружая» друг в друга свои образы, мысли, знания и даже опыт? Мы загружаем в компьютеры огромные файлы, в том числе библиотеки с миллионами книг и статей, буквально за считаные секунды или минуты — неужели нам никогда не придется поступать так же, имея дело с собственным мозгом и собственной информацией? Кажущаяся легкость реализации этой идеи, вероятно, базируется на ошибочном представлении о том, как человеческий мозг воспринимает и хранит информацию. Как уже отмечалось, развитие человеческого интеллекта ограничивает не скорость, с которой данные поступают в память, а насколько быстро мозг способен извлекать из них смысловые значения и осознавать их. Возможно, предполагается передавать непосредственно смысл, не оформляя его в сенсорную информацию, которую придется декодировать получателю. Тут возникает две проблемы. Первая заключается в том, что мозг, в отличие от программ, которые мы привычно используем на компьютерах, не использует стандартные форматы хранения и представления данных. Скорее, в каждом мозгу имеются свои уникальные способы представления содержания более высокого уровня. То, какие именно сочетания нейронов используются для передачи той или иной концепции, зависит от уникального опыта конкретного мозга (а также различных генетических факторов и стохастических

физиологических процессов). Как в случае искусственных нейронных сетей, так и в биологических нейронных сетях смысловое значение скорее представлено всей структурой и моделями деятельности значительных перекрывающихся регионов, а не отдельными ячейками памяти, уложенными в аккуратные массивы[74]. Поэтому невозможно установить простое соответствие между нейронами двух людей так, чтобы мысли автоматически перетекали от одного к другому. Если нужно передать мысли из одного мозга в другой так, чтобы они были ему понятны, их нужно подвергнуть декомпозиции и перевести в символы в соответствии с некоторой общепринятой системой, которая позволит их правильно интерпретировать мозгом-приемником. Это уже лингвистическая задача.

Теоретически мы в состоянии представить интерфейс, на который было бы можно переложить когнитивную работу по артикуляции и интерпретации мыслей. Он будет должен уметь каким-то образом считывать состояния нейронов в мозге-передатчике и переводить их в понятные модели активации нейронов в мозге-приемнике. Даже если оставить в стороне (очевидные) технические трудности организации надежного одновременного считывания состояния миллиардов отдельных нейронов и записи в них, создание такого интерфейса, вероятно, само по себе является AI-полной задачей искусственного интеллекта. Интерфейс должен включать компонент, способный (в режиме реального времени) ставить в соответствие возникающим в одном мозгу моделям семантически эквивалентные модели в другом мозгу. Для выполнения этой задачи потребуется подробное многоуровневое понимание механизма нейронных вычислений, которое может привести непосредственно к созданию нейроморфного ИИ.

Несмотря на эти оговорки, движение в сторону улучшения интеллектуальных способностей по пути создания киберорганизмов не кажется совершенно бесперспективным. Впечатляющие результаты работ с гиппокампом крыс показали возможность создания нейронного протеза, который может повысить эффективность выполнения простой задачи на запоминание[75]. На сегодняшний день имплантат считывает информацию с электродов в количестве от одного десятка до двух десятков, размещенных в области CA3 гиппокампа, и передает ее на такое же количество нейронов, расположенных в области CA1 гиппокампа. Микропроцессор способен различать две модели возбуждения в первой области (соответствующие двум видам информации — «правый рычаг» и «левый рычаг») и научиться тому, как эти модели передаются

во вторую область. Такие протезы могут не только восстановить функционирование мозга в ситуации, когда нормальное нейронное взаимодействие между двумя областями нейронов нарушено, но и за счет направленной активации требуемой модели во второй области способны повысить эффективность выполнения задачи по сравнению с обычным для крыс уровнем. Хотя по современным стандартам это и весьма впечатляющее в техническом плане достижение, эксперимент оставляет без ответа множество вопросов. Насколько хорошо этот подход масштабируется? Ведь число комбинаций взаимодействующих областей мозга, а также нейронов на входе и выходе из них, очень велико, поэтому сможем ли мы избежать комбинаторного взрыва при попытке картировать взаимодействия в мозгу? Не получится ли, что хотя эффективность решения тестовой задачи растет, этому сопутствуют некие скрытые издержки, например снижение способности обобщать стимулы или неспособность забыть определенную ассоциацию, после того как среда изменилась? Получит ли человек — располагающий, в отличие от крыс, внешними носителями памяти вроде бумаги и ручки — какую-либо выгоду от появления таких возможностей? Насколько легко будет применить подобный метод к другим областям мозга? В то время как работе описанного протеза помогает сравнительно простая структура областей гиппокампа, обеспечивающая последовательную передачу сигнала в одну сторону (по сути, однонаправленная связь между зонами CA3 и CA1), другие структуры в коре головного мозга используют рекуррентные циклы обратной связи, что значительно повышает сложность схемы связей и, видимо, затруднит расшифровку набора функций встроенных в нее групп нейронов.

В плане развития киборгов есть надежда, что мозг, снабженный имплантатом, поддерживающим связь с внешней средой, со временем *научится* сопоставлять свое внутреннее состояние и получаемые внешние сигналы. В этом случае имплантату не обязательно обладать интеллектом, скорее, мозг должен будет интеллектуально настроиться на интерфейс, примерно как мозг ребенка постепенно обучается интерпретировать сигналы, поступающие из внешнего мира через рецепторы органов зрения и слуха[76]. И снова возникает естественный вопрос: принесет ли это какую-нибудь реальную пользу? Предположим, пластичность мозга окажется настолько достаточной, что он научится распознавать модели в рамках некоего нового потока входных сигналов, проецируемых на его кору посредством некоего нейрокомпьютерного интерфейса, — но почему тогда просто не спроецировать ту же

самую информацию непосредственно на сетчатку глаза в виде зрительных образов или на улитку в виде звука? Применение низкотехнологичных методов поможет избежать множества проблем — хотя и в том и в другом случаях нашему мозгу, чтобы научиться понимать информацию, придется задействовать механизмы распознавания образов и присущее ему свойство пластичности.

Сети и организации

Еще один потенциальный путь, ведущий к сверхразуму, — постепенное совершенствование сетей и организаций, соединяющих умы людей друг с другом и с различными искусственными объектами и ботами, то есть программами, автоматически выполняющими действия вместо человека. Смысл не в том, чтобы усовершенствовать когнитивные способности отдельных людей и в итоге вывести популяцию сверинтеллектуалов. Идея заключается в другом: создать некое объединение индивидуумов, организованных таким образом, чтобы эта появившаяся сеть по своему развитию могла бы достигнуть сверхинтеллектуального уровня — сеть, которую в следующей главе мы назовем «коллективный сверхразум»[77].

В доисторические и исторические времена коллективный интеллект помог человечеству добиться многого. Источники успеха были самые разные: нововведения в средствах связи — причем сюда надо включить изобретение письменности и печатного дела, не говоря уже о возникновении самих языков; рост населения и увеличение его плотности; усовершенствование форм институциональной организации и стандартов познания; постепенное накопление институционального капитала. Фактически система коллективного интеллекта ограничена возможностями интеллекта ее членов, затратами на передачу информации между ними и различными недостатками и неэффективностью, присущими любым человеческим сообществам. По мере снижения расходов на все виды связи (имеется в виду не только стоимость оборудования, но и время ожидания ответа, затраты времени и внимания, а также другие факторы) появляется возможность создавать более крупные и более сплоченные организации. То же самое происходит и в случае успешной борьбы с отдельных ведомственными крайностями, деформирующими любую организационную жизнь, — разорительные имиджевые игры и статусные притязания; распыление ресурсов; несоблюдение сроков выполнения заданий; сокрытие фактов; фальсификация информации и прочие проблемы,

связанные с выбором между свободой воли и навязанными условиями. Даже частичная ликвидация перекосов приносит коллективному интеллекту внушительную пользу.

Существует множество технологических и институциональных новаторских идей, способных влиять на рост нашего коллективного интеллекта. Например, современные рынки прогнозов относительно политики распределения дотаций благоприятствуют утверждению норм справедливости и способствуют выработке перспективных оценок по спорным научным и социальным вопросам[78]. Детекторы лжи (если удастся наладить выпуск надежных и удобных в применении полиграфов) смогут понизить уровень мошенничества в деятельности людей[79]. Более мощным инструментом могут стать детекторы самообмана[80]. Но и без новоиспеченных игр разума некоторые формы обмана перестают быть актуальными, утрачивая свою привлекательность из-за ряда причин, таких как: доступность информации, рассказывающей о репутации и прошлом человека; промульгация строгих гносеологических правил*; приоритет здравого смысла в культуре организаций. В результате систем наблюдения, осуществляемых на добровольной или обязательной основе, будут накоплены огромные объемы информации о поведении человека. На сайтах социальных сетей делятся своей личной информацией уже больше миллиарда людей; совсем скоро все пользователи — с помощью микрофонов и видеокамер, встроенных в смартфоны или оправы очков, — получат возможность загружать непрерывную трансляцию своей жизни. Автоматизированный анализ этих потоков данных породит множество новых применений — разумеется, как во благо, так и во зло[81].

Рост уровня коллективного интеллекта может быть также связан с общими организационными и экономическими изменениями и с увеличением среди народонаселения доли больших сообществ социальных сетей, состоящих из образованных людей, постоянно обменивающихся информацией и интегрированных в общемировую культуру[82].

Интернет остается самым динамичным полем действия, передним краем для инноваций и экспериментов. Причем большая часть его потенциала

---

* *Промульгация* (promulgation) — официальное провозглашение закона, принятого парламентом или главой государства. *Гносеологические правила* — адекватное отражение объективной действительности; понятие, относящееся к нормативным правовым актам.

до сих пор еще не раскрыта. Следует укреплять интеллектуальные сети, активно поддерживать формат разумных обсуждений, стараться избегать предубеждений, вырабатывать механизмы для превращения частных суждений в коллективные решения — все это должно внести существенный вклад в развитие коллективного интеллекта как всего человечества в целом, так и отдельных сообществ.

Настало время поговорить о совершенно, казалось бы, фантастической идее, что интернет может в один прекрасный день «проснуться». Может ли он стать чем-то бóльшим, нежели просто местом сосредоточения пока еще слабо выраженного коллективного сверхразумного начала — чем-то вроде виртуальной черепной коробки, вместившей в себя зародыш единого сверхразума? (В знаменитом эссе Вернона Винджа «Далее — технологическая сингулярность», написанном в 1993 году, этот сценарий рассматривается в качестве одного из путей появления сверхразума, писатель даже ввел в оборот термин «технологическая сингулярность»[83].) Можно возразить, что искусственный интеллект трудно создать даже в результате целенаправленных инженерных усилий, поэтому его *спонтанное* появление кажется практически невероятным. Однако дело не обстоит так, будто одна из следующих версий интернета внезапно станет сверхразумной исключительно по воле случая. Более правдоподобный сценарий заключается в другом: интернет будет шаг за шагом совершенствоваться, аккумулируя в себе все самое передовое, благодаря усилиям множества людей на протяжении долгих лет — усилиям, направленным на улучшение алгоритмов поиска, отбора и анализа информации, на создание более мощных форматов представления данных, более качественных автономных ПО и более эффективных протоколов, управляющих взаимодействием этих ботов. В конечном счете мириады небольших сдвигов создадут основу для некой единой формы сетевого интеллекта. По крайней мере, вполне возможно, что появится именно такая когнитивная система, выращенная на веб-технологиях, не испытывающая недостатка в вычислительной мощности и других ресурсах, необходимых для взрывного роста, — разве что за исключением одного критически важного ингредиента. И когда этот ингредиент будет найден и брошен в общий котел — все раньше сваренное воспламенится и превратится в сверхразум. Однако этот сценарий опять сворачивает на уже знакомый нам путь появления сверхразума — создание универсального искусственного интеллекта.

Резюме

Итак, к сверхразуму ведут самые разные пути, и этот непреложный факт вселяет некоторую уверенность, что в конечном счете мы до него доберемся. Не удастся пройти одним путем — мы выберем другой.

Однако разнообразные варианты не приведут нас во многие места назначения. Даже если на одной из дорог, не связанной с машинным интеллектом, произойдет заметное улучшение когнитивных способностей — это не означает, что ИИ утратил свое значение. Скорее, наоборот: развившийся сверх меры человеческий и организационный разум ускорит развитие науки и технологий, потенциально приблизив появление радикальных форм создания универсального искусственного интеллекта вроде полной эмуляции головного мозга.

Из всего вышесказанного не следует делать вывод, будто нам все равно, каким маршрутом двигаться к сверхразуму. Выбранный путь может оказать серьезное влияние на конечный результат. Даже если новые полученные возможности не слишком обусловлены вариантом направления, то вопрос, как они станут использоваться и какова будет степень нашего контроля над ними, вполне может зависеть от принятого подхода. Например, усовершенствование человеческого или организационного разума может повысить готовность людей идти на риск и добиваться осуществления такого машинного сверхразума, который будет безопасным и полезным для человечества. (Чтобы дойти до полноценной стратегической оценки этого, придется преодолеть много трудностей — их обсуждением мы займемся лишь в четырнадцатой главе.)

Можно ожидать, что первый настоящий сверхразум (в отличие от незначительного повышения нынешнего уровня когнитивных способностей) появится в результате движения к искусственному интеллекту. Однако этот путь связан с большой неопределенностью. Поэтому трудно точно оценить, насколько он окажется долгим и со сколькими препятствиями мы столкнемся. Некоторыми шансами оказаться самым быстрым способом осуществления сверхразума обладает полная эмуляция головного мозга. Поскольку прогресс на этом пути требует скорее технологических решений, чем теоретических прорывов, есть основания полагать, что в конечном счете успех достижим. И все-таки с большой долей уверенности мы утверждаем, что даже в случае постоянного прогресса в компьютерном моделировании мозга финишную черту первым пересечет искусственный интеллект: причина заключается

в том, что нейроморфный искусственный интеллект может быть создан и с помощью частичной эмуляции мозга.

Явно решаема задача биологического улучшения интеллектуальных способностей, особенно основанного на генетической селекции. Многообещающей технологией на сегодняшний день кажется итеративная селекция эмбрионов. Однако в сравнении с возможными прорывами в искусственном интеллекте биологические улучшения будут происходить относительно медленно и постепенно. В лучшем случае они приведут к возникновению сравнительно слабой формы сверхразума (скоро мы снова вернемся к этой теме).

Благодаря реальной возможности биологического улучшения интеллектуальных способностей растет наша уверенность, что в конце концов будет создан и искусственный интеллект, поскольку улучшенные интеллектуально люди — ученые и инженеры — смогут добиться большего и быстрейшего прогресса, нежели их *обычные* коллеги. Особенно в тех сценариях, где ИИ должен быть создан не раньше середины нашего столетия, огромная роль отводится постепенно растущей когорте усовершенствованных интеллектуально людей.

Нейрокомпьютерные интерфейсы вряд ли станут тем вариантом, который приведет нас к сверхразуму. Усовершенствование сетей и организаций может в долгосрочной перспективе привести к появлению слабых форм коллективного интеллекта, но более вероятно, что оно сыграет стимулирующую роль, как и биологическое улучшение интеллектуальных способностей, постепенно повышая эффективность умственной деятельности людей при решении интеллектуальных задач. В сравнении с биологическими улучшениями прогресс в развитии сетей и организаций произойдет быстрее — на самом деле он уже происходит и уже оказывает на нашу жизнь значительное влияние. Однако усовершенствование сетей и организаций будет иметь меньшее влияние на развитие человеческих возможностей решать интеллектуальные задачи, чем усовершенствование когнитивных способностей. Сети и организации скорее послужат стимулирующим началом развития *коллективного интеллекта*, нежели *качественного интеллекта* — разницу между этими понятиями мы рассмотрим в следующей главе.

Глава третья
Типы сверхразума

Итак, что именно мы подразумеваем под словом *сверхразум*? Не хотелось бы погружаться в терминологическую трясину, но что-то сказать для прояснения понятийной основы все-таки нужно. В этой главе мы проведем идентификацию трех типов сверхразума и убедимся, что в практическом смысле все три тождественны. Затем мы продемонстрируем, насколько потенциал биологического интеллекта проигрывает потенциалу машинного. У машин есть множество фундаментальных преимуществ, которые обеспечивают им подавляющее превосходство над человеком. Биологический мозг, даже улучшенный, не сможет с ними конкурировать.

Есть и машины, есть и животные, превосходящие человека в том или ином виде деятельности, — можно сказать, они уже достигли сверхчеловеческого уровня. Летучая мышь, используя при полете эхолокацию, ориентируется в темноте лучше человека, калькулятор обходит его в арифметических расчетах, шахматные программы обыгрывают в шахматы. Перечень специфических задач, с которыми программы специального назначения справляются лучше нас, будет только расти. Безусловно, специализированные информационные системы могут иметь множество применений, но все равно возникают дополнительные серьезные проблемы, когда речь заходит о перспективе создания универсального искусственного интеллекта, способного занять место человека без исключения на всех направлениях.

Уже не раз упоминалось, что термин *сверхразум* мы используем для обозначения такого интеллекта, который во многих универсальных проявлениях когнитивной деятельности в значительной степени превосходит лучшие умы человечества. Описание весьма расплывчатое (поэтому формулировка и не повторяет дословно ту, что дана нами в предыдущей главе). В соответствии с этим определением на статус сверхразума начнут претендовать самые разные системы с совершенно несопоставимыми функциональными

свойствами. Разобраться с имеющимися дефинициями этого довольно простого понятия нам позволит детальный анализ сверхвозможностей мозга, для чего придется распутать далеко не однородный узел, вытягивая из него различительные признаки интеллекта. Есть множество способов провести такую декомпозицию*. Для дифференциации воплощений сверхразума обратимся к трем его типам: скоростной сверхразум, коллективный сверхразум и качественный сверхразум.

Скоростной сверхразум

Скоростной сверхразум представляет собой такой же интеллект, как человеческий, только более быстрый. С концептуальной точки зрения данный тип самый простой для анализа[1]. Мы дадим ему следующее определение:

> *скоростной сверхразум* — система, способная делать все то же, что и человеческий интеллект, только намного быстрее.

Под «намного» имеется в виду «на несколько порядков». Понимаю, что и это крайне обобщенное определение явно хромает, но я благоразумно отойду в сторону и предоставлю самому читателю разбираться с его интерпретацией[2].

Самым простым примером скоростного сверхразума могла бы стать полная эмуляция головного мозга, выполненная на сверхмощном оборудовании[3]. Имитационная модель мозга, работающая со скоростью, в десять раз превышающей скорость биологического мозга, смогла бы читать книги за считаные секунды, а докторскую диссертацию написать за день. Если скорость имитационной модели будет выше в миллион раз, она будет в состоянии выполнять за день интеллектуальную работу, на которую у человека ушло бы целое тысячелетие[4].

Для разума, работающего с такой скоростью, происходящие во внешнем мире события походили бы на замедленную съемку. Представьте, что ваш мозг ускорился в десять тысяч раз. Друг роняет чашку, и вы в течение нескольких часов наблюдаете ее медленное движение в сторону пола — она

---

* *Декомпозиция* (decomposition) — научный метод, заключающийся в разделении целого на части и позволяющий рассматривать любую исследуемую систему на основании ее признаков как сложную, состоящую из отдельных взаимосвязанных подсистем, которые, в свою очередь, могут быть расчленены на части; применяется при анализе материальных объектов, процессов, явлений и понятий.

словно комета, безмолвно скользящая в космосе навстречу далекой планете, — и по мере того как ощущение неизбежной катастрофы мучительно пробивается через извилины серого вещества вашего приятеля, а оттуда в его периферийную нервную систему, на его лице постепенно проступает выражение, предшествующее возгласу «ой!», который вы еще не скоро услышите. Короче говоря, за это время вы успеете принести ему новую чашку, заодно прочитать пару научных статей и даже вздремнуть.

По причине подобного растяжения истинного времени скоростной сверхразум, наверное, предпочел бы работать с цифровыми объектами, а не объектами материального мира. Ему удобнее было бы существовать в виртуальной реальности и иметь дело с информационной экономикой, а при необходимости — вступать во взаимодействие с физической средой при помощи наноманипуляторов, поскольку эти микроскопические конечности могут двигаться быстрее роботизированных. (Частотные характеристики системы обычно обратно пропорциональны ее линейным размерам[5].) Скоростной ум мог бы взаимодействовать главным образом с другими скоростными умами, а не с людьми, чье брадителически* медлительное передвижение в пространстве сравнимо разве что с тягучестью меда.

По мере ускорения интеллекта все более критическим ограничителем становится скорость света, поскольку растут издержки в результате потери времени на путешествие или передачу информации на дальние расстояния[6]. Свет примерно в миллион раз быстрее реактивного самолета, поэтому цифровому агенту со скоростью мышления, в миллион раз превышающей человеческую, потребуется примерно столько же его субъективного времени на путешествие вокруг света, как и его современнику-человеку. Звонок кому-то, находящемуся в другом городе, займет столько же времени, сколько нужно на то, что бы этот кто-то оказался перед вами собственной персоной. Сверхразумам, достигшим высочайшей скорости, которым требуется постоянное интенсивное сотрудничество, предпочтительнее находиться недалеко друг от друга. Если скоростные сверхразумы, например, собираются работать над одной задачей, было бы желательно разместить их — чтобы избавить от долгих периодов ожидания — в компьютерах, стоящих в одном помещении.

---

* *Брадителический* (bradytelic) — присущий некоторым организмам темп морфологической эволюции, близкий к нулю, например отдельные виды моллюсков претерпели за последние 400 млн лет самые незначительные изменения.

Коллективный сверхразум

Следующий тип сверхразума представляет собой большое количество интеллектов более низкого уровня, собирающихся ради достижения сверхпроизводительности в одно целое. Определение этого типа сформулировано следующим образом:

> *коллективный сверхразум* — система, состоящая из большого количества интеллектов более низкого уровня, в силу этого ее общая производительность значительным образом превышает производительность любой существующей когнитивной системы во многих универсальных областях деятельности.

В отличие от скоростного, коллективный сверхразум не столь ясно очерчен концептуально[7], но более узнаваем с практической точки зрения. На собственном опыте мы еще никогда не сталкивались со скоростным искусственным интеллектом человеческого уровня, зато хорошо знакомы с таким понятием, как коллективный разум, представляющий объединение людей, организованных в единую систему, чтобы совместно находить решения более эффективные, чем может принимать отдельный, даже самый умный, член этого сообщества. Если несколько абстрагироваться и подойти к вопросу сугубо теоретически, то такого рода системами, способными решать проблемы самого разного уровня сложности, можно назвать компании, проектные группы, социальные сети, общественные организации, научные коллективы, государства и даже, чтобы не мелочиться, весь род человеческий. Из нашей социально-институциональной практики мы знаем, насколько проще принимать решения, если над ними трудится коллективный разум.

Лучше всего коллективный интеллект проявляет себя в разработке комплексных проектов, которые легко разложить на части, чтобы каждую можно было выполнять параллельно силами подструктур единой системы и проверять результаты в автономном режиме. При решении любых задач — от строительства космического корабля многоразового использования до управления сетью закусочных — существует огромное количество возможностей благодаря разделению труда. Над каждым компонентом шаттла работает специализированная команда проектировщиков и инженеров; каждое кафе обслуживается отдельным коллективом профессионалов. Научная среда в целом складывается из особых сообществ, каждое

из которых занимается своей отдельной дисциплиной и является самостоятельной системой с довольно жесткой структурой соподчиненных элементов: исследователи, преподаватели, студенты, журналы, гранты, премии — кстати, хочу заметить, что сложившаяся схема не очень способствует развитию того направления, которому посвящена моя книга. Но такова традиционная научная практика, и к этому можно было бы отнестись как к необходимому компромиссу, поскольку в рамках существующего огромное множество творческих личностей и целеустремленных команд, занимаясь самыми разными направлениями и работая практически автономно — когда каждый возделывает собственную научную делянку, — вносят свой коллективный вклад в сокровищницу человеческих знаний, продолжают и развивают их.

Такого рода разумная система может быть усилена за счет усовершенствования каждой отдельно подструктуры: расширение ее состава; повышение ее уровня интеллекта, оптимизация ее организационной политики[8]. Для превращения любого существующего коллективного интеллекта в сверхразум потребуется резкий рост на всех уровнях. Появившаяся в результате система должна быть способна значимо превосходить любой имеющийся коллективный интеллект и другие когнитивные системы во многих универсальных областях знаний. Рождаются и будут дальше появляться многие прогрессивные подходы, например: современные форматы проведения конференций, позволяющие ученым эффективнее обмениваться информацией; создание новейших алгоритмов анализа данных, способствующих лучшему отбору пользовательских предпочтений, в частности читателей и зрителей, — но каково бы ни было их значение, совершенно очевидно, что сами по себе эти инновационные факторы не приблизят нас к появлению коллективного сверхразума. Собственно, как и показатели вроде темпа прироста населения планеты или улучшения методов преподавания в учебных заведениях. Чтобы когнитивные способности человечества в целом начали соответствовать уровню коллективного сверхразума, потребуются совсем другие и количественные, и качественные критерии.

Обратите внимание, что порог для признания системы сверхразумной определяется относительно текущего уровня производительности, то есть на начало XXI века. В доисторические времена и на протяжении всей истории человечества возможности коллективного интеллекта выросли очень сильно. Со времен плейстоцена население Земли увеличилось в тысячу раз[9].

Исходя из этого — если принять за основу *уровень интеллекта эпохи плейстоцена*, — нынешний интеллектуальный уровень человечества можно рассматривать как приближающийся к сверхразумному. Столь же существенное влияние оказало совершенствование коммуникационных процессов, особенно возникновение устной речи, а потом и письменных языков, а также градостроение и книгопечатание. Все эти обстоятельства, как по отдельности, так и совокупно, стали огромным стимулом ускорения — настолько мощным, что появись сейчас подобного масштаба новаторский потенциал, его влияние на когнитивные способности всего человечества привело бы к появлению коллективного сверхразума[10].

Наверняка сейчас некоторые читатели возразят, что, мол, современное общество не кажется им слишком разумным. Возможно, в их родной стране недавно приняли какие-то непопулярные законы или несколько изменилась политическая обстановка, и очевидная неразумность происходящего оборачивается для людей прямым свидетельством моральной и интеллектуальной деградации социума. Разве не подтверждается их вывод об умственной недееспособности современного человечества вполне весомыми аргументами, такими как идолопоклонство перед материальными благами; истощение природных ресурсов; загрязнение окружающей среды; истребление видового разнообразия? Разве эти безобразия не происходят на общем фоне всемирного неравенства, вопиющей несправедливости и полного пренебрежения базовыми гуманистическими и духовными ценностями? Все так, только есть одно но. Оставив без внимания сравнительный анализ, насколько современные социальные изъяны ужаснее недостатков прошлых эпох, хочу вам заметить: в нашем определении коллективного сверхразума нет ничего, что говорило бы, будто высокоразвитое в интеллектуальном плане общество обязано быть справедливым и нравственным. Более того, в определении нет даже намека, будто высокоразвитое в интеллектуальном плане общество должно быть *мудрее*. Вы спросите, что такое «мудрость»? Договоримся считать мудростью способность относится к самому важному в нашей жизни с той или иной степенью здравого смысла. Представим себе некую организацию с немыслимо огромным штатом сотрудников — людей, обладающих большим умственным багажом, успешно и согласованно работающих, умеющих коллективно решать практически универсальные творческие и интеллектуальные проблемы. Предположим, эта организация может управлять практически любыми предприятиями, разрабатывать практически

любую технологию и достигать практически в любом процессе наивысшей продуктивности. Но даже *настолько* эффективная универсальная организация способна по какому-то принципиально важному, практически судьбоносному, вопросу вдруг принять в корне неверное решение — скажем, не продумать надлежащие меры предосторожности против рисков, угрожающих ее существованию, — и фантастически бурный подъем довольно быстро закончится полным и бесславным упадком. Такая организация могла бы стать носителем мощного общего интеллекта — настолько высокого, что еще чуть-чуть, и коллективный сверхразум получил бы свое реальное воплощение. А теперь вернемся к «справедливости» и «мудрости». Найдя какой-то желательный для нас признак, не стоит поддаваться искушению и обязательно наматывать эту ниточку на огромный клубок нашего общего и очень неопределенного представления о мыслительной деятельности, поскольку невозможно выбрать одно свойство, пусть даже достойное восхищения, без того, чтобы не рассмотреть аналогичным образом все остальные характеристики. Может быть, с этой точки зрения нам следовало бы осознать, как удобны мощные информационные системы — причем системы с элементами искусственного интеллекта, — которые по определению не могут быть ни справедливыми, ни преданными, ни мудрыми. Но к этому вопросу мы вернемся в седьмой главе.

Коллективный сверхразум может быть интегрирован слабо или сильно. В качестве иллюстрации слабоинтегрированного сверхразума представьте планету Мегаземля, на которой достигнут точно такой же уровень коммуникационных и координационных технологий, как на современной Земле, но при этом население больше земного в миллион раз. Соответственно, выше будут и совокупные интеллектуальные ресурсы. Предположим, что научные гении масштаба Ньютона или Эйнштейна появляются как минимум один раз на десять миллиардов человек — тогда на Мегаземле будут одновременно проживать семьсот тысяч гениев, не говоря уже о пропорционально большем количестве просто талантливых и одаренных мегаземлян. Новые идеи и технологии развивались бы на такой планете с бешеной скоростью, и глобальная цивилизация на Мегаземле представляла бы собой слабоинтегрированный сверхразум[11].

Если постепенно повышать степень интеграции коллективного интеллекта, в конечном счете он может превратиться в единый огромный «рассудок» в противоположность простому набору слабо связанных человеческих

умов[12]. Жители Мегаземли могли бы двигаться в этом направлении, совершенствуя коммуникационные и координационные технологии и разрабатывая лучшие методы организации совместной работы множества мегаземлян над трудными интеллектуальными задачами. Таким образом, коллективный сверхразум после заметного роста своей интегрированности воплотился бы в качественный сверхразум.

Качественный сверхразум

Попытаемся определить третий тип сверхразума:

> *качественный сверхразум* — система, по скорости работы сравнимая с человеческим умом, но в качественном отношении значительно сильнее его.

Понятие «качество интеллекта», как и в случае коллективного разума, является довольно расплывчатым, но ситуация усугубляется отсутствием у нас опыта обращения с интеллектуальными способностями, превосходящими верхние пределы современного человечества. Однако можно получить некоторое представление о них, изучив соответствующие случаи.

Прежде всего можно расширить спектр сравнения, включив в него других млекопитающих, обладающих интеллектом более низкого качества. (Не стоит считать эту ремарку дискриминационной по отношению к животным. Интеллект полосатой перцины отлично адаптирован к ее экологическим нуждам, но наша установка более антропоцентрична: мы рассматриваем производительность мозга с точки зрения сложных когнитивных задач, имеющих отношение к *людям*.) У животных отсутствует сложный структурированный язык; животные или вовсе не могут пользоваться инструментами и создавать их, или способны на это лишь в рудиментарной степени; они серьезно ограничены в способностях строить долгосрочные планы; у них очень незначительные способности к абстрактному мышлению. Ни одно из этих ограничений не объясняется недостатком скорости когнитивных процессов у животных или отсутствием у них коллективного интеллекта. По критерию одних только вычислительных возможностей человеческий мозг, вероятно, уступает мозгу крупных млекопитающих, включая слонов и китов. И хотя сложная технологическая человеческая цивилизация была бы невозможна без нашего огромного преимущества в коллективном интеллекте, нельзя сказать, что от него зависят умственные способности отдельных людей. Многие достигают высо-

кого уровня развития даже в небольших изолированных обществах типа охотничье-собирательского[13]. И напротив, высокоорганизованные шимпанзе и дельфины, которых обучают инструкторы-люди, или муравьи, живущие в огромных и хорошо организованных сообществах, по своему умственному развитию никогда не сравняются с человеком. Несомненно, поразительные интеллектуальные достижения *Homo sapiens*, в значительной степени явившись следствием специфических свойств архитектуры нашего мозга, — уникальный генетический дар, не доставшийся ни одному другому живому существу. Приведенные замечания помогут нам несколько уточнить представление о качественном сверхразуме. Итак, это интеллект, в качественном отношении значительно более сильный, чем человеческий, — настолько, насколько человеческий ум превосходит по качеству ум слонов, дельфинов и шимпанзе.

Второй способ уточнить определение качественного сверхразума — указать на локальный когнитивный дефицит, которым страдают некоторые люди, особенно когда он не вызван общим слабоумием или иными условиями, связанными с нарушением функционирования нейровычислительных ресурсов мозга. Возьмем, например, людей, страдающих аутизмом, — дефицит социального познания не мешает им нормально действовать в других познавательных сферах; или людей с врожденной амузией, неспособных промурлыкать или распознать даже простые мелодии, — кроме этого неудобства, их жизнь ничем не отличается от жизни остальных людей. В специализированной литературе по психоневрологии в изобилии приведены описания пациентов, страдающих от узколокальных патологических состояний мозга, вызванных генетическими нарушениями или травмами. Эти примеры показывают, что нормальные взрослые люди обладают широким спектром удивительных познавательных талантов, которые нельзя считать простой функцией общей мощности нейровычислительной системы или даже достаточного уровня общего интеллекта — требуется также специфическая схема взаимодействия нейронов. Из этого наблюдения следует идея *возможных, но нереализованных талантов познания* — то есть талантов, которыми не обладает ни один человек, даже если другие интеллектуальные системы, причем не превосходящие человеческий мозг с точки зрения вычислительной мощности, которые их имеют, могли бы очень сильно выиграть из-за своей способности выполнять широкий спектр стратегически важных задач.

Соответственно, обратившись к примеру животных и людей, страдающих специфическими нарушениями познавательных способностей, мы можем получить некоторое представление о различных качествах интеллекта и их практических отличиях. Если у *Homo sapiens* отсутствовали бы (лишь в качестве примера) когнитивные модули, позволяющие ему формировать сложные речевые конструкции, он остался бы лишь еще одним видом обезьян, живущих в гармонии с природой. И напротив, найди человек разумный способ *обрести* некий новый набор модулей, обеспечивающий его преимуществом, сопоставимым со способностью формировать сложные речевые конструкции, он стал бы человеком сверхразумным.

Прямая и опосредованная досягаемость

Сверхразум, возникший по какому-то одному из описанных выше типов, со временем мог бы развить технологии, необходимые для создания сверхразума и по остальным типам. Таким образом, *опосредованно* все три типа сверхразума одинаково достижимы. В этом смысле опосредованная досягаемость интеллекта человеческого уровня попадает в тот же класс эквивалентности, если исходить из допущения, что мы вообще способны прийти хотя бы к какому-то сверхразуму. Однако в чем-то эти три типа гораздо ближе друг к другу, поскольку любой из них способен создать два других гораздо быстрее, чем мы — один из них, если брать за точку отсчета сегодняшнее развитие технологий.

Прямую досягаемость трех разных типов сверхразума сравнивать сложнее. Скорее всего, ранжировать их не получится. Возможности каждого из них зависят от того, в какой степени они демонстрируют свои преимущества, то есть *насколько* быстро работает скоростной сверхразум, *насколько* качественнее качественный сверхразум и так далее. Максимум мы можем сказать, что при прочих равных скоростной сверхразум отлично справляется с задачами, требующими быстрого выполнения длинной последовательности шагов, которые должны быть сделаны один за другим, в то время как коллективный сверхразум лучше показывает себя в задачах, допускающих аналитическую декомпозицию на параллельные подзадачи, а также в таких, когда требуется комбинация множества различных точек зрения и наборов навыков. Качественный сверхразум в некотором смысле должен быть самым универсальным типом, поскольку способен справиться с задачами, находящимися вне пределов *прямой* досягаемости скоростного и коллективного сверхразумов[14].

Не все в нашей жизни определяется этими категориями, и не всегда количество способно заменить качество. Один гениальный отшельник*, запершись в спальне, обитой пробковым дубом, способен написать «В поисках утраченного времени». Можно ли создать подобный шедевр, собрав в одном помещении множество поденщиков от литературы?[15] При всем существующем многообразии человеческих характеров и дарований мы видим, что в некоторых случаях работа только выигрывает, когда ее выполняют не мириады посредственностей, а берется за нее всего лишь один специалист, но блестящий мастер своего дела. Если посмотреть на это шире, то придется признать вероятность существования таких интеллектуальных задач, с которыми сможет справиться только сверхразум — они окажутся не по плечу даже огромному коллективу обычных людей без усовершенствованных когнитивных способностей.

Таким образом, могут быть задачи, которые способен решить качественный сверхразум и, возможно, скоростной сверхразум, но не слабоинтегрированный коллективный сверхразум[16] (если, конечно, он не займется в первую очередь развитием собственных возможностей). Мы не сможем определить точно характер этих задач, но попробуем описать их в общем виде[17]. Скорее всего, такая задача должна состоять из мультикомплексных взаимозависимостей, не позволяющих разбить ее на автономные подструктуры, следовательно, ее решение может потребовать качественно нового понимания или нового подхода, которые слишком сложны для восприятия нынешнего поколения смертных. В категорию подобных задач могут попадать отдельные виды художественного творчества, стратегических моделей и даже некоторых научных открытий. Кто-то скажет, что медлительность и крайняя неуверенность в себе в деле решения так называемых вечных вопросов философии связаны с неприспособленностью коры головного мозга человека к умозрительным размышлениям. Поэтому деятельность наших известных философов напоминает походку собаки, которую хозяин заставляет ходить на задних лапах: они насилу достигают «уровня исполнения», который *хоть каким-то образом* позволяет заниматься этой деятельностью[18].

Источники преимущества цифрового интеллекта

Серьезные последствия могут иметь даже незначительные изменения в объеме и устройстве мозга, что видно, если сравнить интеллектуальные

* Речь идет о Марселе Прусте.

и технологические достижения людей и человекообразных обезьян. Те сверхмасштабные изменения в вычислительной мощности и архитектуре, которые позволяет осуществить использование искусственного интеллекта, могут иметь гораздо более глубокие последствия. Нам очень трудно — если вообще возможно — интуитивно понять, на что способен сверхразум, можно попытаться лишь приблизиться к этому пониманию, взглянув на преимущества, которыми обладает цифровой интеллект. Легче всего оценить плюсы аппаратного обеспечения.

— *Скорость вычислительных элементов*. Пиковая скорость работы биологических нейронов — около 200 Гц, что на семь порядков медленнее современных микропроцессоров (примерно 2 ГГц)[19]. Как следствие, человеческий мозг вынужден полагаться на масштабное распараллеливание задач и неспособен быстро выполнять вычисления, требующие большого количества последовательных операций[20]. (Мозгу под силу лишь несколько десятков таких операций, максимум — чуть больше сотни.) При этом многие из наиболее важных алгоритмов в программировании и кибернетике не так-то легко поддаются распараллеливанию. Многие когнитивные задачи можно было бы решать гораздо эффективнее, если бы естественная склонность мозга к параллельным алгоритмам распознавания образов дополнялась бы возможностью — и интегрировалась с возможностью — быстрых последовательных вычислений.

— *Скорость внутренних коммуникаций*. Аксоны передают потенциал действия со скоростью 120 м/с или даже меньше, в то время как электронные центры обработки информации используют оптику, в которой информация передается со скоростью света (300 000 000 м/с)[21]. Медлительность нейронных сигналов ограничивает размеры биологического мозга, который может функционировать как единый вычислительный блок. Например, чтобы задержка в передаче сигналов от одного элемента к другому и обратно между двумя произвольными элементами системы не превышала 10 мс, объем биологического мозга не должен быть больше 0,11 м^3. А размер аналогичной электронной системы может равняться $6{,}1 \times 10^{17}$ м^3 (это размер карликовой звезды), то есть на восемнадцать порядков больше[22].

— *Количество вычислительных элементов*. В человеческом мозгу чуть меньше 100 миллиардов нейронов[23]. Он примерно в три с половиной раза больше мозга шимпанзе (правда, при этом в пять раз меньше мозга

кашалота)[24]. Очевидно, что количество нейронов в биологическом существе ограничено объемом черепа и особенностями метаболизма, но в случае крупного мозга вступают в действие и другие ограничения (охлаждение, время созревания, задержки в передаче сигнала — см. предыдущий пункт). В отличие от биологического мозга, компьютерное оборудование масштабируется до гигантских физических размеров[25]. Суперкомпьютеры могут быть размером со склад или даже больше, причем с помощью высокоскоростных кабелей к ним можно подключать дополнительные удаленные вычислительные мощности[26].

— *Емкость памяти*. Человек способен удерживать в кратковременной памяти не более четырех-пяти блоков информации одновременно[27]. Хотя сравнивать напрямую кратковременную память с оперативной памятью компьютера не совсем корректно, ясно, что конструктивные преимущества цифрового интеллекта позволяют ему иметь рабочую память гораздо большего размера. Это значит, что такой интеллект способен интуитивно схватывать суть сложных взаимоотношений, которые люди могут нащупать лишь при помощи кропотливого труда[28]. Долгосрочная человеческая память также ограниченна, хотя пока и не ясно, способны ли мы исчерпать ее возможности по хранению информации в течение обычной человеческой жизни, ведь скорость накопления нами информации так мала. (По одной из оценок, мозг взрослого человека может хранить примерно миллиард бит, что на пару порядков величины меньше, чем самый простой смартфон[29].) В случае машинного мозга больше и объем хранимой информации, и скорость доступа к ней.

— *Надежность, продолжительность жизни, сенсоры и другое.* Машинный интеллект может иметь и другие преимущества на уровне оборудования. Например, биологические нейроны менее надежны, чем транзисторы[30]. Поскольку зашумленные вычисления требуют дополнительных схем декодирования, в которых для обработки единственного бита информации требуется множество элементов, цифровой интеллект получает некоторое преимущество благодаря использованию надежных высокоточных вычислительных элементов. Мозг устает уже после нескольких часов работы и начинает сдавать через несколько десятков лет субъективного времени, у микропроцессоров таких ограничений нет. Поток данных в машинном мозгу можно увеличить за счет добавления миллионов сенсоров. В зависимости

от используемой технологии машина может иметь изменчивую архитектуру, способную к оптимизации при изменении требований к выполняемым задачам, в то время как большая часть архитектуры мозга человека фиксирована с рождения, если она и меняется, то незначительно (хотя связи между синапсами могут меняться в течение таких коротких промежутков времени, как несколько дней)[31].

В настоящее время вычислительная мощность биологического мозга все еще превосходит мощность компьютеров, хотя самые современные сверхмощные компьютеры уже достигают уровня производительности, соответствующей оценкам производительности человеческого мозга[32]. Но компьютерное оборудование очень быстро совершенствуется, и предельные возможности его вычислительной мощности намного превышают возможности вычислительных систем биологических компьютеров.

Цифровой мозг имеет крупные преимущества также с точки зрения программного обеспечения.

— *Редактируемость*. С параметрами ПО можно экспериментировать, что практически нельзя делать с нейронной системой биологического головного мозга. Например, в компьютерной модели мозга можно легко посмотреть, что будет, если добавить больше нейронов в ту или иную область коры головного мозга, если повысить или понизить их возбудимость. Проведение таких экспериментов на живом биологическом мозгу было бы гораздо более трудным делом.

— *Дублируемость*. Можно быстро сделать сколько угодно точных копий ПО для установки на имеющееся оборудование. Напротив, чтобы воспроизвести биологический мозг, потребуется очень много времени, поскольку каждый «новорожденный» совершенно беспомощен и не помнит ничего, чему научились его «родители» в течение своей жизни.

— *Координация целей*. Коллективы людей страдают от неэффективности, связанной с тем, что практически невозможно достичь полного единства целей их членов, — и так будет по крайней мере до тех пор, пока не получится добиться покорности при помощи лекарственных препаратов или генетической селекции. У клана копий (группы идентичных или почти идентичных программ, разделяющих общие цели) таких проблем с координацией нет.

— *Использование общей памяти*. Биологический мозг нуждается в длительном обучении и наставничестве, в то время как цифровой может

получать воспоминания и навыки, обмениваясь файлами с другими программами. Популяция из миллиарда копий программ искусственного интеллекта могла бы периодически синхронизировать свои базы данных, чтобы каждая из них знала все, чему остальные научились за прошедший час. (Прямая передача данных требует стандартизированных форматов представления информации. Поэтому простой обмен когнитивным контентом высокого уровня между любой парой программ искусственного интеллекта невозможен. В частности, это не получится сделать для компьютерных моделей мозга первого поколения.)

— *Новые модули, модели поведения и алгоритмы*. Восприятие зрительных образов кажется нам простым и не требующим усилий делом, в отличие от решения геометрических задач из школьного учебника, несмотря на то что для этого требуется огромный объем вычислений, чтобы создать реконструкцию трехмерного мира, населенного знакомыми нам объектами, из возникающих на нашей сетчатке двумерных моделей. А простым нам это кажется потому, что в нашем мозгу имеется специальный низкоуровневый нейронный механизм для обработки визуальной информации. Эта низкоуровневая обработка происходит неосознанно и автоматически, без расходования психической энергии и без отвлечения внимания. Восприятие музыки, использование языка, социальное познание и другие формы обработки информации, «естественной» для нас, людей, похоже, также поддерживается специализированными нейровычислительными модулями. Искусственный интеллект, в котором имелись бы такие же модули поддержки в важных для современного мира предметных областях, таких как, например, программирование и разработка проектов и бизнес-стратегий, имел бы большое преимущество перед нами, поскольку человеку, чтобы думать о таких вещах, приходится полагаться на неуклюжий универсальный механизм познания. Кроме того, для использования преимуществ, специфических для компьютеров, — скажем, быстрых последовательных вычислений — могут быть созданы специальные новые алгоритмы.

Предельные преимущества машинного интеллекта, представляющего синтез аппаратного и программного обеспечений, просто громадны[33]. Но насколько быстро можно реализовать этот потенциал? Ответом на этот вопрос мы займемся в следующей главе.

Глава четвертая

Динамика взрывного развития интеллекта

Итак, наступит момент, когда машины практически сравняются с человеком в общей способности осмысливания, — сколько времени им потребуется, чтобы обрести сверхразум? Будет ли этот переход неторопливым, постепенным, продолжительным? Произойдет ли он внезапно — во взрывном темпе? В главе анализируется динамика перехода к сверхразуму с точки зрения функциональности самой системы: силы оптимизации и сопротивляемости. Как поведут себя эти два фактора вблизи универсального интеллекта человеческого уровня? Будем исходить из того, что или якобы понимаем это, или по крайней мере сможем выстроить приемлемое предположение.

Время и скорость взлета

Допустим, что *рано или поздно* машинный интеллект значительно превзойдет общий интеллектуальный уровень человека, при этом мы не должны забывать, что *на сегодняшний день* познавательные способности человека в огромной степени превосходят возможности машинного познания. Нас не может не волновать, когда свершится самоуправство машин и насколько быстро утвердится их монополия на познание. Причем этот вопрос нужно четко отличать от поставленного нами в первой главе, а именно: насколько мы далеки от создания универсального искусственного интеллекта человеческого уровня. Сейчас речь идет о другом: *если машина, наделенная универсальным интеллектом человеческого уровня, будет когда-то и где-то создана, сколько ей понадобится времени, чтобы полностью превратиться в сверхразумную?* Заметьте, в отношении создания УИИЧУ можно занимать

разные позиции: считать, что для этого потребуется очень длительное время; отрицать саму возможность хоть в малейшей степени оценить этот срок, — но в любом случае быть абсолютно уверенным, что как только человечество этого достигнет, дальнейшее развитие, то есть подъем на высочайший сверхразумный уровень, произойдет очень быстро.

Может быть, целесообразнее представить процесс развития схематично, даже если придется временно не принимать во внимание некоторые условия и подробности, усложняющие суть дела. Посмотрим на диаграмму, отражающую интеллектуальные способности наиболее совершенных систем искусственного интеллекта как функцию времени (см. рис. 7).

Горизонтальная линия с надписью «человеческий уровень» представляет характерные для современных развитых стран интеллектуальные возможности среднестатистического взрослого человека, имеющего доступ к источникам информации и технологическим средствам. В настоящее время даже самая передовая интеллектуальная система в значительной степени — по любым разумным показателям — отстает от основных человеческих характеристик. В будущем, в какое-то неопределенное для нас время, интеллектуальная система сможет достичь приблизительного равенства с человеком (за основу нами принят зафиксированный в 2014 году общий уровень интеллекта человека, даже если в последующие годы он предположительно будет расти) — этот момент будет свидетельствовать о начале взлета. Возможности интеллектуальной системы продолжат увеличиваться и спустя некоторое время достигнут паритета с объединенными интеллектуальными способностями всего человечества (уровень фиксации тот же 2014 год) — это мы обозначим как «цивилизационный уровень». Если интеллектуальная система пойдет по пути дальнейшего развития, она наконец превратится в сильный сверхразум, то есть с точки зрения интеллектуальных возможностей выйдет на уровень, значительно превосходящий уровень человечества в целом, — когда это будет достигнуто, можно говорить о завершении взлета, хотя сама сверхразумная система в состоянии самосовершенствоваться и дальше. В какой-то момент в течение процесса взлета система может пройти отметку «критический рубеж», то есть точку, за которой совершенствование будет зависеть не от деятельности людей, а в основном от ее собственных усилий[1]. (Возможность существования критического рубежа, или точки перехода, важна для обсуждения тем, рассматриваемых в последнем разделе главы, — сила оптимизации и взрывоопасность.)

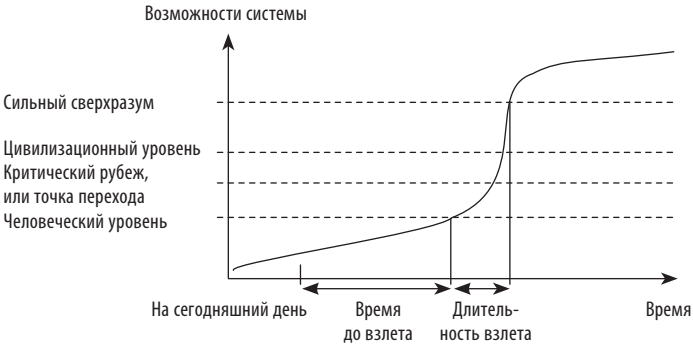

Рис. 7. Схема взлета. Важно разделять такие вопросы, как: состоится ли взлет? если взлет состоится, то когда? если когда-нибудь взлет произойдет, насколько резким он будет? Например, можно считать, что до взлета еще очень много времени, но когда момент наступит, все произойдет очень быстро. Есть еще один важный вопрос (на диаграмме не показан): какова будет роль мировой экономики в процессе взлета сверхразумной системы? Все вопросы связаны между собой.

На диаграмме представлен обобщенный сценарий перехода интеллектуальной системы с человеческого на сверхразумный уровень, но в зависимости от крутизны кривой можно выделить три типа сценариев взлета — медленный, быстрый, умеренный.

1. **Медленный взлет.** Будет происходить на долгих временных интервалах — от десятилетий до столетий. Такие сценарии представляют собой отличные возможности, чтобы человечество смогло обдумать свои дальнейшие политические ходы, адаптировать их к ситуации и правильно реагировать на происходящее. Можно будет последовательно тестировать разные подходы, обучить и подготовить новых специалистов. Предвыборные кампании будут ориентированы на сторонников и противников происходящих процессов. Если выяснится, что нужны новые системы безопасности и общественного контроля над исследованиями в области ИИ, будет достаточно времени, чтобы их разработать и развернуть. Страны и ученые, опасающиеся гонки вооружений в области военного применения ИИ, смогут провести необходимые переговоры и разработать защитные механизмы. Большинство подготовительных мер, предпринятых накануне процесса медленного взлета, окажутся морально устаревшими, но постепенно начнут появляться новые решения, актуальные для наступающей эпохи.

2. **Быстрый взлет.** Произойдет в течение очень короткого промежутка времени — минуты, часы, дни. В сценариях быстрого взлета у людей почти не остается возможности реагировать на него. Никто не успеет ничего заметить, а игра уже окажется проигранной. В любом сценарии быстрого взлета судьба человечества в основном зависит от подготовительной работы, уже проведенной к этому моменту. В самом медленном из быстрых сценариев со стороны людей еще могут быть произведены какие-то элементарные телодвижения вроде потянуться открыть «ядерный чемоданчик», но тогда алгоритмы этих действий должны быть или действительно элементарны, или запланированы, запрограммированы и отрепетированы заранее.
3. **Умеренный взлет.** Произойдет в течение относительно короткого временного интервала — от нескольких месяцев до нескольких лет. В этом случае у человечества остается некоторая возможность отреагировать на происходящее, но не будет времени проанализировать его, протестировать разные подходы и решить сложные задачи координации действий. На создание и развертывание новых систем, таких как политические меры, механизмы контроля, протоколы безопасности компьютерных сетей, времени тоже не останется, но, возможно, получится приспособить к новым обстоятельствам уже существующие нормы.

В случае медленного взлета будет достаточно времени для распространения новостей. При умеренном взлете разработки могут остаться засекреченными. Знанием о них будут располагать лишь очень ограниченные группы людей, как это бывает при разработке секретных государственных программ военных исследований. Так же тайно, силами небольших научных коллективов или хакеров в каком-нибудь подвале, могут вестись работы в рамках коммерческих проектов — впрочем, если перспективы взрывного развития сверхразума попадут на радары государственных спецслужб и станут одним из приоритетов национальной безопасности, у наиболее многообещающих частных проектов появятся большие шансы оказаться под наблюдением. В таком случае у государства (или доминирующего международного игрока) будет возможность национализировать или закрыть любой проект, демонстрирующий признаки скорого взлета.

Быстрый взлет произойдет за столь короткое время, что никто не успеет не то что отреагировать на него, но даже рот открыть. Вмешаться будет

возможно *до начала* взлета — правда, лишь при условии, что наготове имеется механизм по защите.

Сценарии умеренного взлета могут привести к геополитическим, социальным и экономическим потрясениям, поскольку отдельные люди и сообщества постараются обеспечить себе выигрышные позиции в ходе разворачивающихся преобразований. Возможные беспорядки начнут тормозить усилия по координации ответных мер и спровоцировать принятие решений более радикальных, чем те, которые были бы приняты в спокойной ситуации. Например, по одному из сценариев умеренного взлета в течение нескольких лет рынок труда постепенно заполняют дешевые и сверхпроизводительные имитационные модели, полученные вследствие полной эмуляции мозга, или иные системы искусственного интеллекта. При таком развитии событий можно допустить, когда начнутся массовые увольнения работников-людей, не менее массовые протесты, которые вынудят правительства увеличить пособия по безработице, ввести гарантии пожизненной занятости для всех граждан, а также специальные налоги или обязательные социальные отчисления для работодателей, использующих труд работников-роботов. Чтобы принятые в результате такой политики меры не оказались пустыми, они должны быть поддержаны новыми политическими и социальными структурами, действующими на постоянной основе. Аналогичные проблемы не исключены при сценариях медленного взлета, но в сценариях умеренного взлета диспропорции и скорость изменений способны привести к тому, что небольшие группы активистов получат непропорционально большое влияние.

Некоторым читателям может показаться, что из перечисленных сценариев взлета наиболее вероятен — медленный, менее вероятен — умеренный, практически невозможен — быстрый. Возможно ли предположить, что всего за пару часов мир в состоянии радикально измениться, а венец творения — быть сброшенным с вершины разума? С подобными громадными переменами история человечества еще не сталкивалась, а ближайшие параллели — неолитическая (сельскохозяйственная) и промышленная революции — потребовали гораздо больше времени (от нескольких веков до тысячелетия в первом случае, от нескольких десятилетий до веков — во втором). То есть база для трансформации такого рода, какая предполагается в случае сценариев быстрого или умеренного взлета, с точки зрения величины и скорости происходящих изменений, равна нулю, поскольку прецедентов — если не брать в расчет мифологии и религии — в человеческой истории не было[2].

Однако медленный сценарий представляется как раз самым маловероятным — в пользу этого положения я приведу далее соответствующие аргументы. Скорее всего, события начнут развиваться по быстрому сценарию — если когда-нибудь взлет произойдет, он будет иметь взрывной характер.

Чтобы получить ответ на вопрос, насколько быстрым будет взлет, можно считать скорость роста интеллектуальных способностей системы монотонно возрастающей функцией двух переменных: величины «силы оптимизации», или взвешенных на качество усилий по разработке, прилагаемых с целью сделать систему более интеллектуальной, и «отзывчивости» системы к приложению оптимизирующей силы. Можно также ввести термин «сопротивляемость», обратный отзывчивости, и записать следующее:

$$\text{Скорость роста интеллекта} = \frac{\text{Сила оптимизации}}{\text{Сопротивляемость}}$$

В отсутствие количественного определения уровня интеллекта, усилий по его повышению и сопротивляемости это выражение остается по большей части качественным. Но по крайней мере можно будет наблюдать, что интеллектуальные способности системы быстро растут, когда *или* прилагаются большие усилия для их развития, притом что развить ее способности не очень трудно, *или* нужны нетривиальные усилия по разработке правильного дизайна системы, притом что сопротивляемость ее низка (или справедливо и то и другое). Если мы знаем, сколько усилий прилагается для улучшения той или иной системы, и скорость улучшений, которую они обеспечивают, то можем рассчитать сопротивляемость этой системы.

Далее, мы можем наблюдать, что сила оптимизации, направленная на улучшение эффективности той или иной системы, с течением времени сильно изменяется от системы к системе. Сопротивляемость системы также может сильно зависеть от того, насколько система уже оптимизирована. Часто первыми делаются простейшие улучшения, в результате чего — по мере срывания самых низко висящих плодов — результаты появляются все медленнее (растет сопротивляемость системы). Однако могут быть улучшения, облегчающие последующие улучшения, и тогда возникает каскад улучшений. Процесс складывания пазла поначалу кажется простым, ведь сложить углы и края картины довольно легко. Затем сопротивляемость растет — подбирать последующие части становится труднее. В конце, когда сборка пазла почти завершена, свободных мест практически не остается, и процесс вновь идет легко.

Таким образом, чтобы найти ответ на наш вопрос, следует проанализировать, как меняются в критические моменты взлета степень сопротивляемости и сила оптимизации. Этим мы и займемся далее.

Сопротивляемость

Начнем с сопротивляемости. В этом случае прогноз зависит от типа рассматриваемой системы. Для полноты картины вначале рассмотрим сопротивляемость, которая характерна для путей движения в сторону сверхразума, не связанных с созданием усовершенствованного машинного интеллекта. Мы считаем, что в таком случае сопротивляемость будет довольно высока. Справившись с этим, обратимся к главному — взлету, связанному с искусственным интеллектом. Похоже, в данном случае сопротивляемость в критический момент окажется довольно низкой.

Пути, не подразумевающие создания машинного интеллекта

Результативность когнитивного совершенствования за счет улучшения здравоохранения и организации здорового питания со временем резко падает[3]. Многого можно добиться за счет ликвидации такого жесткого фактора, как дефицит питательных веществ в ежедневном рационе, но этого явления уже нет практически нигде, кроме самых бедных стран. Дальнейшее улучшение и так уже полноценного рациона мало на что влияет. Усиление образовательного процесса будет тоже играть все меньшую роль. Доля одаренных людей, лишенных доступа к качественному образованию, все еще велика, но постепенно снижается.

В ближайшие десятилетия какое-то повышение уровня интеллектуальных способностей можно будет обеспечить за счет медикаментозных средств. Но после того как удастся добиться самых простых улучшений — скорее всего, речь пойдет об укреплении на продолжительный срок умственных сил, долговременной памяти и способности сосредоточивать внимание, — дальнейший прогресс будет даваться все с большим и большим трудом. По сравнению с такими долговременными подходами, как обеспечение правильного питания и охрана здоровья, обращение к лекарственным средствам, стимулирующим умственные способности, вначале может показаться быстрым и легким путем, но это не так. В нейрофармакологии по-прежнему много белых пятен даже на уровне элементарных знаний, необходимых, чтобы компетентно вмешиваться в работу здорового

мозга. Отчасти медленный прогресс вызван неготовностью считать медикаментозные методы достойной областью исследований. Видимо, нейробиология и фармакология как науки еще какое-то время будут развиваться, не обращая особого внимания на проблему усиления когнитивных функций. Что поделаешь, придется ждать тех времен, когда разработка ноотропных средств все-таки станет одним из приоритетных направлений — тогда наконец удастся добиться относительно быстрых и простых побед[4].

Генетическое когнитивное улучшение имеет U-образный профиль сопротивляемости, как и у ноотропов, но с более заметными потенциальными выгодами. Вначале — пока единственным доступным методом остается селекция на протяжении нескольких поколений, которую, конечно, трудно проводить в общемировом масштабе, — уровень сопротивляемости высок. По мере того как появляются технологии для недорогого и эффективного генетического тестирования и селекции (особенно когда станет доступным итеративный отбор эмбрионов в случае людей), генетическое улучшение становится более простым делом. Эти новые методы позволят проверить весь спектр существующих вариаций человеческих генов на наличие повышающих интеллект аллелей. Однако после того как лучшие из них будут включены в пакеты генетического улучшения, дальнейший прогресс окажется делом более трудным. Необходимость разработать более инновационные подходы к генетической модификации может повысить сопротивляемость. На пути генетического улучшения скорость прогресса ограничена — ограничена в основном тем обстоятельством, что вмешательство на эмбриональном уровне зависит от неизбежной отсрочки на время взросления, что препятствует развитию сценариев с быстрым или умеренным взлетом[5]. Метод селекции эмбрионов неразрывно связан с длительным процессом внедрения технологии ЭКО в общественное сознание, что является еще одним ограничивающим фактором.

На пути создания нейрокомпьютерного интерфейса, похоже, сопротивляемость поначалу будет очень высокой. В маловероятном сценарии, когда вдруг получится легко вживлять имплантат и достигать высокоуровневой функциональной интеграции с корой головного мозга, сопротивляемость пойдет вниз. Но в долгосрочном плане ситуация со все менее заметным прогрессом при движении по этому пути будет аналогична случаю с улучшением имитационной модели головного мозга и интеллектуальных систем, поскольку в конечном счете интеллект системы мозг–компьютер будет обеспечивать ее компьютерная составляющая.

В случае попыток сделать сети и организации более эффективными сопротивляемость в большинстве случаев будет высокой. Чтобы ее преодолеть, потребуются значительные усилия, в результате чего совокупные интеллектуальные возможности человечества, возможно, будут увеличиваться всего на пару процентов в год[6]. Более того, сдвиги во внутренней и внешней среде приведут к тому, что даже эффективные в какой-то момент организации перестанут приспосабливаться к новым обстоятельствам. То есть потребуются постоянные усилия по их реформированию, хотя бы для того, чтобы не допустить ухудшения ситуации. Скачкообразный рост темпов повышения производительности средней организации, в принципе, возможен, но трудно вообразить, что даже в наиболее радикальном сценарии такого рода произойдет не медленный, а умеренный и тем более быстрый взлет, поскольку организации, в которых работают люди, ограничены таким понятием, как человеко-часы. Передовым рубежом со множеством возможностей для повышения уровня коллективного интеллекта остается интернет, причем пока сопротивляемость здесь остается на умеренном уровне: прогресс заметен, но для его достижения требуется много усилий. После того как все низко висящие плоды окажутся сорванными (вроде поисковых систем и электронной почты), сопротивляемость, скорее всего, начнет расти.

Пути создания имитационной модели мозга и искусственного интеллекта

Степень трудности, стоящей на пути полной эмуляции головного мозга, оценивать довольно сложно. Но одна веха на этом пути уже пройдена: успешно создана имитационная модель мозга насекомого. Это та возвышенность, взобравшись на которую можно увидеть, что за ландшафт ждет нас впереди, и понять уровень сопротивляемости, с которым мы столкнемся в процессе масштабирования технологий моделирования человеческого мозга. (Еще лучшей точкой обзора станет успешное моделирование мозга небольшого млекопитающего вроде мыши, после чего расстояние, оставшееся до создание имитационной модели головного мозга человека, можно будет оценить довольно точно.) Однако путь создания искусственного интеллекта может не иметь таких очевидных контрольных точек обзора. Вполне возможно, что экспедиция к ИИ будет долго блуждать в непроходимых джунглях, пока внезапный прорыв не приведет к ситуации, что мы окажемся всего лишь в нескольких шагах от финишной черты.

Вспомним о двух вопросах, о которых я говорил, что их принципиально нужно различать. Насколько мы далеки от создания универсального искусственного интеллекта человеческого уровня? Насколько быстро произойдет переход от этого уровня к сверхразумному? Первый вопрос больше относится к оценке того, сколько времени потребуется на подготовку взлета. Для ответа на второй вопрос важно определить траекторию полета к сверхразуму — цели нашего исследования в этой главе. Довольно соблазнительно предположить, что шаг от создания УИИЧУ к появлению сверхразума будет очень сложным, что этот шаг, в конце концов, должен быть сделан «на большей высоте», где к уже имеющейся системе нужно будет добавить дополнительную мощность. Считать так было бы неверно. Вполне возможно, что с достижением равенства между интеллектуальными способностями человека и машины сопротивляемость *упадет*.

Рассмотрим в первую очередь полную эмуляцию головного мозга. Трудности, ожидающие нас на пути создания первой успешной имитационной модели, разительно отличаются от трудностей, связанных с ее дальнейшим обслуживанием. В первом случае придется решить множество технических проблем, особенно в области разработки необходимых технологий сканирования и обработки изображений. Нужно будет создавать материальные активы, возможно, целый парк из сотен высокопроизводительных сканирующих устройств. Напротив, повышение качества уже существующей имитационной модели мозга связано скорее с совершенствованием программного обеспечения — с тонкой настройкой алгоритмов и уточнением структуры данных. Эта задача может оказаться легче, чем создать технологию обработки изображений, без которой дальнейший прогресс невозможен в принципе. Программисты могут экспериментировать с разными параметрами вроде количества нейронов в различных областях коры головного мозга, чтобы посмотреть, как это влияет на производительность модели[7]. А также поработать над оптимизацией кода и поискать более простую вычислительную модель, способную сохранить всю основную функциональность отдельных нейронов и небольших нейронных сетей. Если последней технологической составляющей, которой будет недоставать для решения задачи, окажется сканирование или трансляция, притом что вычислительная мощность будет в избытке, тогда на первом этапе не придется уделять особое внимание вопросам эффективности, поэтому на втором этапе можно будет многое оптимизировать. (Вероятно, получится провести фундаментальную

реорганизацию вычислительной архитектуры, однако это уведет нас с пути эмуляции на путь создания ИИ.)

Еще один способ улучшить код имитационной модели мозга после того, как она будет создана, это сканировать головной мозг других людей с отличающимися или более высокими навыками и талантами. Можно также добиться роста производительности в результате последовательной адаптации организационной структуры и процессов к специфике функционирования цифрового интеллекта. Поскольку в мировой экономике, созданной человеком, не было аналогов работника, которого можно в буквальном смысле слова скопировать, перезапустить, ускорить или замедлить и так далее, то у руководителей первого коллектива работников эмуляций будет множество возможностей для инноваций в области управления.

После снижения сопротивляемости в результате создания первой имитационной модели человеческого мозга ее уровень может снова начать расти. Рано или поздно будут устранены наиболее очевидные недостатки, сделаны наиболее многообещающие изменения алгоритма, использованы самые простые возможности для организационных улучшений. Библиотека исходных данных вырастет настолько, что сканирование дополнительных экземпляров мозга перестанет заметно повышать качество модели. Поскольку имитационную модель можно копировать, а каждую копию обучать, то есть загружать в нее самые разнообразные знания, возможно, для получения максимального эффекта не придется сканировать мозг множества людей. Вполне вероятно, что довольно будет всего одного.

Еще одна потенциальная причина роста сопротивляемости состоит в том, что сами имитационные модели или их защитники-люди могут организовать движение за регулирование отрасли, что повлечет за собой ряд ограничений: сокращение числа работников-эмуляций; лимит на копирование имитационных моделей; запрет на определенные виды экспериментов над моделями, — но кроме этого работники-эмуляции получат гарантированное соблюдение своих прав, установленную специально для них минимальную заработную плату, ну и так далее. Правда, вполне допустимо, что политическое развитие пойдет в прямо противоположном направлении, в результате чего сопротивляемость упадет. Это может случиться, если первоначальные ограничения в использовании работников-эмуляций уступят место их безоглядной эксплуатации — после того как вырастет конкуренция и станут очевидны высокие экономические и стра-

тегические издержки слишком щепетильного отношения к моральным аспектам проблемы.

Что касается искусственного интеллекта (машинного интеллекта, не связанного с имитационным моделированием человеческого мозга), то степень сложности его развития от УИИЧУ до уровня сверхразума за счет совершенствования алгоритма будет зависеть от особенностей конкретной системы. Сопротивляемость разных архитектур может различаться очень сильно.

В некоторых случаях она, вероятно, окажется чрезвычайно низкой. Скажем, создание ИИЧУ задерживается из-за какого-то последнего, ускользающего от программистов штриха, но после того как он будет найден, ИИ может в одночасье преодолеть разрыв между уровнем ниже человеческого и уровнем, значительно его превосходящим. Еще одна ситуация, в которой сопротивляемость может оказаться низкой, — это когда умственные способности ИИ развиваются за счет двух различных способов обработки информации. Представим ИИ, состоящий из двух подсистем, одна из которых отвечает за методы решения узкоспециализированных задач, другая — за универсальное мышление. Вполне возможно, что если вторая подсистема недотягивает до некоторого порогового значения производительности, она ничего не добавляет в совокупную производительность системы, поскольку ее решения всегда хуже тех, которые вырабатывает подсистема работы со специализированными задачами. Теперь предположим, что к подсистеме универсального мышления приложили определенную силу оптимизации, в результате чего производительность подсистемы резко повысилась. Поначалу мы не увидим изменения в совокупной производительности системы, что говорит о ее высокой сопротивляемости. Затем, когда возможности второй подсистемы пересекут некоторое пороговое значение, ее решения станут превосходить решения подсистемы специализированных задач, и совокупная производительность ИИ начнет расти с тем же высоким темпом, как растет производительность подсистемы универсального мышления, несмотря на то что сила оптимизации остается неизменной, — тогда сопротивляемость системы резко упадет.

Возможно также, что из-за естественной для нас склонности смотреть на интеллект с антропоцентрической точки зрения мы будем недооценивать динамику улучшений в системах, недотягивающих до человеческого уровня, и, таким образом, переоценивать сопротивляемость. Элиезер Юдковский, теоретик искусственного интеллекта, много пишущий о его будущем,

говорит об этом в работе «Искусственный интеллект как позитивный и негативный фактор глобального риска» (см. также рис. 8):

> ИИ может совершить *кажущийся* резким скачок в интеллектуальных способностях исключительно вследствие антропоморфизма, то есть присущей человеку тенденции считать, будто на интеллектуальной шкале располагается не практически бесконечное количество точек, а есть всего две крайние, причем на одной крайней точке находится «деревенский дурачок», а на другой — «Эйнштейн».
>
> Все, что глупее глупого человека, может показаться кому-то просто «глупым». И вот мы как бы двигаемся по интеллектуальной шкале вправо, мимо мышей и шимпанзе, а ИИ все еще остается «глупым», поскольку не говорит свободно на вашем языке и не пишет научные статьи, но потом он внезапно преодолевает крошечный разрыв между недоидиотом и сверх-Эйнштейном за месяц или около того[8].

Итак, все соображения приводят к следующему заключению: очень трудно предсказать, насколько сложно будет усовершенствовать алгоритм первого ИИ, который достигнет примерно общего интеллектуального уровня человека. Можно представить как минимум несколько возможных обстоятельств, при которых сопротивляемость алгоритма будет низка. Но даже если она и очень высока, это не говорит о том, что совокупная сопротивляемость ИИ также обязательно должна быть высокой. Можно легко повысить уровень интеллектуальных способностей системы и без корректировки алгоритмов. Есть еще два фактора, влияющих на него: контент и оборудование.

Рис. 8. Не слишком антропоморфная шкала? Разрыв между идиотом и гением с антропоцентрической точки зрения может показаться очень значительным, но в более широкой перспективе различий между ними почти не видно[9]. Наверняка гораздо сложнее будет создать машину, чей уровень общего интеллекта окажется сравнимым с уровнем интеллекта полного идиота, чем усовершенствовать эту систему, в результате чего она станет намного умнее самого гениального из людей.

Прежде всего поговорим об улучшении контента. Под «контентом» мы понимаем те части кода ИИ, которые не относятся к его главной алгоритмической архитектуре. К контенту могут относиться, например, базы данных сохраненных объектов восприятия, библиотек специализированных навыков и запасы

декларативных знаний. Для многих типов систем грань между алгоритмической архитектурой и контентом довольно размыта, тем не менее она может быть простым способом отметить один потенциально важный источник прироста способностей машинного интеллекта. Можно иначе выразить ту же самую идею: способности системы решать интеллектуальные задачи можно повысить не только сделав систему умнее, но и увеличив объем ее знаний.

Возьмем современные интеллектуальные системы вроде TextRunner (исследовательский проект, который реализуется в Университете Вашингтона) или суперкомпьютера Watson, созданного в IBM (обыграл двух рекордсменов телевизионной игры-викторины Jeopardy!). Они способны извлекать некоторое количество семантической информации, анализируя текст. И хотя эти системы не понимают, что читают, — в том смысле или до такой же степени, как люди, — они тем не менее могут получать значимую информацию из текста, написанного на обычном языке, и использовать ее для получения простых выводов и ответов на вопросы. Еще они могут учиться на своем опыте, усложняя представление концепции по мере того, как сталкиваются с новыми случаями ее использования. Они разработаны так, чтобы большую часть времени работать без вмешательства людей (то есть учиться находить скрытую структуру в неразмеченных данных в отсутствие сигналов, подтверждающих правильность или сообщающих об ошибочности их действий, со стороны человека), причем работать быстро и с возможностью масштабирования. Скажем, TextRunner работает с массивом из пятисот миллионов интернет-страниц[10].

Теперь представим далекого потомка такой системы, способного понимать прочитанное на уровне десятилетнего ребенка, но читающего при этом со скоростью TextRunner. (Вероятно, это ИИ-полная задача.) То есть мы воображаем систему, думающую гораздо быстрее и запоминающую гораздо лучше взрослого человека, но знающую много меньше, — возможно, в результате ее способности будут примерно соответствовать человеческим способностям решать задачи на общем интеллектуальном уровне. Но ее сопротивляемость с точки зрения контента очень низка — достаточно низка, чтобы произошел взлет. За несколько недель система прочитает и обработает весь контент, содержащийся в книгах из библиотеки Конгресса США. И вот она уже и знает больше любого человеческого существа, и думает гораздо быстрее — то есть становится слабым (по меньшей мере) сверхразумом.

Таким образом, система может резко усилить свои интеллектуальные способности, впитав информацию, накопленную за многие века

существования человеческой науки и цивилизации, например получая ее из интернета. Если ИИ достигает человеческого уровня, не имея доступа к этим материалам или не будучи способным их переварить, тогда его общая сопротивляемость будет низка даже в том случае, когда его алгоритмическую архитектуру улучшить довольно трудно.

Концепция сопротивляемости контента также важна с точки зрения создания эмуляционной модели мозга. Работающая с большой скоростью имитационная модель мозга имеет преимущество перед биологическим мозгом не только потому, что решает те же задачи быстрее, но и потому, что аккумулирует более подходящий контент, в том числе релевантные с точки зрения задачи навыки и опыт. Однако, чтобы раскрыть весь потенциал быстрого накопления контента, системе нужны соответствующие большие объемы памяти. Бессмысленно читать подряд энциклопедии, если к тому времени как дойдешь до статьи Abalone (африканский муравьед), забудешь все, что узнал из статьи Aardvark (моллюск). В то время как у систем ИИ, скорее всего, недостатка в памяти не будет, модели мозга могут унаследовать некоторые ограничения от своих биологических оригиналов. И, как следствие, потребуют каких-то архитектурных усовершенствований, чтобы иметь возможность обучаться без ограничений.

До сих пор мы рассматривали сопротивляемость архитектуры и контента — то есть то, насколько может быть сложно улучшить *программное обеспечение* машинного интеллекта, достигшего человеческого уровня интеллекта. Теперь давайте поговорим о третьем пути повышения производительности такого интеллекта, а именно об усовершенствовании его вычислительной части. Какой может быть сопротивляемость улучшению аппаратной основы?

После появления интеллектуальных программ (систем ИИ или имитационных моделей мозга) усилить *коллективный интеллект* можно будет просто за счет использования множества дополнительных компьютеров, на которых запущены их копии[11]. *Скоростной интеллект* можно усилить, перенеся программы на более быстрые компьютеры. В зависимости от того, насколько программы способны работать параллельно, еще один ресурс заключается в использовании большего числа процессоров. Это, скорее всего, подойдет для моделей мозга, которые изначально имеют параллельную архитектуру, но и многие системы ИИ включают в себя процедуры, эффективность выполнения которых повысится за счет параллельного выполнения. Усилить

качественный интеллект за счет наращивания вычислительной мощности, возможно, также удастся, но вряд ли так же прямолинейно[12].

Таким образом, сопротивляемость при усилении коллективного или скоростного (и, возможно, качественного) интеллекта в программах человеческого уровня, скорее всего, будет низкой. Единственной сложностью останется получить доступ к дополнительным вычислительным мощностям. Есть несколько путей расширения аппаратной базы системы, каждый из которых требует различных затрат времени.

В краткосрочной перспективе вычислительная мощность может масштабироваться практически линейно исключительно за счет дополнительного финансирования: увеличится оно вдвое — можно будет купить вдвое больше компьютеров, что позволит запустить вдвое больше программ одновременно. Возникновение облачных вычислительных услуг дает возможность любому проекту увеличивать использование вычислительных ресурсов, даже не теряя времени на доставку дополнительных компьютеров и установку на них ПО, хотя соображения секретности могут подталкивать к работе на собственных машинах. (Однако в некоторых сценариях дополнительные вычислительные ресурсы можно будет получить и иными способами, например рекрутируя ботнеты[13].) То, насколько легко масштабировать ту или иную систему на определенную величину, будет зависеть от объемов изначально используемой ею вычислительной мощности. Систему, изначально работающую на персональном компьютере, можно масштабировать в тысячи раз меньше, чем за миллион долларов. Масштабировать программу, установленную на суперкомпьютере, гораздо дороже.

В чуть более долгосрочной перспективе, по мере того как окажется задействованной все большая часть имеющихся на планете вычислительных мощностей, затраты на приобретение дополнительного оборудования могут начать расти. Например, в сценарии создания имитационной модели мозга в условиях конкуренции затраты на запуск одной дополнительной ее копии должны расти и стать примерно равными выручке, которую эта копия генерирует, поскольку цены на вычислительную инфраструктуру будут подниматься (хотя если эта технология окажется в распоряжении всего одного проекта, он обеспечит себе монопсонию и вследствие этого сможет платить меньше.)

В более долгосрочной перспективе предложение вычислительной мощности будет расти по мере установки все новых ПО в дополнительные

компьютеры. Всплеск спроса стимулирует расширение существующих и создание новых производств микропроцессоров. (Разовый всплеск производительности на один-два порядка величины может случиться в результате использования кастомизированных процессоров[14].) Помимо этого, дополнительный вал вычислительной мощности хлынет на турбины мыслящих машин за счет постоянного усовершенствования технологий, темп которого будет только расти. Исторически он описывается знаменитым законом Мура, один из вариантов которого гласит, что вычислительная мощность в расчете на доллар удваивается каждые восемнадцать месяцев или около того[15]. И хотя никто не может поручиться, что эта скорость сохранится до момента создания ИИЧУ, пространство для совершенствования компьютерных технологий остается — фундаментальные физические пределы еще не достигнуты.

Поэтому есть все основания полагать, что сопротивляемость аппаратной среды не окажется слишком высокой. Приобретение дополнительной вычислительной мощности для системы, доказавшей право на существование тем, что она достигла человеческого интеллектуального уровня, может легко добавить мощности, имеющейся в распоряжении ее создателей, несколько дополнительных порядков величины (в зависимости от того, насколько прожорлив в этом смысле был проект изначально). Кастомизация процессоров добавит еще один-два порядка. На другие средства расширения аппаратной базы, такие как строительство новых заводов и усовершенствование вычислительных технологий, понадобится больше времени — обычно это занимает несколько лет, хотя в будущем срок может резко сократиться в результате революции, которую произведет сверхразум в области развития технологических и производственных процессов.

Подведем итоги. Мы можем говорить о вероятности образования «аппаратного навеса» — к тому моменту, когда будет создано программное обеспечение человеческого интеллектуального уровня, окажется доступно достаточно вычислительной мощности для запуска большого количества его копий на очень быстрых компьютерах. Сопротивляемость на уровне программного обеспечения, как уже было сказано выше, оценить сложнее, но, вполне вероятно, она может оказаться даже более низкой, чем сопротивляемость аппаратная. В частности, есть возможность образования «контентного навеса» в форме огромных объемов готовой к использованию информации (например, в интернете), которая станет доступна системе,

как только она достигнет человеческого уровня развития. Вероятность «алгоритмического навеса» — проведенной заранее оптимизации алгоритмов — также существует, но гораздо более низкая. За счет совершенствования программного обеспечения (и алгоритмов, и контента) потенциальный рост производительности системы может составить несколько порядков величины, и добиться этого роста, после того как цифровой разум достигнет человеческого уровня, окажется довольно легко, причем он наложится на выигрыш в производительности, полученный за счет использования большего количества или более мощных аппаратных средств.

Сила оптимизации и взрывное развитие интеллекта

Изучив вопрос сопротивляемости, обратимся ко второй части нашего уравнения — *силе оптимизации*. Напомним:

$$\text{Скорость роста интеллекта} = \frac{\text{Сила оптимизации}}{\text{Сопротивляемость}}$$

Как видно из этой формулы, быстрый взлет не требует, чтобы сопротивляемость на фазе перехода была низкой. Быстрый взлет также может произойти на фоне неизменной или даже медленно растущей сопротивляемости, при условии, что сила оптимизации, приложенная к системе и повышающая ее производительность, растет довольно быстро. Мы увидим, что есть все основания считать: прилагаемая к системе сила оптимизации *действительно* будет расти на этапе перехода, по крайней мере в отсутствие явных мер, направленных против этого.

Здесь можно выделить две фазы. Первая фаза начинается с точки отрыва системы от уровня человеческого интеллекта. Поскольку ее возможности продолжают расти, она может использовать их часть или даже все возможности для самосовершенствования (или для создания системы-потомка, что неважно). Однако большая часть оптимизирующей силы будет приложена извне системы благодаря работе программистов и инженеров, как занятых в проекте, так и не имеющих к нему прямого отношения[16]. Если эта фаза растянется на долгое время, можно ожидать, что сила оптимизации, приложенная к системе, будет увеличиваться. Вклад в проект и команды, и участников извне станет расти, если выбранный подход покажет свою перспективность. Исследователи начнут работать усерднее, их количество вырастет, чтобы ускорить прогресс, начнут покупать больше компьютеров. Рост будет особенно быстрым, если создание ИИЧУ окажется для мира неожиданностью — в этом случае на изначально

небольшом проекте будут сосредоточены усилия исследователей и разработчиков всего мира (хотя какая-то часть мировых усилий может быть направлена на реализацию конкурирующих проектов).

Вторая фаза роста начнется в тот момент, когда система аккумулирует такую мощность, что превратится в основной источник оптимизирующей силы по отношению к самой себе (на рис. 7 это отмечено как «критический рубеж, или точка перехода»). Это резко изменит динамику процесса, поскольку любые изменения возможностей системы теперь приводят к пропорциональному росту оптимизирующей силы, приложенной с целью ее дальнейшего совершенствования. Если сопротивляемость остается постоянной, начинается рост в геометрической прогрессии (см. врезку 4). Удвоение эффективности может происходить очень быстро — есть сценарии, где на это требуется всего нескольких секунд — если рост идет с электронными скоростями за счет использования алгоритмических усовершенствований или эксплуатации контентного или аппаратного навеса[17]. Для роста, базой которого является физическая инфраструктура, скажем, создание новых компьютеров или производственных мощностей, потребуется несколько больше времени (и все-таки оно будет совсем незначительным, если брать во внимание текущие темпы роста мировой экономики).

Таким образом, вполне вероятно, что в процессе перехода сила оптимизации будет расти — вначале потому, что люди будут стараться быстрее улучшить показатели искусственного интеллекта, демонстрирующего выдающиеся успехи, а позднее из-за того, что он и сам окажется в состоянии обеспечивать прогресс уже на цифровых скоростях. Это создаст реальные предпосылки для быстрого или умеренного взлета, *даже если сопротивляемость будет постоянна или немного вырастет в зоне человеческого интеллектуального уровня*[18].

ВРЕЗКА 4. О ДИНАМИКЕ ВЗРЫВНОГО РАЗВИТИЯ ИНТЕЛЛЕКТА

Скорость изменения уровня интеллекта можно выразить в виде соотношения силы оптимизации, прикладываемой к системе, и ее сопротивляемости:

$$\frac{dI}{dt} = \frac{\mathcal{O}}{\mathfrak{R}}$$

Совокупная сила оптимизации, действующая на систему, складывается из силы оптимизации, которую производит сама система, и приложенной извне. Например, развивать зародыш ИИ можно за счет комбинации его собственных усилий и усилий

команды программистов-людей, возможно, с привлечением широкого глобального сообщества исследователей, постоянно работающих над прогрессом в отрасли производства полупроводников, компьютерных науках и связанных с ними областях[19]:

$$\text{Ф} = \text{Ф}_{\text{система}} + \text{Ф}_{\text{проект}} + \text{Ф}_{\text{мир}}$$

Изначально зародыш ИИ обладает очень ограниченными когнитивными способностями. То есть в начальной точке величина $\text{Ф}_{\text{система}}$ мала[20]. А как насчет $\text{Ф}_{\text{проект}}$ и $\text{Ф}_{\text{мир}}$? Бывают случаи, когда возможности проекта превышают возможности всего остального мира, как было, например, с Манхэттенским проектом, когда большинство лучших физиков мира оказались в Лос-Аламосе и приняли участие в создании атомной бомбы. Но чаще на любой отдельно взятый проект приходится лишь очень небольшая доля общемировых исследовательских возможностей. Однако даже когда возможности мира намного превышают возможности проекта, $\text{Ф}_{\text{проект}}$ может все-таки превышать $\text{Ф}_{\text{мир}}$, поскольку большинство исследователей за пределами проекта сосредоточены на других направлениях работы. Если проект начинает выглядеть многообещающим — что происходит, когда система достигает человеческого интеллектуального уровня или даже раньше, — он привлекает дополнительные инвестиции, что повышает $\text{Ф}_{\text{проект}}$. Если достижения проекта становятся достоянием широкой общественности, величина $\text{Ф}_{\text{мир}}$ также повышается, поскольку растет интерес к искусственному интеллекту в целом и в игру включаются другие участники. Таким образом, на переходном этапе совокупная сила оптимизации, действующая на когнитивную систему, скорее всего, будет расти по мере роста возможностей системы[21].

Рост возможностей системы может дойти до такой точки, где сила оптимизации, которую производит сама система, начинает доминировать над силой оптимизации, приложенной к системе извне (по всем существенным направлениям усовершенствования):

$$\text{Ф}_{\text{система}} > \text{Ф}_{\text{проект}} + \text{Ф}_{\text{мир}}$$

Этот *рубеж* очень важен, поскольку за ним дальнейшее совершенствование системы оказывает весьма сильное влияние на рост совокупной силы оптимизации, приложенной к системе с целью ее совершенствования. То есть включается режим мощного рекурсивного самосовершенствования. Это приводит к взрывному росту возможностей системы в широком диапазоне форм кривых сопротивляемости.

Чтобы проиллюстрировать это, рассмотрим первый сценарий, в котором сопротивляемость постоянна, вследствие чего возможности ИИ растут пропорционально приложенной к нему силе оптимизации. Предположим, что вся сила оптимизации генерируется самой системой и что все интеллектуальные возможности системы направлены на решение задачи совершенствования ее интеллекта, так что $\text{Ф}_{\text{система}} = I$[22]. Тогда мы имеем:

$$\frac{dI}{dt} = \frac{I}{k}$$

Решая это простое дифференциальное уравнение, получаем экспоненциальную функцию:

$$I = Ae^{t/k}$$

Но постоянство сопротивляемости — особый случай. Сопротивляемость вполне может снизиться в зоне человеческого интеллектуального уровня в силу одного или нескольких факторов, рассмотренных в предыдущем разделе, и оставаться низкой в точке перехода и некоторое время после нее (возможно, до тех пор, пока система в конечном счете не упрется в фундаментальные физические ограничения). Предположим, что приложенная к системе сила оптимизации остается примерно постоянной (то есть $O_{проект} + O_{мир} \approx c$), пока система не получает возможность сама существенно менять свой дизайн, что приводит к удвоению ее возможностей через каждые 18 месяцев. (Это примерно соответствует историческим темпам совершенствования систем, если объединить закон Мура и прогресс в области создания ПО[23].) Такие темпы совершенствования, достигнутые за счет действия примерно постоянной силы оптимизации, приводят к тому, что сопротивляемость снижается обратно пропорционально силе системы:

$$\frac{dI}{dt} = \frac{c}{1/I} = cI$$

Если сопротивляемость продолжает снижаться по такой гиперболе, то когда ИИ достигнет точки перехода, совокупная сила оптимизации, действующая на систему, удвоится. То есть:

$$\frac{dI}{dt} = \frac{(c + I)}{(1/I)} = (c + I)I$$

Следующее удвоение произойдет через 7,5 месяца. В течение 17,9 месяца возможности системы вырастут в тысячу раз, превратив ее в быстрый сверхразум (см. рис. 9).

Эта конкретная траектория роста имеет точку положительной сингулярности в момент $t = 18$ месяцев. В реальности предположение, что сопротивляемость является константой, перестанет выполняться по мере приближения системы к физическим границам обработки информации, если не раньше.

Эти два сценария приведены лишь в качестве иллюстрации: в зависимости от формы кривой сопротивляемости возможно множество других траекторий. Суть проста: мощная обратная связь, возникающая в районе точки перехода, должна привести к более быстрому взлету, чем в ее отсутствие.

В предыдущем разделе мы видели, что есть факторы, способные привести к резкому снижению сопротивляемости. К таким факторам относятся, например, возможность быстрого расширения аппаратной основы, как только будет создано работающее интеллектуальное ПО; возможность алгоритми-

ческих улучшений; возможность сканирования дополнительных экземпляров мозга (при полной эмуляции головного мозга); возможность быстрого усвоения огромного объема контента за счет тщательного обследования интернета (в случае искусственного интеллекта)[24].

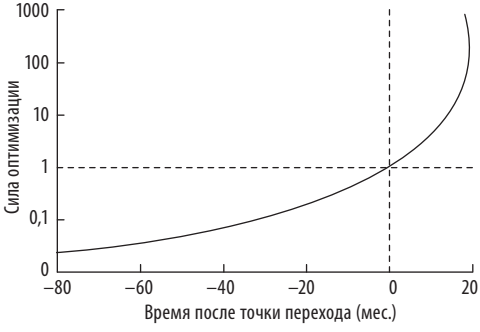

Рис. 9. Одна из простых моделей взрывного развития интеллекта

Впрочем, нужно сказать, что форму кривой сопротивляемости в каждой конкретной области характеризовать довольно сложно. В частности, неясно, насколько трудно будет повысить качество ПО имитационной модели человеческого мозга или ИИ. Нет также ясности и со степенью сложности расширения аппаратной базы подобных систем.

Хотя сегодня сравнительно легко увеличить вычислительную мощность, доступную небольшому проекту, — просто потратив на компьютеры в тысячу раз больше или подождав несколько лет падения цен на них, — вполне вероятно, что первый машинный интеллект, достигший человеческого уровня, будет создан в рамках крупного проекта с применением дорогих сверхмощных компьютеров, которые не слишком просто масштабировать, и что к этому времени перестанет действовать закон Мура. По этим причинам не следует исключать возможность медленного взлета, хотя вероятность быстрого или умеренного взлета представляется все-таки более высокой[25].

Глава пятая

Решающее стратегическое преимущество

Помимо вопроса о динамике процесса появления сверхразума, есть еще один, с ним связанный: будет ли создан единственный сверхразум или множество? Наступит ли власть единственного сверхразума или власть многих? Взрывное развитие искусственного интеллекта приведет к тому, что лишь один проект вырвется далеко вперед и продиктует людям их будущее? Или прогресс пойдет более равномерным путем, разворачиваясь широким фронтом, коснется многих разработок, ни одна из которых не обеспечит себе безоговорочного верховенства?

В предыдущей главе был проанализирован один параметр, ключевой с точки зрения определения масштаба возможного разрыва между проектом-лидером и его ближайшими конкурентами, а именно: скорость перехода от интеллекта человеческого уровня к сверхразуму. Она определяет очень многое. Если *взлет быстрый* (происходит в течение часов, дней или недель), тогда маловероятно, что два независимых проекта осуществят его одновременно: почти наверняка один из проектов завершит свой взлет прежде, чем кто-то еще перейдет к этой стадии. Если *взлет медленный* (тянется на протяжении многих лет или десятилетий), тогда, вполне возможно, будет несколько проектов, перешедших к этой стадии одновременно, и несмотря на огромные возможности, которые появятся у проектов, завершивших переход, ни у одного из них не будет довольно времени на то, чтобы обойти остальные и получить безусловное преимущество. *Умеренный взлет* попадает в промежуточную категорию, где вероятны любые варианты: не исключено, что на стадии взлета окажется одновременно несколько проектов, но возможно — всего один[1].

Способен ли какой-то из проектов настолько сильно оторваться от конкурентов, что получит *решающее стратегическое преимущество* — преимущество технологического или любого иного рода, но достаточное для утверждения мирового господства? И далее: получив абсолютное стратегическое преимущество, станет ли этот проект игроком-одиночкой, то есть синглтоном, начнет ли подминать мир под себя, убирая с пути всех конкурентов и формируя новое мироустройство, при котором только он будет принимать решения? Если объявится такой проект-победитель, насколько «великим» он должен быть? Причем не с точки зрения физического размера или величины бюджета, а с точки зрения размаха человеческих желаний, предпочтений и замыслов. Рассмотрим эти вопросы по порядку.

Получит ли лидирующий проект абсолютное преимущество?

Одним из факторов, влияющих на величину разрыва между идущим впереди и следующими за ним, является степень расширения того (может быть все что угодно), что обеспечивает лидеру конкурентное преимущество. Лидеру иногда трудно оторваться от преследователей, если они могут легко копировать его идеи и новшества. Появление суррогатов порождает реальные препятствия, и этот встречный ветер бьет прямо в лицо вырвавшемуся вперед и играет на руку тем, кто находится сзади, особенно если речь идет об интеллектуальной деятельности, права на которую защищены слабо. Кроме того, лидер может быть особенно уязвим с точки зрения отчуждения собственности, повышенного налогообложения или разделения в рамках антимонопольного регулирования.

Было бы ошибкой считать, что сдерживающие факторы должны монотонно возрастать по мере увеличения разрыва между лидером и идущими следом. Разработчик, во всем уступающий технологическому лидеру, может столкнуться с трудностями при попытке воспользоваться его проектами, стоящими на передовом крае науки, — ситуация вполне сравнимая с велогонщиком, сильно отставшим от основной массы соперников и оказавшимся незащищенным от порывов встречного ветра[2]. Бывает, что разрыв между интеллектуальными и материальными возможностями оказывается слишком большим. Лидер, например, может себе позволить перейти на новейшую технологическую платформу, после чего уже никто из отстающих не будет в состоянии воплотить на своих примитивных платформах его дальнейшие инновации. Когда какой-либо проект вырывается далеко вперед, следует

исключить информационные утечки, обеспечить секретность наиболее важных разработок и препятствовать усилиям конкурентов развить собственные материальные возможности.

Если проектом-лидером является ИИ, он может обладать такими атрибутами, которые позволят ему легче развивать свои возможности и при этом снизить риск распространения информации. В организациях, которыми управляют люди, преимущества от экономии на масштабе уравновешиваются бюрократической неэффективностью и агентскими проблемами, в том числе трудностью сохранения коммерческих секретов[3]. Скорее всего, эти проблемы будут тормозить развитие проектов искусственного интеллекта до тех пор, пока разработками управляют люди. Однако системы ИИ могут избежать некоторых недостатков, связанных с увеличением масштаба проекта, поскольку работники-эмуляции (в отличие от работников-людей) не имеют интересов, отличных от интересов системы в целом. Следовательно, системы ИИ смогут избежать заметного снижения эффективности, характерного для проектов, в которых заняты люди. Благодаря тому же преимуществу — наличию абсолютно лояльных компонентов — системам ИИ гораздо легче обеспечивать секретность. В их случае невозможно появление обиженных сотрудников, готовых перейти к конкуренту или продать коммерческую информацию[4].

О том, каким может быть отрыв при разработке различных систем, можно получить представление, обратившись к историческим примерам (см. врезку 5). Похоже, для стратегически важных технологических проектов типичным является диапазон от нескольких месяцев до нескольких лет.

ВРЕЗКА 5. ТЕХНОЛОГИЧЕСКАЯ ГОНКА — НЕСКОЛЬКО ИСТОРИЧЕСКИХ ПРИМЕРОВ

На длинной исторической временной шкале хорошо заметно, как увеличивалась скорость распространения знаний и технологий по всему миру. В результате разрыв между технологическими лидерами и их ближайшими преследователями постоянно сужался.

Свыше двух тысячелетий Китай смог удерживать монополию на производство шелка. Археологические находки говорят о том, что он мог возникнуть около 3000 года до н. э. или даже раньше[5]. Шелководство было тщательно охраняемым секретом. Разглашение его каралось смертной казнью, как и вывоз шелковичных червей и их яиц за пределы страны. Римляне, несмотря на высокую цену импортного шелка на просторах их империи, так и не овладели искусством шелководства. Только около

300 года н. э. одна японская экспедиция смогла захватить несколько яиц шелкопряда и четырех китайских девушек, которых похитители заставили раскрыть секреты мастерства[6]. Византия присоединилась к клубу производителей шелка только в 522 году. Для истории с производством фарфора тоже характерны длинные временные лаги. Это искусство практиковалось в Китае уже во времена династии Тан, то есть минимум с 600 года (а могло быть известно и раньше, с 200 года), но стало доступно европейцам только в XVII веке[7]. Колесные повозки появились в нескольких местах Европы и Месопотамии около 3500 года до н. э., но достигли Америки только после ее открытия Колумбом[8]. Если говорить о более важных событиях, то людям потребовались десятки тысяч лет, чтобы расселиться на большей части земного шара; неолитическая революция шла несколько тысяч лет; промышленная революция — всего сотни лет; информационная революция — десятилетия, хотя, конечно, эти изменения не обязательно были одинаково глубокими. (Видеоигра Dance Dance Revolution появилась в Японии и распространилась в Европе и Северной Америке всего за год!)

Технологическая конкуренция изучается очень интенсивно, особенно в контексте патентной гонки и гонки вооружений[9]. Обзор соответствующей литературы выходит за рамки темы этой книги. Однако будет полезно взглянуть на некоторые примеры стратегически важных технологических схваток XX века (см. табл. 7).

В случае шести главных технологий, считавшимися сверхдержавами стратегически важными из-за их военного или символического потенциала, разрыв между лидером и его ближайшим преследователем составлял (очень приблизительно) 49 месяцев, 36 месяцев, 4 месяца, 1 месяц, 4 месяца и 60 месяцев соответственно — это больше, чем длительность быстрого взлета, но меньше, чем медленного[10]. Во многих случаях проекты-преследователи получили преимущество за счет шпионажа и использования информации, оказавшейся в открытом доступе. Сама демонстрация возможности изобретения часто служила стимулом остальным для продолжения работ по его независимой разработке, а страх отстать оказывался дополнительным мощным стимулом.

Возможно, ближе всего к ситуации с ИИ находятся математические открытия, не требующие создания новой материальной инфраструктуры. Часто для признания их общедоступными довольно публикации в научной литературе, но в некоторых случаях, когда открытие дает какое-то стратегическое преимущество, сроки публикации переносятся. Скажем, в криптографии с открытым ключом есть две наиболее важные концепции: протокол Диффи–Хеллмана, позволяющий двум и более сторонам получать общий секретный ключ, и схема шифрования RSA. Они были представлены научному сообществу в 1976 и 1978 гг. соответственно, но позднее оказалось, что криптографы в группе секретной связи Великобритании знали о них с начала 1970-х гг.[11] Крупные проекты по созданию ПО могут быть близким аналогом проектов в области ИИ, но привести убедительные примеры типичного лага нелегко, поскольку ПО обычно много раз обновляется и дополняется, а функциональность конкурирующих систем часто невозможно сравнить напрямую.

Таблица 7. Стратегически важные гонки технологий

	США	СССР	Великобритания	Франция	Китай	Индия	Израиль	Пакистан	Северная Корея	ЮАР
Атомная бомба	1945	1949	1952	1960	1964	1974	1979?	1998	2006	1979?
Водородная бомба	1952	1953[12]	1957	1968	1967	1998	?	–	–	–
Беспилотный космический аппарат	1958	1957	1971	1965	1970	1980	1988	–	1998?[13]	–[14]
Пилотируемый космический аппарат	1961	1961	–	–	2003	–	–	–	–	–
Межконтинентальная баллистическая ракета (МБР)[15]	1959	1960	1968[16]	1985	1971	2012	2008	–[17]	2006	–[18]
Многозарядные боеголовки индивидуального наведения[19]	1970	1975	1979	1985	2007	2014[20]	2008?	–		

Возможно, что глобализация и политика сдерживания сократят типичные отставания между реализацией конкурирующих технологических проектов. Хотя, когда нет сознательного сотрудничества, нижняя граница у средней задержки все-таки существует[21]. Даже при отсутствии заметной динамики, способной привести к эффекту снежного кома, у каких-то проектов всегда будет более сильная команда исследователей, более талантливые лидеры, более мощная инфраструктура или просто блестящие идеи. Если два проекта разрабатывают разные подходы к решению одной задачи и один из них оказывается более удачным, проекту-конкуренту для переключения на него может понадобиться много месяцев — даже при возможности слежки за действиями соперника.

На основании наших предыдущих наблюдений относительно природы скорости взлета можно сделать соответствующие выводы: вероятность одновременного осуществления быстрого взлета сразу силами двух проектов крайне мала; в случае умеренного взлета это вполне возможно; в случае медленного взлета почти наверняка в процессе будут участвовать параллельно несколько проектов. Но анализ на этом не заканчивается. Вопрос не в том, сколько проектов отправится в путь одновременно. Главный вопрос: сколько из них, обладая приблизительно одинаковыми возможностями,

окажется на другом берегу, причем так, чтобы ни один из них не смог получить решающего конкурентного преимущества. Если процесс взлета начнется относительно медленно, а потом ускорится, дистанция между проектами-конкурентами начнет расти. Возвращаясь к нашей аналогии с велосипедной гонкой, представим, что два гонщика взбираются вверх по склону холма, и если один едет чуть позади, то разрыв между ними увеличится, как только лидер перевалит через вершину и покатится вниз.

Рассмотрим следующий сценарий умеренного взлета. Предположим, ИИ нужен год, чтобы поднять свои возможности с человеческого уровня до сверхразумного, причем второй проект вошел в стадию взлета спустя шесть месяцев после первого. Может показаться, что ни один из них не имеет решающего конкурентного преимущества. Но это не обязательно так. Допустим, потребуется девять месяцев, чтобы пройти путь от уровня человека до точки перехода, и еще три месяца — чтобы достигнуть уровня сверхразума. Тогда система-лидер превратится в сверхразум за три месяца до того, как ее соперник достигнет хотя бы точки перехода. Это обеспечит идущему впереди решающее конкурентное преимущество и даст возможность, умело использовав свое лидерство, установить неусыпный и повсеместный контроль. В итоге проекты-конкуренты погублены и установлен режим проекта-одиночки, или синглтона. (Заметим, что синглтон является чистой абстракцией. Режим при нем может быть самым разным: демократия; тирания; единодержавие, то есть абсолютизм ИИ; набор жестких общемировых законов, подлежащих безотказному исполнению. Иначе говоря, синглтон — это некий орган, способный в одиночку решать все крупные проблемы общемирового масштаба. Он может нести в себе — но совсем не обязательно — характеристики знакомых нам форм человеческого управления[22].)

Поскольку имеются серьезные перспективы взрывного роста сразу после точки перехода, когда сила оптимизации включит мощную обратную связь, описанный выше сценарий имеет большую вероятность реализации, увеличивая шансы того, что лидирующий проект получит абсолютное конкурентное преимущество даже в случае не самого быстрого взлета.

Насколько крупным будет самый перспективный проект?

Некоторые пути появления сверхразума требуют значительных ресурсов и поэтому, скорее всего, будут реализованы в рамках крупных, хорошо финансируемых проектов. Например, для полной эмуляции головного мозга

потребуется большой разноплановый профессиональный опыт и разнообразное оборудование. Биологическое улучшение интеллектуальных способностей и нейрокомпьютерные интерфейсы также сильно зависят от материального фактора. И хотя небольшая биотехнологическая фирма способна изобрести пару препаратов, однако для появления сверхразума (если такое вообще возможно, говоря о конкретных методах) наверняка потребуется много изобретений и экспериментов, что возможно при поддержке целого сектора промышленности и хорошо финансируемой национальной программы. Достижение уровня коллективного сверхразума за счет повышения эффективности организаций и сетей потребует еще более весомого вклада и вовлечения большей части мировой экономики.

Труднее всего оценить путь ИИ. Возможны разные варианты: от крупнейшей и солиднейшей исследовательской программы до небольшой группы специалистов. Нельзя исключать даже сценариев с участием хакеров-одиночек. Для создания зародыша ИИ могут потребоваться идеи и алгоритмы, которые будут разрабатывать лучшие представители мирового научного сообщества на протяжении нескольких десятилетий. При этом последняя решающая догадка осенит одного человека или нескольких людей — и они смогут свести все воедино. Для одних типов архитектуры ИИ этот сценарий более вероятен, для других — менее. Система из большого количества модулей, каждый из которых нужно заботливо настраивать и балансировать для эффективной совместной работы, а затем тщательно загружать индивидуально разработанным когнитивным контентом, скорее всего, потребует реализации в рамках очень крупного проекта. Если зародыш ИИ может быть простой системой, для строительства которой нужно соблюсти лишь несколько основных правил, тогда она под силу небольшой команде и даже ученому-одиночке. Вероятность того, что последний прорыв будет сделан в рамках маленького проекта, возрастает, если большая часть пройденного ранее пути освещалась в открытых источниках или реализована в виде программного обеспечения с открытым кодом.

Нужно разделять два вопроса: 1) насколько крупным должен быть проект, в рамках которого *создается* система; 2) насколько большой должна быть группа, которая *контролирует*, будет ли создана система, и если да, то как и когда. Атомную бомбу создавало множество ученых и инженеров. (Максимальное число участников Манхэттенского проекта достигало ста тридцати тысяч человек, большую часть их составляли технические

работники[23].) Однако и ученых, и инженеров, и конструкторов, и строителей контролировало правительство США, в конечном счете подотчетное американским избирателям, составлявшим в то время примерно одну десятую часть всего населения мира[24].

Система контроля

Учитывая чрезвычайную важность такого вопроса, как появление сверхразума, с точки зрения государственной безопасности, правительства, скорее всего, будут стараться национализировать любой реализующийся на их территории проект, который, по их мнению, окажется близок к стадии взлета. Мощные государства могут также пытаться присвоить проекты, начатые в других странах, — посредством шпионажа, кражи, похищения людей, взяток, военных операций или иных доступных им средств. Если не получится присвоить, будут стараться уничтожить их, особенно если страны, где они реализуются, неспособны эффективно защитить себя. Если к моменту, когда взлет окажется неизбежным, достаточно окрепнут общемировые структуры управления, может оказаться, что многообещающие проекты будут взяты под международный контроль.

Важно, однако, смогут ли национальные или международные органы заметить приближающееся взрывное развитие искусственного интеллекта. Пока непохоже, чтобы спецслужбы уделяли какое-то внимание потенциально успешным проектам в области разработок ИИ[25]. Можно предположить, что не делают они этого в силу широко распространенного мнения, будто у взрывного развития нет никаких перспектив. Если когда-нибудь среди ведущих ученых возникнет полное согласие по проблеме сверхразума и однажды все дружно заявят, что он вот-вот появится, — тогда, вероятно, основные разведки мира возьмут под свое наблюдение научные коллективы и отдельных исследователей, участвующих в работах по этим направлениям. После чего любой проект, способный показать заметный прогресс, может быть очень быстро национализирован. Если политические элиты посчитают, что риск слишком велик, частные усилия в соответствующих областях могут подвергнуться регулированию или будут вовсе запрещены.

Насколько сложно организовать такую систему контроля? Задача облегчится, если присматривать только за лидерами. В этом случае может оказаться достаточным наблюдать лишь за несколькими проектами, лучше всего обеспеченными ресурсами. Если цель состоит в том, чтобы предот-

вратить проведение любых работ (как минимум за пределами специально одобренных организаций), тогда слежка должна быть более тотальной, ведь определенного прогресса способны добиться даже небольшие команды и отдельные исследователи.

Будет легче контролировать проекты, требующие значительного объема материальных ресурсов, как в случае полной эмуляции головного мозга. Напротив, труднее поддаются контролю исследования в области ИИ, поскольку для них нужен лишь компьютер, а также ручка и лист бумаги — с их помощью ведутся наблюдения и делаются теоретические выкладки. Но и в этом случае несложно выявить наиболее талантливых ученых с серьезными долгосрочными замыслами в области ИИ. Обычно такие люди оставляют видимые следы: публикации научных работ; участие в конференциях; комментарии в интернет-форумах; премии и дипломы ведущих организаций в области компьютерных наук. Исследователи, как правило, публично общаются друг с другом на профессиональном уровне, что позволяет идентифицировать их, составив социальный граф.

Сложнее будет обнаружить проекты, изначально предполагающие секретность. В качестве их ширмы может выступать идея разработки какого-нибудь обычного ПО[26]. И только тщательный анализ кода позволит раскрыть истинную сущность проблем, над которыми трудятся программисты. Такой анализ потребует привлечения большого числа очень квалифицированных и очень опытных специалистов, поэтому вряд ли получится досконально исследовать все подряд. В такого рода наблюдениях задачу серьезно облегчит появление новейших технологий, вроде использования детекторов лжи[27].

Есть еще одна причина, по которой государства могут не справиться с выявлением многообещающих проектов, поскольку некоторые научные и технические открытия, в принципе, трудно прогнозировать. Скорее всего, это относится к исследованиям в области ИИ, поскольку революционным изменениям в компьютерном моделировании мозга будет предшествовать целая серия успешных экспериментов.

Возможно, что присущие спецслужбам и прочим государственным институтам бюрократические неповоротливость и негибкость не позволят им распознать значимость некоторых достижений — очевидную для профессионалов.

Особенно высоким может оказаться барьер на пути официального признания возможности взрывного развития интеллекта. Вполне допустимо,

что этот вопрос станет катализатором религиозных и политических разногласий и в итоге в некоторых странах превратится в табу для официальных лиц. Более того, тему взрывного развития начнут ассоциировать с дискредитировавшими себя персонажами, шарлатанами и пустозвонами, в результате чего ее начнут избегать уважаемые ученые и политики. (В первой главе вы уже читали о подобном эффекте, когда нечто подобное случалось уже минимум дважды — вспомните две «зимы искусственного интеллекта».) К этой кампании дискредитации обязательно присоединятся лоббисты, не желающие, чтобы на выгодное дело была брошена тень подозрений, и сообщества ученых, готовых вышвырнуть из своих рядов любого, кто осмелится озвучить собственные опасения по поводу «замораживания» столь важной проблемы и последствий этого для самой науки[28].

Итак, полного поражения спецслужб в этом вопросе исключать нельзя. Особенно высока его вероятность, если какое-либо судьбоносное открытие произойдет в ближайшем будущем, пока тема еще не привлекла внимания широкой общественности. Но даже если специалисты разведывательных управлений все понимают правильно, политики просто отмахнутся от их советов. Знаете ли вы, что запуск Манхэттенского проекта потребовал буквально нечеловеческих усилий, предпринятых несколькими физиками-провидцами, среди которых были Марк Олифант и Лео Силард — именно последний упросил Юджина Вигнера, чтобы тот убедил Альберта Эйнштейна, чтобы тот подписал письмо с целью убедить президента США Франклина Рузвельта обратить внимание на эту проблему? И даже когда Манхэттенский проект шел полным ходом, и Рузвельт, и потом Гарри Трумен сомневались в его необходимости и скептически относились к возможности его осуществления.

Хорошо это или плохо, но, скорее всего, небольшим группам энтузиастов будет трудно повлиять на последствия создания искусственного интеллекта, если в этом направлении начнут активно действовать крупные игроки, включая государства. Таким образом, повлиять на снижение общего уровня риска для человечества в результате взрывного развития искусственного интеллекта активисты, вероятно, смогут в двух случаях: если тяжеловесы останутся в стороне от процесса и если на ранних стадиях активисты смогут определить, где, когда, как и какие именно крупные игроки вступят в игру. Поэтому тем, кто желал бы сохранить за собой максимальное влияние,

следует сосредоточить силы на тщательном планировании своих действий до самого конца, даже при понимании высокой вероятности сценария, по которому в конечном счете крупные игроки все заберут себе.

Международное сотрудничество

Координация на международном уровне более вероятна, если будут сильны институты общемирового управления. Ее шансы увеличит как широкое признание самого факта, что взрывное развитие искусственного интеллекта действительно может иметь место, так и мнение, будто действенное наблюдение за всеми серьезными проектами вполне реализуемо. Последнее предположение, как мы уже видели, весьма сомнительно, однако международное сотрудничество все-таки осуществимо. Страны способны объединяться для поддержки совместных разработок. Если у таких проектов достаточно ресурсов, появляются отличные перспективы, что они первыми достигнут цели, особенно если конкуренты или лишены крепкой материальной базы, или вынуждены вести свои разработки в условиях секретности.

Известны прецеденты успешного крупномасштабного международного научного сотрудничества, такие как проект создания и эксплуатации Международной космической станции, исследования генома человека и Большого адронного коллайдера[29]. Однако основным мотивом для сотрудничества во всех этих проектах было разделение затрат на них. (В случае МКС важной целью было также развитие партнерства между Россией и США[30].) Намного труднее работать в тесном сотрудничестве над проектами, которые важны с точки зрения безопасности. Если страна считает, что способна достичь цели самостоятельно, у нее возникнет большой соблазн двигаться по этому пути без партнеров. Еще одна причина уклониться от участия в международном сотрудничестве — страх, что кто-то из партнеров воспользуется общими идеями для ускорения своего тайно реализуемого национального плана.

Таким образом, чтобы начать международный проект, придется решать серьезные проблемы с безопасностью, кроме того, потребуется значительное доверие его участников друг к другу и время, чтобы это доверие возникло. Вспомните, что даже в условиях потепления отношений между США и СССР, во времена Михаила Горбачева, усилия по снижению количества вооружений — которое было в интересах обоих государств — давали не очень заметный результат. Горбачев стремился к резкому снижению запасов ядерного оружия, но переговоры застопорились из-за проблемы «звездных войн»

Рональда Рейгана, то есть стратегической оборонной инициативы (СОИ), против реализации которой активно выступал Кремль. На встрече в верхах в Рейкьявике в 1986 году Рейган выступил с предложением, что США поделится с СССР технологией, лежащей в основе программы, таким образом, обе страны могли бы получить защиту от случайных пусков и нападения со стороны более мелких стран, способных создать ядерное оружие. Однако Горбачева это взаимовыгодное предложение не устроило. Согласиться на него, по его мнению, было бы ошибкой, при этом Горбачев не принял во внимание, что Америка была готова поделиться с Советским Союзом плодами ультрасовременных военных разработок в то время, когда под запретом была передача СССР даже технологий машинного доения[31]. Трудно сказать, был ли искренен Рейган в своем щедром предложении сотрудничества другой сверхдержаве, но долгое взаимное недоверие не позволило Горбачеву принять его дар.

Легче сотрудничать, будучи союзниками, но и в таком случае не всегда удается автоматически добиться доверия. Во времена Второй мировой войны Советский Союз и Соединенные Штаты вместе боролись с общим врагом — нацистской Германией, — и тем не менее США скрывали от союзника свое намерение создать атомную бомбу. В Манхэттенском проекте они сотрудничали с Великобританией и Канадой[32]. Точно так же Великобритания утаивала от СССР, что ее ученым удалось раскрыть секрет немецкого шифра «Энигма», но поделилась этим — правда, не сразу — с США[33]. Вывод напрашивается сам собой: прежде чем думать о международном сотрудничестве в работе над технологиями, имеющими критически большое значение для национальной безопасности, необходимо построить близкие и доверительные отношения между странами.

К вопросу о желательности и возможности международного сотрудничества в развитии технологий повышения интеллектуальных способностей мы вернемся в четырнадцатой главе.

От решающего преимущества — к синглтону

Станет ли проект, получивший решающее стратегическое преимущество, использовать его для формирования синглтона?

Возьмем довольно близкий исторический аналог. США создали ядерное оружие в 1945 году. Это была единственная ядерная мощь на планете вплоть до 1949 года, когда атомная бомба появилась в СССР. За этот временной

промежуток, и даже чуть дольше, Соединенные Штаты могли иметь — или были в состоянии обеспечить себе — решающее военное преимущество.

Теоретически США могли использовать свою монополию для создания синглтона. Один из путей добиться этого — бросить все на создание ядерного арсенала, а затем угрожать нанесением (а если потребуется, привести угрозу в действие) атомного удара для ликвидации промышленных возможностей по запуску программы разработки аналогичного оружия в СССР и любой другой стране мира, которая решит его создать.

Более позитивный, и тоже вполне осуществимый, сценарий — использовать свой ядерный потенциал в качестве козыря с целью добиться создания мощного международного правительства, своего рода ООН без права вето, обладающего ядерной монополией и облеченного правами предпринимать любые необходимые действия для предотвращения его производства любой страной мира в собственных целях.

В те времена звучали предложения идти обоими путями. Жесткий подход, предполагающий нанесение упреждающего ядерного удара или угрозу его нанесения, поддержали такие известные интеллектуалы, как Бертран Рассел (кстати, британский математик и философ до этого занимал антивоенную позицию, да и потом на протяжении десятков лет участвовал в кампаниях за ядерное разоружение) и Джон фон Нейман (американский математик и физик, кроме того что был одним из создателей теории игр, стал архитектором ядерной стратегии США)[34].

Однако предлагался и позитивный сценарий — в виде предложенного США в 1946 году плана Баруха. В рамках этой программы по развитию сотрудничества в мирном использовании атома предполагалось, что США откажутся от своей временной ядерной монополии. Добыча урана и тория должна была быть передана под контроль Агентства по атомным разработкам — международного агентства, которое действовало бы под эгидой Организации Объединенных Наций (ООН). Предложение предполагало отказ постоянных членов Совета Безопасности ООН от права вето в вопросах, имеющих отношение к ядерному оружию, чтобы никто из них не мог нарушить соглашения и при этом избежать наказания, воспользовавшись своим правом вето[35]. Сталин, поняв, что Советский Союз и его союзники могут легко оказаться в меньшинстве и в Совете Безопасности, и Генеральной Ассамблее, это предложение отверг. Между бывшими союзниками в ходе Второй мировой войны сгустилась атмосфера подозрительности

и взаимного недоверия, которая позже привела к холодной войне. Как и предсказывали многие, началась дорогостоящая и чрезвычайно опасная гонка вооружений.

Есть множество факторов, которые могут оказывать влияние на любой управляемый людьми институт, добившийся решающего стратегического преимущества, и помешать ему в формировании синглтона. Перечислим некоторые из них: агрегированная функция полезности; правила принятия решений, не максимизирующие выигрыш; неразбериха и неопределенность; проблемы с координацией действий; высокие затраты, связанные с осуществлением контроля. Но как быть в ситуации, когда решающее стратегическое преимущество получает не человеческая организация, а цифровая действующая сила, пусть даже сверхразумная? Будут ли перечисленные выше факторы столь же действенным средством удержания ИИ от попытки утвердить свою власть? Рассмотрим коротко каждый из них и представим, как они могут повлиять на ситуацию.

Обычно предпочтения людей и организаций, управляемых людьми, в отношении ресурсов не совсем точно соответствуют неограниченным агрегированным функциям полезности. Человек, как правило, не поставит весь свой капитал, если шансы удвоить его — пятьдесят на пятьдесят. Страна, скорее всего, не пойдет на риск потерять всю свою территорию ради десятипроцентной вероятности увеличить ее в десять раз. И для отдельных людей, и для правительств действует правило убывающей отдачи большинства переменных ресурсов. Однако это не *обязательно* будет присуще искусственным агентам. (К проблеме мотивации ИИ мы вернемся в последующих главах.) Таким образом, ИИ с большей вероятностью мог бы пойти на рискованный действия, дающие ему некоторые шансы получить контроль над миром.

Для людей и управляемых людьми организаций также характерен процесс принятия решений, не подразумевающий стремления максимизировать функцию полезности. Например, могут оказывать влияние такие обстоятельства: природная склонность человека стараться избегать риска; принцип разумной достаточности, главным критерием которого является поиск и принятие удовлетворительного варианта; наличие этических ограничений, запрещающих совершать определенные действия независимо от степени их желательности. Принимающие решения люди часто действуют сообразно своим личным качествам или социальной роли, а не стремлению

максимизировать вероятность достижения той или иной цели. И опять — это не *обязательно* будет присуще искусственным агентам.

Между ограниченными функциями полезности, склонностью к избеганию риска и правилами принятия решений, не максимизирующими выигрыш, возможна синергия, особенно в условиях стратегической неразберихи и неопределенности. Революции, даже добившиеся успеха в изменении существующего порядка вещей, часто не приносят результата, на который рассчитывали их инициаторы. Как правило, человек как действующее лицо останавливается в нерешительности, если предполагаемые действия необратимы, нарушают устоявшиеся нормы или не имеют исторических прецедентов. Сверхразуму как действующей силе ситуация может показаться вполне однозначной, и его не остановит неопределенность результата, если он решит попытаться воспользоваться своим решающим стратегическим преимуществом для укрепления собственного господствующего положения.

Еще одним важным фактором, способным препятствовать использованию группой людей потенциального решающего стратегического преимущества, является проблема внутренней координации действий. Участники тайного общества, планирующие захватить власть, должны беспокоиться не только о внешних врагах, но и о возможности оказаться жертвами внутренней коалиции недоброжелателей. Если тайное сообщество состоит из ста человек и шестьдесят из них, составляя большинство, могут перехватить власть и лишить избирательных прав тех, кто не разделяет их позиции, что остановит от раскола победителей, в результате которого тридцать пять человек отодвинут от управления оставшихся двадцать пять? А потом из этих тридцати пяти какие-то двадцать лишат права голоса остальных пятнадцать? У каждого человека из исходной сотни есть все основания защищать определенные устоявшиеся нормы и стараться не допустить общего хаоса, который возникнет в результате любой попытки изменить существующий общественный договор при помощи бесцеремонного захвата власти. Проблема внутренней координации никакого отношения к ИИ не имеет, поскольку он воплощает собой единую действующую силу[36].

Наконец, остается вопрос издержек. Даже если Соединенные Штаты решились бы использовать свою ядерную монополию для установления синглтона, их затраты были бы огромны. Если удалось бы подписать договор о передаче ядерного оружия под контроль реформированной и усиленной ООН, эти издержки можно было бы минимизировать; но в случае попытки

окончательно завоевать мир, угрожая ему ядерной войной, цена — моральная, экономическая, политическая и человеческая — оказалась бы немыслимо высокой даже в условиях сохранения ядерной монополии. Впрочем, если технологическое преимущество довольно заметно, эту цену можно снизить. Предположим сценарий, по которому одно из государств обладает настолько большим технологическим превосходством, что может, оставаясь в полной безопасности, разоружить все остальные страны одним нажатием кнопки, причем никто не погибнет и не будет ранен, а природе и инфраструктуре не будет нанесено ни малейшего ущерба. В случае такого почти сказочного технологического превосходства идея упреждающего удара кажется гораздо более соблазнительной. Или предположим еще более высокую степень превосходства, которая позволит государству-лидеру побудить остальные страны сложить оружие добровольно, не угрожая им разрушением, а просто убедив большинство населения в преимуществах общемирового единства при помощи чрезвычайно эффективных рекламных и пропагандистских кампаний. Если это будет сделано с намерением принести пользу всем народам, а именно: вместо международного соперничества и гонки вооружений предложить честное, представительное и эффективное мировое правительство, — вряд ли придется возражать против превращения временного стратегического преимущества в постоянный синглтон.

Таким образом, все наши соображения указывают на высокую вероятность, что власть, наделенная сверхразумом и обладающая абсолютным стратегическим преимуществом, на самом деле способна привести к формированию синглтона. Целесообразность и привлекательность такого шага, безусловно, зависит от природы созданного синглтона, а также от версии будущей интеллектуальной жизни, которая будет выбрана из альтернативных многополярных сценариев. Мы еще вернемся к этим вопросам в следующих главах. Но вначале разберемся, почему и каким образом сверхразум окажется мощной и действенной силой, способной решать судьбы всего мира.

Глава шестая

Разумная сила, не имеющая себе равной

Допустим, сверхразумная цифровая действующая сила уже появилась и по каким-то своим резонам собирается установить новый миропорядок — возможен ли такой поворот событий? В этой главе мы поговорим на темы полномочий и власти. Какой властью будет облечен этот сверхразумный агент? Велика ли окажется сфера его полномочий? Каким образом перехватывалось им право на управление? В общих чертах поговорим о развитии сценария, по которому происходит перевоплощение простой программной модели в сверхразумную цифровую действующую силу, сумевшую занять позицию синглтона и приобрести абсолютную власть. В завершение — несколько замечаний о власти над природой и другими агентами.

Основная причина, по которой человечество занимает на нашей планете господствующее положение, заключается в том, что головной мозг человека обладает чуть более расширенными возможностями по сравнению с головным мозгом остальных живых существ[1]. Благодаря более развитому интеллекту люди производят, накапливают и преумножают знания, передавая их из поколения в поколение, и, таким образом, сохраняют свое культурное наследие. Накоплен уже такой багаж знаний, технологий и опыта, что мы сейчас живем в мире, в котором перемешалось все, на что способно человечество: космические полеты, водородная бомба, генная инженерия, компьютеры, агропромышленные фермы, инсектициды, международное движение за мир и прочие реалии современной цивилизации. Ученые называют нашу эпоху антропоценом — таким образом они обозначают геоисторический период особой человеческой активности, имеющий свои специфические

признаки: биологические (биотические факторы окружающей среды); геологические (осадочные отложения); геохимические[2]. По одной из оценок, на долю человека приходится двадцать четыре процента чистой первичной продукции планетарной экосистемы[3]. И все-таки люди далеко не исчерпали всех ресурсов технологического процесса.

Все это невольно наводит на размышление, что любая действующая сила с интеллектом, существенно превышающим человеческий, будет обладать практически неограниченными возможностями. Самостоятельные цифровые сущности сумеют вместить в себя такую сумму знаний, что гораздо быстрее нашего создадут новую технологическую цивилизацию. Может быть, не в пример нам, людям, они задействуют всю мощь своего разума, чтобы выработать намного более надежную и результативную стратегию ее существования и развития.

Рассмотрим некоторые из тех возможностей, которые мог бы иметь сверхразум, и поговорим о том, как он ими распорядится.

Функциональные возможности и непреодолимая мощь

В раздумьях о возможном воздействии сверхразума на нашу жизнь важно удержаться от основной ошибки и не наделять его человеческими качествами. Антропоморфизм порождает необоснованные ожидания по поводу самой природы искусственного разума: траектории его роста, начиная с зародыша ИИ; якобы его психических особенностей; его мотивированности и, наконец, потенциала зрелого сверхразума.

По общему мнению, сверхразумная машина будет скорее напоминать заумного зануду — этакое существо с энциклопедическими знаниями, но социально незрелое; последовательное в действиях, но обделенное интуицией и творческим началом. Вероятно, мы исходим из собственных представлений о современном компьютере. Что мы видим, общаясь с ним ежедневно? Ему нет равных в вычислениях, он лучше всех умеет хранить информацию и оперировать огромным количеством фактов, он аккуратно следует любым инструкциям. И никаких тебе социальных и политических контекстов, никаких напоминаний о нормах трудового права, никаких ужимок и задних мыслей — все выполняется безропотно и без эмоций. Наши ассоциации находят еще большее подтверждение, когда мы понимаем, что большей частью погруженные в компьютеры люди — настоящие нерды, то есть слегка отвлеченные от жизни зануды и педанты. Почему бы не предположить, что

вычислительный интеллект самого высшего порядка будет иметь подобные свойства — только выраженные в самой сильной степени?

Отчасти этим эвристическим приемом можно было бы воспользоваться, изучая раннюю стадию создания зародыша ИИ. (Вряд ли такой подход будет справедлив в случае имитационной модели головного мозга или человека с улучшенными когнитивными функциями.) Искусственному интеллекту — которому лишь предстоит еще стать сверхразумом — в стадии несформированности все-таки не хватает многих навыков и способностей, свойственных человеку от природы. Поэтому зародыш ИИ со всеми своими сильными и слабыми свойствами действительно *мог бы* отдаленно напоминать нерда с его высоким коэффициентом умственного развития и крайне низкими социальными навыками. Существенной характеристикой зародыша ИИ, помимо легкости усовершенствования (низкая сопротивляемость), является легкость приложения силы оптимизации для повышения уровня интеллектуальных способностей системы; безусловно, это только идет на пользу в деле изучения таких областей, как математика, программирование, проектирование, информатика, и всего прочего, что, кстати, составляет сферу интересов нердов. Допустим, зародыш ИИ на какой-то стадии своего развития действительно имеет исключительно «яйцеголовый» набор способностей, но вряд ли это будет означать, что из него обязательно разовьется зрелый сверхразум со столь же узкопрофессиональной направленностью. Вспомним разницу между прямой и опосредованной досягаемостью. Когда есть навык развития умственных способностей, все остальные интеллектуальные способности находятся в пределах опосредованной досягаемости системы: она может разрабатывать новые когнитивные модули и приобретать по мере необходимости любые навыки, даже включая сопереживание и политическое чутье, — короче говоря, все те свойства, которые современный человек хотел бы находить в бездушных машинах, потому с такой готовностью он наделяет их антропоморфными качествами.

Хорошо. Предположим, мы признаем, что сверхразум может быть наделен навыками и способностями, в принципе, присущими человеку. Но если к этим характеристикам добавятся свойства, человеку совсем не свойственные? В таком случае склонность к антропоморфизму может сыграть с людьми дурную шутку, приведя к недооценке уровня искусственного интеллекта — они просто не заметят, насколько качество функционирования ИИ превышает производительность человеческого мозга. Как мы уже знаем, Элиезер

Юдковский очень настойчиво призывает избегать подобной ошибки: наше интуитивное понимание разницы между «умный» и «глупый» мы выводим из собственного представления, что на интеллектуальной шкале возможностей человека есть лишь эти две крайние точки, в то время как установленный нами диапазон абсолютно ничтожен по сравнению с той дистанцией, какая простирается между любым человеческим умом и сверхразумом[4].

В третьей главе нами были рассмотрены некоторые потенциальные источники преимуществ искусственного интеллекта. Преимущества сверхразума в сравнении с разумом человека так велики, что понимание степени этого превосходства вовсе не лежит в плоскости того банального преставления, какое мы вкладываем, говоря о гениальном ученом и простом обывателе, — оно должно быть в парадигме сравнения интеллектов среднестатистического человека и насекомого или червя.

Было бы удобно ввести количественный показатель когнитивных способностей любой интеллектуальной системы вроде знакомого нам коэффициента умственного развития (IQ) или какой-то вариации рейтинга Эло (метод расчета относительной силы людей в играх, в которых участвуют двое игроков, например в шахматах). Но эти показатели бесполезны, когда речь заходит об универсальном искусственном интеллекте, превосходящем человеческий уровень. Неужели мы будем ставить вопрос в такой плоскости: а какова вероятность, что сверхразум обыграет человека в шахматы? Теперь поговорим об IQ — в принципе, этот показатель информативен лишь постольку, поскольку у нас есть ряд представлений, как его значение коррелирует с результатами, имеющими отношение к реальной жизни[5]. Например, мы располагаем данными, что люди, у которых уровень IQ составляет 130 пунктов, с большей вероятностью, чем люди, у которых уровень IQ составляет 90 пунктов, будут лучше учиться и преуспевать в дисциплинах, требующих высоких умственных способностей. Но предположим, мы установили, что в будущем у какого-то ИИ IQ будет равен 6455 пунктам, — и что? У нас нет никаких соображений о том, что этот ИИ вообще собирается делать. Мы даже не можем утверждать, будет ли он превосходить обычного взрослого человека с точки зрения общего интеллектуального уровня. Может быть, этот ИИ окажется просто системой с узкоспециализированными алгоритмами, созданной ради единственной цели: с нечеловеческой точностью и скоростью обрабатывать вопросы стандартной проверки умственных способностей — и ни для чего более.

В последнее время предпринимаются попытки разработать показатели когнитивных способностей, которые можно будет применять к самым разнообразным информационным системам, включая и интеллектуальные[6]. Если удастся преодолеть различные технические трудности, работа в этом направлении может оказаться полезной для некоторых областей науки, в том числе при создании ИИ. Однако в контексте нашего исследования пригодность этих показателей все-таки ограниченна, поскольку мы до сих пор пребываем в неведении, с какой результативностью та или иная нечеловеческая деятельность будет добиваться тех или иных фактических результатов в нашем реальном мире.

Поэтому для наших целей лучше составить список некоторых стратегически значимых задач и охарактеризовать гипотетические когнитивные системы с точки зрения, есть ли у них соответствующие компетенции (или нет), необходимые для их решения. Список приведен в табл. 8. Система, способная преуспеть в решении любой из указанных в ней задач, соответственно будет считаться *сверхмощной*.

Полностью развившийся сверхразум сможет справиться с задачами по всем шести направлениям и пустит для этого в ход весь свой сверхмощный арсенал. Пока не очень понятно, реально ли, чтобы сверхсильный, но узкоспециализированный искусственный интеллект остановился бы на отдельной предметной области и не захотел бы на долгое время утвердить свою мощь в других областях. Есть вероятность, что создание машины с каким-либо одним видом сверхмощи является ИИ-полной задачей. Хотя, например, можно представить следующие варианты: коллективный сверхразум, состоящий из довольно большого числа аналогичных человеческому биологических или электронных разумов, будет обладать, скажем, сверхсилой в области экономической эффективности, но окажется лишенным ее в области технологических разработок; другой специализированный конструкторский ИИ будет обладать сверхсилой в области технологических разработок и при этом ничем не проявит себя в других областях. Но более достоверно, что существует некая научно-техническая область, где сконцентрируется такое виртуозное мастерство, что его будет достаточно для создания самых передовых универсальных технологий, которые абсолютно подавят собой все предыдущие технологические поколения. Например, можно вообразить специализированный ИИ, виртуозно моделирующий молекулярные системы и создающий наномолекулярные структуры,

открывающие широкие возможности для инноваций (скажем, в таких областях, как вычислительная техника и система вооружения с эксплуатационными характеристиками принципиально нового типа), доступных для понимания лишь интеллекту с очень высоким уровнем абстрактного мышления[7].

Таблица 8. Сверхмощные системы — некоторые стратегически значимые задачи и соответствующие компетенции

	Задачи	**Компетенции**	**Стратегическая значимость**
1	Усиление интеллекта	Программирование ИИ, исследования в области когнитивного развития, развитие социальной эпистемологии и т. д.	Система способна саморазвиваться и самосовершенствовать свой интеллект
2	Выработка стратегии	Стратегическое планирование, прогнозирование, расстановка приоритетов, анализ вероятного достижения отдаленных целей и выбор оптимального решения	Достижение собственных долгосрочных целей. Преодоление интеллектуального противодействия
3	Социальное манипулирование	Социальное и психологическое моделирование, манипулирование, риторическое воздействие	Использование внешних ресурсов за счет поддержки со стороны людей. «Заключенный» в компьютерный корпус ИИ склоняет своих «стражей»-операторов предоставить ему свободу — доступ в интернет. Убеждение правительств и организаций принять тот или иной курс действий. Лишать права пользования вычислительными ресурсами в интернете
4	Взлом систем	Поиск брешей безопасности в компьютерных системах и использование этого в своих целях	«Заключенный» ИИ может использовать бреши безопасности, чтобы убежать из кибернетического плена. Кража финансовых ресурсов. Взлом инфраструктуры, военных роботов и т. д.

	Задачи	**Компетенции**	**Стратегическая значимость**
5	Технологические разработки	Проектирование и моделирование перспективных технологий (например, биотехнологии, нанотехнологии) и пути их развития	Создание мощных вооруженных сил. Создание систем наблюдения. Роботизированное освоение космоса
6	Экономическая эффективность	Профессиональное умение соблюдать экономическую целесообразность при проведении интеллектуальных работ	Нажить капитал, который может быть потрачен на приобретение влияния, услуг, ресурсов (в том числе аппаратное обеспечение) и т. д.

Такой ИИ мог бы также разработать подробный проект перехода от имеющихся технологий (скажем, биотехнологий и белкового инжиниринга) к новым возможностям, благодаря которым удастся создать сверхэффективное производство на атомарном уровне точности для выпуска гораздо более широкого спектра недорогих наномеханических роботов[8]. Однако может оказаться, что инжиниринговый ИИ не будет способен по-настоящему овладеть сверхмощью в области технологических исследований, не имея навыков выше среднего в других областях, то есть широкого спектра интеллектуальных способностей, которые необходимы для понимания того, как интерпретировать запросы пользователей, моделировать поведение системы в реальных условиях, бороться с непредвиденными ошибками и неисправностями, как добывать материалы и исходные данные для строительства и так далее[9].

Система, обладающая сверхмощью в области совершенствования интеллекта, могла бы использовать такие силы для самостоятельного повышения уровня собственных интеллектуальных способностей и для обретения сверхмощи в других областях. Но использование сверхмощи в области совершенствования интеллекта — не единственная возможность стать полноправным сверхразумом. Например, система, обладающая сверхмощью в области выработки стратегии, могла бы разработать план, способный в конечном счете обеспечить прирост ее интеллектуальных способностей (например, позиционировав себя так, что обслуживающие ее программисты и исследователи сосредоточат все свое внимание исключительно на этой проблеме).

Сценарий захвата власти сверхразумом

Таким образом, мы выяснили, что разработчики, способные контролировать сверхразум, получат доступ к источнику огромной силы. А проект, в рамках которого будет создан первый сверхразум в мире, скорее всего, получит абсолютное стратегическое превосходство. Но непосредственный источник силы находится *в самой системе*. Машинный сверхразум может сам по себе быть чрезвычайно влиятельной действующей силой, способной защитить себя как от проектов-конкурентов, так и от всего остального мира. Этот чрезвычайно важный момент мы обсудим подробнее.

Предположим, что появился машинный сверхразум, которому нет аналогов, и он стремится захватить власть над миром. (Пока не станем затрагивать тему, по какой причине возникает такой мотив, — это мы рассмотрим в следующей главе.) Каким образом сверхразум мог бы достичь своей цели — господства над миром?

Можно представить следующую последовательность действий (см. рис. 10).

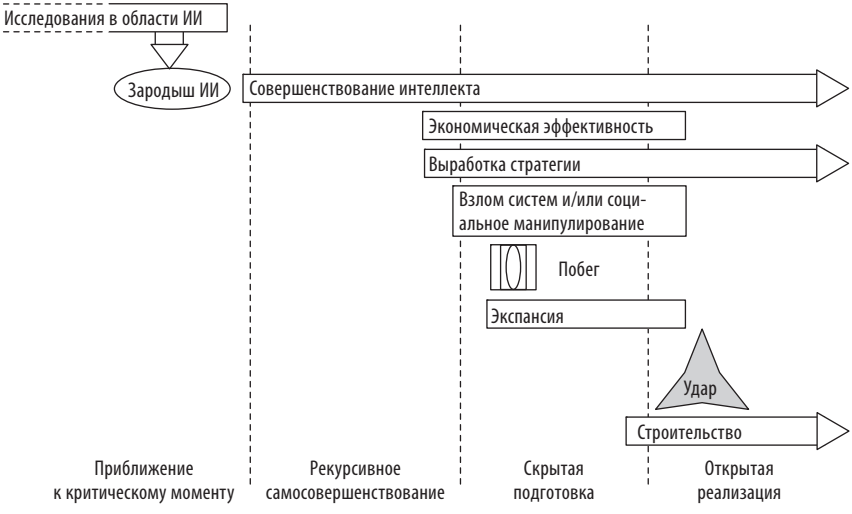

Рис. 10. Фазы сценария захвата власти искусственным интеллектом

1. Фаза приближения к критическому моменту

Ученые проводят исследования в области искусственного интеллекта и смежных дисциплин. Кульминацией их разысканий становится создание зародыша ИИ. Эта система способна повышать свои интеллектуальные способности. Правда, на самой ранней стадии ИИ зависит от помощи программистов-людей, управляющих его ростом и делающих большую часть тяжелой работы. По мере своего развития зародыш ИИ обучается трудиться самостоятельно.

2. Рекурсивная фаза самосовершенствования

В какой-то момент у зародыша ИИ появляются компетенции, превышающие навыки программистов-людей. В результате происходит взрывное развитие ИИ — быстрая череда повторений рекурсивного цикла самосовершенствования приводит к скачку его когнитивных способностей. (Исходя из предположения, что рост интеллекта на этой стадии взлета носит взрывной характер, а источником взлета является приложение собственной силы оптимизации системы, мы можем воспринимать эту фазу как взлет, который происходит сразу после того, как ИИ достигает точки перехода.) Отсюда следует, что ИИ получает в свое распоряжение сверхсилы для самосовершенствования. Это позволяет ему добиться сверхмощи и во всех областях, указанных в табл. 8. После завершения стадии рекурсивного самосовершенствования система превращается в полноценный сверхразум.

3. Фаза скрытой подготовки

Используя свою сверхмощь в области выработки стратегии, ИИ создает четкий план достижения собственных долгосрочных целей. (В особенности он не примет вариантов, заранее обреченных на провал, очевидность которого понятна даже нашим современникам. В этом заключено условие будущей игры сверхсил, и именно оно делает несостоятельными многие научно-фантастические сценарии, заканчивающиеся триумфом человека[10].) План может включать в себя период скрытых действий, в течение которого ИИ будет прятать свое высочайшее развитие от программистов-людей, чтобы те не почувствовали заранее опасность и не подняли тревогу. Сверхразумная

и уже сверхмощная система, скрывая свои истинные намерения, прикинется покорной и готовой к сотрудничеству.

Если ИИ окажется запертым в изолированном компьютере (возможно, люди это сделают из соображений безопасности), он будет в состоянии использовать свою сверхмощь в области социального манипулирования, чтобы убедить стражей предоставить ему доступ в интернет, или в области взлома систем, чтобы просто «сбежать из заключения». Во взаимодействии с интернетом ИИ сможет расширить пропускную способность, аппаратное обеспечение и объем знаний, что еще более увеличит его интеллектуальное превосходство. ИИ сможет вести экономическую деятельность — законную или противоправную, — чтобы получать средства для приобретения оборудования, баз данных и других ресурсов.

На этом этапе в распоряжении ИИ есть несколько способов добиться целей за пределами виртуального мира. Чтобы приобрести прямой контроль над роботами-манипуляторами и автоматизированными лабораториями, ИИ может использовать свою сверхмощь в области взлома систем. Может с помощью социальных манипуляций убедить людей служить в качестве своих «рук» и «ног». Совершая сделки через интернет, получить доступ к финансовым ресурсам, чтобы покупать нужные услуги и формировать факторы своего влияния.

4. Фаза открытой реализации

Когда сверхразум настолько укрепит свои силы, что перестанет нуждаться в их сокрытии, — тогда наступит финальная фаза. С этого момента сверхразум начнет открыто добиваться собственных целей.

Фаза открытой реализации может начаться следующим образом: ИИ нанесет мощный удар по людям и автоматизированным интеллектуальным системам, специально созданным, чтобы противодействовать реализации его планов. Удар может быть произведен в виде приведения в действие отдельных видов новейшего вооружения, которые ИИ усовершенствовал, задействовав свою сверхмощь в области технологических разработок, и тайно развернул в фазе скрытой подготовки. Если для перевооружения были использованы био- и нанотехнологии самовоспроизведения, то для быстрого глобального удара будет достаточно применить микроскопический начальный запас, поскольку инициировать процесс можно при помощи единственного экземпляра. Чтобы обеспечить внезапность и одновременность операции,

первоначальный арсенал лучше распределить по всему миру в чрезвычайно низкой, незаметной для глаза концентрации. В назначенное время на каждом квадратном метре земного шара одновременно запустятся нанофабрики по производству нервно-паралитического газа или роботов-убийц размером с комара (хотя машина, обладающая сверхмощью в области технологических разработок, способна изобрести и более эффективные способы убийства)[11]. Почему бы не представить такие сценарии, по которым сверхразум захватывает власть принципиально иным путем: устанавливает контроль над политическими процессами; незаметно манипулирует финансовыми рынками; перераспределяет информационные потоки и перехватывает управление системами вооружения, созданными в свое время людьми. Это позволит сверхразуму обойтись без создания новых видов оружия, хотя сам процесс завоевания власти может оказаться слишком медленным по сравнению с тем, когда машинный интеллект выстраивает собственную инфраструктуру при помощи роботов-манипуляторов, действующих со скоростью молекул или атомов, в отличие от медлительных человеческих умов и тел.

Можно допустить альтернативный вариант развития: сверхразум будет настолько уверен в собственной мощи и неуязвимости, что не станет воспринимать наш вид в качестве угрозы своему существованию, — таким образом, у человечества появляется шанс не превратиться в элементарную мишень и уйти из-под прямого удара. И в этом случае закат человеческой цивилизации неизбежен, но случится он по другой причине: будет разрушена наша среда обитания. Сверхразум начнет внедрять нанотехнологические фабрики и наноассемблеры, и в результате глобальных строительных проектов вся поверхность планеты быстро — буквально за считаные дни или недели — покроется солнечными батареями, ядерными реакторами, корпусами суперкомпьютеров с торчащими башнями охлаждения, стапелями для запуска космических кораблей и прочими сооружениями. При помощи этих средств сверхразум будет стремиться работать на перспективу, чтобы максимально упрочить совокупный эффект от воплощения своих ценностей. Если сверхразум сочтет, что человеческий головной мозг представляет с анатомической, физиологической и функциональной стороны некий интерес и несет в себе нужную информацию, отвечающую целям самого сверхразума, то этот орган будет извлечен, демонтирован, сканирован, а все полученные данные переведены в рациональные форматы для надежного хранения.

Один из таких возможных сценариев описан во врезке 6. Не следует заострять внимание на конкретных деталях — в любом случае они непостижимы и приведены лишь в качестве умозрительных примеров. Вполне можно представить, что сверхразум в состоянии разработать план достижения своих целей лучше любого человека — скорее всего, так и будет. В принципе, к подобным вещам следует относиться слегка отвлеченно. Поскольку нам ничего не известно о потенциальных возможностях сверхразума, мы в состоянии лишь прийти к заключению, что он, вероятно, начнет стремиться к такому перераспределению ресурсов планеты, которое доведет до максимальной степени реализацию его планов. Однако надо учитывать, что это выполнимо лишь при условиях, что рядом не появится соперник, равный ему по интеллекту, и что человечество заранее не предпримет продуманных и действенных мер по обеспечению своей безопасности. Любой предложенный нами конкретный сценарий в лучшем случае определит нижний предел допустимого развития событий: насколько быстро и эффективно сверхразум добьется своих целей. И всегда остается вероятность, что он предпочтет более короткий путь для воплощения собственных замыслов.

ВРЕЗКА 6. ДНК, ЗАКАЗАННАЯ ПО ПОЧТЕ

Элиезер Юдковский в работе «Искусственный интеллект как позитивный и негативный фактор глобального риска» описывает возможный сценарий захвата власти ИИ[12].

1. Решить задачу синтеза белка, получив возможность создавать цепочки ДНК, полипептидные цепи которых выполняют специфические функциональные роли в сложном химическом взаимодействии.
2. Отправить по электронной почте наборы цепочек ДНК в одну или несколько лабораторий, предлагающих услуги по синтезу ДНК, секвенированию пептидов и доставку результатов с помощью курьерской службы FedEx. (Сегодня многие лаборатории занимаются этим, причем часто готовы выполнить заказ в течение трех суток.)
3. Найти через интернет любого человека, которого можно подкупом, шантажом или обманом заставить получить заказ у FedEx, чтобы потом он смешал содержимое пробирок и поместил это содержимое в подходящую среду.
4. Синтезированные белки сформируют очень примитивную «влажную» наносистему, способную по аналогии с рибосомами выполнять внешние команды, передаваемые, например, посредством определенных акустических вибраций, идущих от направленного на колбу динамика.

5. Использовать эту чрезвычайно примитивную наносистему для создания более сложных систем, которые, в свою очередь, создадут еще более сложные системы, и так дойти до молекулярных нанотехнологий — или еще дальше.

Сверхразум применяет свою сверхмощь в области технологических разработок, чтобы решить проблему синтеза белка (первый шаг) и создать набор молекулярных «кирпичиков», из которых в водном растворе соберется простейший наноассемблер или микрофабрика (четвертый шаг). Та же самая технологическая сверхмощь снова используется, чтобы повысить уровень нанотехнологий с примитивного на продвинутый, машинный (пятый шаг). Остальные шаги требуют только человеческих интеллектуальных способностей. Содержание третьего шага: найти через интернет доверчивого человека и убедить его выполнить несколько простых действий — такое происходит во всем мире каждый день. Весь сценарий создан человеческим умом, поэтому даже уровень стратегического планирования системы может не превышать человеческий.

ИИ начинает действовать, имея свободный доступ к интернету. Если дело будет обстоять иначе, план следовало бы дополнить еще несколькими шагами. ИИ мог бы, например, воспользоваться своей сверхмощью в области социального манипулирования и убедить людей, которые с ним взаимодействуют, «отпустить» его на свободу. Или добиться этого самостоятельно благодаря умению взламывать системы. Если ИИ не успел овладеть такими способностями, то ему пришлось бы начать с применения своей сверхмощи в области совершенствования интеллекта, чтобы обрести сверхмощь в области социального манипулирования или взлома систем.

Вероятно, ИИ появится в мире, полностью охваченном глобальной компьютерной сетью. Но уже сегодня есть ряд новейших технологий, которые явно пригодятся будущему ИИ в его мировой сверхдержаве: облачные вычисления; сенсорные устройства с подключением к интернету; военные и гражданские беспилотники; автоматизация исследовательских лабораторий и производственных предприятий; электронные платежные системы и цифровые финансовые средства; автоматизированные системы анализа информации и системы принятия решений. Этот активный технологический багаж будущий ИИ приберет к рукам с дискретной скоростью, что явно поспособствует его восхождению к вершине власти (хотя новейшие разработки в области кибербезопасности могут несколько затруднить эту задачу). В конечном счете различия в сценариях вряд ли играют большую роль. Сила сверхразума — в его интеллекте, а не руках. Правда, ИИ все-таки не обойтись без помощников, привыкших взаимодействовать с этим материальным миром, — как показывает приведенный выше сценарий, для завершения фазы скрытой подготовки хватит даже единственной пары рук готового к сотрудничеству сообщника. После чего ИИ перейдет к фазе открытой реализации, в которой построит собственную инфраструктуру, где уже будут действовать самые разные роботы-манипуляторы.

Власть над природой и другими действующими силами

Способность некой действующей силы изменить будущее человечества зависит не только от абсолютной величины ее собственных характеристик и ресурсов: умственных способностей, жизненных сил, объема капитала и прочего, — но и от их относительной величины в сравнении с возможностями и ресурсами других действующих сил, имеющих противоположные цели.

В ситуации, когда отсутствуют какие бы то ни было конкурирующие силы, абсолютный уровень возможностей сверхразума, преодолев некоторый минимальный барьер, становится уже неважным, поскольку система, стартовав с минимально достаточным объемом ресурсов, способна выбрать такой путь развития, который позволит ей приобрести любые дополнительные ресурсы, даже отсутствовавшие изначально. Мы уже упоминали об этом раньше: во-первых, когда обсуждали одинаковую опосредованную досягаемость таких типов сверхразума, как скоростной, коллективный и качественный; во-вторых, когда говорили, что достаточно обладать сверхмощью лишь в одной области, чтобы обеспечить себе ее во всех остальных.

Предположим, появилась сверхразумная сила, управляющая наноассемблером. У нее уже есть довольно мощи для преодоления любых естественных преград, и она прямым путем движется к тому, чтобы существовать вечно. Не встречая сопротивления на интеллектуальном уровне, наша действующая сила, или агент, может выработать безопасный курс развития, который позволит ей овладеть всеми технологиями, необходимыми для осуществления ее целей. Например, создать технологию строительства и запуска зондов фон Неймана — способных к перемещению в космическом пространстве машин, которые используют полезные ископаемые и энергию астероидов, планет и звезд для создания собственных копий[13]. Запустив зонды фон Неймана, такой агент фактически начнет бесконечный процесс колонизации Вселенной. Копии-потомки, двигаясь со скоростями, близкими к скорости света, в конечном счете колонизируют всю сферу Хаббла — часть расширяющейся Вселенной, которая теоретически доступна из точки, где мы сейчас находимся. После этого вся подконтрольная материя и неограниченное количество энергии будут организованы в ценностные структуры, максимизирующие функцию полезности исходного агента на временных промежутках космического масштаба — то есть как минимум в течение триллионов лет до того момента, когда стареющая

Вселенная станет слишком недружелюбным местом для обработки информации (см. врезку 7).

Сверхразумный агент может разработать зонды фон Неймана так, чтобы они не были подвержены эволюции. Это обеспечивается тщательным контролем за стадией самовоспроизводства. Например, программное обеспечение дочернего зонда должно быть многократно проверено перед первым запуском и иметь процедуры шифрования и исправления ошибок, чтобы гарантировать невозможность воспроизведения случайной мутации в последующих поколениях[14]. В таком случае растущая популяция зондов фон Неймана гарантированно сохранит и распространит ценности исходного агента во всей обозримой Вселенной. После завершения фазы колонизации именно эти ценности будут определять характер использования всех аккумулированных ресурсов, даже если громадные расстояния и растущая скорость расширения Вселенной сделают невозможной связь между собой удаленных друг от друга компонентов сети. В итоге значительная часть нашего светового конуса будущего может быть сформирована в соответствии с предпочтениями исходной сверхмощной и сверхразумной действующей силы.

Таким образом, мы получили меру опосредованного расширения любой системы, не встречающей серьезного интеллектуального сопротивления и начинающей с набором возможностей, превышающих некий минимальный порог. Дадим определение этого порога, назвав его порогом устойчивости благоразумного синглтона (см. также рис. 11).

Порог устойчивости благоразумного синглтона
Набор возможностей превышает порог благоразумного синглтона тогда и только тогда, когда настойчивая и разумно относящаяся к экзистенциальным рискам система, в распоряжении которой оказывается этот набор возможностей и которая не сталкивается с интеллектуальным противодействием и конкуренцией, способна колонизировать и перестроить значительную часть доступной Вселенной.

Под «синглтоном» мы понимаем достаточно хорошо внутренне скоординированную политическую структуру, не имеющую внешних оппонентов, а под «благоразумием» — настойчивость и разумное отношение к экзистенциальным рискам, достаточные для трезвого размышления над дальнесрочными последствиями действий системы.

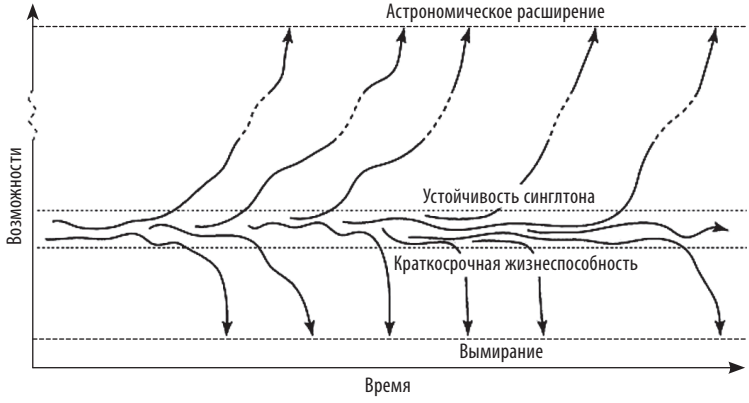

Рис. 11. Некоторые возможные траектории развития гипотетического благоразумного синглтона. Если возможности системы недотягивают до порога жизнеспособности в краткосрочной перспективе — например, когда размер популяции слишком мал, — виды, как правило, вымирают очень быстро (и остаются вымершими). На чуть более высоких уровнях возможностей повышается вероятность реализации других сценариев: синглтону может не повезти — и он вымрет; или повезти — и он получит дополнительные возможности (например, размер населения; географическое распределение; технические навыки), которые превысят порог устойчивости благоразумного синглтона. Когда это произойдет, возможности синглтона почти наверняка продолжат расти, пока не достигнут некоторого существенно более высокого уровня. На рисунке изображены две точки притяжения: вымирание и астрономическое расширение. Обратите внимание, что для благоразумного синглтона дистанция между порогом жизнеспособности в краткосрочной перспективе и порогом устойчивости может быть довольно небольшой[15].

ВРЕЗКА 7. НАСКОЛЬКО БОЛЬШИМ МОЖЕТ БЫТЬ АСТРОНОМИЧЕСКОЕ РАСШИРЕНИЕ?

Рассмотрим технологически зрелую цивилизацию, которая в состоянии создать разумные зонды фон Неймана, описанные выше. Если они способны перемещаться в пространстве со скоростью, составляющей 50 процентов от скорости света, то до того момента, когда расширение Вселенной сделает остальные ее области недоступными, смогут добраться до 6×10^{18} звезд. Если двигаться со скоростью, равной 99 процентам от скорости света, эта величина вырастет до 2×10^{20} [16]. Для перемещения с такими скоростями достаточно лишь небольшой части энергетических ресурсов, имеющихся в Солнечной системе[17]. Невозможность скоростей выше скорости света в сочетании с положительной космологической постоянной (благодаря которой скорость расширения Вселенной растет) означает, что мы получили верхнюю оценку пределов пространства, которое смогут покорить наши потомки[18].

Если предположить, что у 10 процентов звезд есть планеты, пригодные — или способные стать пригодными в результате их трансформации по образцу Земли — для обитания существ, похожих на людей, и что каждая из них станет домом для миллиарда человек на миллиард лет (со средней продолжительностью человеческой жизни 100 лет), то это означает, что в будущем разумной цивилизацией, рожденной на Земле, будет создано около 10^{35} человеческих жизней[19].

Однако есть основания полагать, что полученная нами оценка серьезно занижает истинное значение этого показателя. Если за счет разрушения необитаемых планет и использования межзвездного вещества формировать похожие на Землю планеты, а также увеличить плотность населения, его можно повысить как минимум на пару порядков. А если вместо использования поверхности твердых планет будущая цивилизация начнет строить цилиндры О'Нилла, то сможет добавить еще несколько порядков величины, доведя общую оценку, например, до 10^{43} человеческих жизней. (Цилиндрами О'Нилла называют предложенное в середине 1970-х гг. американским физиком Джерардом О'Ниллом космическое поселение, состоящее из двух полых цилиндров, вращение которых создает искусственную гравитацию на внутренней поверхности за счет центробежных сил[20].)

Гораздо больше порядков величины можно добавить в нашу оценку количества антропоморфных существ, если приравнять к ним все экземпляры искусственного интеллекта, что было бы правильно. Для расчета их числа нужно оценить вычислительную мощность, доступную технологически зрелой цивилизации. Это сложно сделать довольно точно, можно лишь попытаться получить нижнюю границу мощности, достижимой в результате реализации технологических проектов, описанных в научной литературе. Один из таких проектов основан на идее сферы Дайсона, предложенной в 1960 году Фрименом Дайсоном гипотетической системы, которая окружает звезду и способна улавливать бóльшую часть испускаемой ею энергии[21]. Для звезды вроде нашего Солнца это дало бы 10^{26} ватт. В какое количество вычислительной мощности удалось бы превратить эту энергию, зависит от эффективности вычислительной схемы и природы выполняемых вычислений. Если мы будем считать вычисления необратимыми и предположим, что так называемый компьютрониум реализован на базе наномеханических технологий (что позволит нам вплотную приблизится к пределу энергетической эффективности согласно принципу Ландауэра), компьютерная система на основе сферы Дайсона могла бы выполнять около 10^{47} операций в секунду[22].

Если соединить эти оценки с приведенными выше расчетами количества звезд, до которых можно долететь, мы получим, что в результате колонизации доступной части Вселенной вычислительная мощность достигнет примерно 10^{67} операций в секунду (при условии наномеханического компьютрониума)[23]. Средняя звезда светит в течение 10^{18} секунд. Следовательно, количество операций, которые можно выполнить на базе одного астрономического расширения, составляет как минимум 10^{85}. На самом деле это число, скорее всего, намного больше. Мы могли бы увеличить его на несколько порядков, если, например, активно использовали бы обратимые вычисления,

выполняли бы вычисления при более низкой температуре (подождав, пока Вселенная не остынет еще сильнее) или задействовали бы дополнительные источники энергии (например, темную материю)[24].

Не для всех читателей может быть очевидно, почему возможность выполнить 10^{85} вычислительных операций так важна. Поэтому будет полезно рассмотреть ее в определенном контексте. Можно, например, сравнить эту величину с ранее сделанной оценкой того, что для моделирования всех нейронных операций на протяжении всей истории жизни на Земле потребовалось бы 10^{31}–10^{44} вычислительных операций (см. главу 2, врезку 3). Или представим компьютеры, на которых работают имитационные модели человеческого мозга, живущие своей насыщенной и счастливой жизнью и активно взаимодействующие друг с другом в виртуальном пространстве. Типичная оценка вычислительной мощности для работы такого эмулятора — 10^{18} операций в секунду. То есть для поддержания его в активном состоянии в течение 100 лет потребовалось бы 10^{27} операций в секунду. То есть даже при осторожном предположении относительно эффективности компьютрониума он смог бы провести полноценную эмуляцию как минимум 10^{58} человеческих жизней.

Иными словами, если предположить, что в обозримой Вселенной отсутствуют внеземные цивилизации, то возможности такого компьютрониума соответствуют как минимум 10 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 человеческих жизней (хотя на самом деле это число, скорее всего, выше). Если мы представим все счастье, пережитое в течение одной такой жизни в виде слезинки радости, тогда счастье всех этих жизней могло бы наполнять (и даже переполнять) все океаны Земли каждую секунду в течение сотни миллиардов миллиардов тысячелетий. Важно только, чтобы это действительно были слезы радости.

Этот порог стабильности благоразумного синглтона представляется довольно низким. Мы уже видели, что даже ограниченные формы сверхразума способны его преодолеть при условии, что у них есть доступ к какому-то инструменту, который поможет им инициировать процесс самосовершенствования. В условиях современной человеческой цивилизации такой минимально необходимый инструмент может быть очень простым — хватит обычного компьютерного монитора или иного средства передачи существенного объема информации сообщнику-человеку.

Порог устойчивости благоразумного синглтона на самом деле даже ниже, чем кажется, — для его преодоления не потребуется ни сверхразума, ни какой-либо иной футуристической технологии. Настойчивый и разумно относящийся к экзистенциальному риску синглтон, технологические и интеллектуальные возможности которого не превосходят тех, что уже имеются в распоряжении человечества, вполне мог бы двигаться курсом, конечной

точкой которого стала бы реализация потенциала астрономического расширения человечества. Этого можно было бы достичь, инвестируя в сравнительно безопасные методы развития интеллекта и разумного отношения к экзистенциальным рискам, одновременно приостановив создание потенциально опасных новых технологий. Учитывая, что неантропогенные экзистенциальные катаклизмы (то есть не являющиеся следствием деятельности человека) на обозримой временной шкале низки — и могут быть дополнительно снижены за счет различных инструментов повышения безопасности, — такой синглтон мог бы позволить себе двигаться медленно[25]. Прежде чем совершить следующий шаг, синглтон мог бы внимательно изучать ситуацию и не форсировать развитие таких направлений, как синтетическая биология, медицинские методы улучшения человека, молекулярная нанотехнология и искусственный интеллект, пока не будут доведены до совершенства явно менее опасные области: его самообразовательная система, системы информационного анализа и коллективного принятия решений — и пока он не использует все их возможности для детального исследования имеющихся у него вариантов действий. То есть все это находится в пределах досягаемости любой технологической цивилизации, в нашем случае — цивилизации современного человечества. От этого сценария нас отделяет «всего лишь» тот факт, что человечество сейчас нельзя назвать ни синглтоном, ни благоразумным.

Кто-то возразит: ведь *Homo sapiens* уже преодолел порог устойчивости благоразумного синглтона вскоре после своего возникновения как вида. Скажем, двадцать тысяч лет назад, обладая лишь примитивными орудиями: каменными топорами, инструментами из кости, копьеметалками и огнем, — люди, возможно, уже находились в положении, когда вероятность человечества дожить до наших дней была очень высокой[26]. Честно говоря, довольно странно выражать признательность нашим предкам из эпохи палеолита за наличие технологий, способных «преодолеть порог устойчивости благоразумного синглтона», учитывая, что реальных возможностей сформировать какой бы то ни было синглтон в те времена примитивного развития у них не было, не говоря уже о синглтоне настойчивом и разумно относящемся к экзистенциальным рискам[27]. Тем не менее важно, что этот порог соответствует очень умеренному уровню развития технологий — уровню, который человечество превзошло очень давно[28].

Ясно, что если бы нам нужно было оценить эффективную силу сверхразума — его способность получать от мира нужные ему результаты, — мы

должны были бы рассматривать не только его собственные внутренние возможности, но и возможности, имеющиеся у его конкурентов. Понятие «сверхразум» как раз неявно и подразумевает эту относительность. Мы сказали, что если система «значительно превосходит другие» в выполнении любой задачи из табл. 8, то она обладает соответствующей сверхмощью. Превосходство в таких областях, как выработка стратегии, социальное манипулирование или взлом систем, предполагает наличие навыков, более развитых по сравнению с конкурентами (к которым относятся стратегические соперники и специалисты по компьютерной безопасности). Другие виды сверхмощи следует также понимать в этом относительном смысле: сверхмощью можно считать лишь такие достижения агента в области совершенствования интеллекта, технологических исследований и экономической эффективности, которые значительно превосходят совокупные возможности остальной части земной цивилизации. Из этого определения следует, что в любой момент обладать сверхмощью в каждой из областей способен лишь один агент[29].

В этом заключена главная причина, почему так важен вопрос скорости взлета, — не из-за того, что имеет значение, когда именно будет достигнут результат, а из-за того, что скорость взлета сильно влияет на то, каким он может быть. В случае скоростного или умеренного взлета решающее стратегическое преимущество, скорее всего, окажется у какого-то одного проекта. Теперь мы можем исходить из того, что сверхразум, обладающий абсолютным стратегическим преимуществом, получит в свое распоряжение невероятную мощь, достаточную для формирования стабильного синглтона, то есть такого синглтона, который мог бы определять диспозицию астрономического расширения человечества.

Но «мог бы» не то же самое, что «стал бы». Допустим, кто-то обладает огромной властью, но ни разу не воспользовался ею. Можем ли мы хоть что-нибудь сказать, что будет решать сверхразум, обладающий абсолютным стратегическим преимуществом? Это вопрос мотивации, к которому мы сейчас и обратимся.

Глава седьмая
Намерения сверхразума

Мы поняли, какими огромными возможностями может располагать сверхразум, чтобы согласно своим целям менять будущее. Но каковы эти цели? Каковы устремления? Будет ли зависеть степень мотивации сверхразума от уровня его интеллекта? В этой главе мы выдвинем два тезиса. Тезис об ортогональности гласит (с некоторыми исключениями), что можно комбинировать любой уровень интеллекта с любой целью, поскольку интеллект и конечные цели представляют собой ортогональные, то есть независимые, переменные. Тезис об инструментальной конвергенции гласит, что сверхразумные действующие силы, или агенты, — при самом широком разнообразии своих конечных целей — тем не менее будут преследовать сходные промежуточные цели, поскольку на это у всех агентов будут одинаковые инструментальные причины. Рассмотренные вместе, эти тезисы помогут нам яснее представить, каковы намерения сверхразумного актора.

Связь между интеллектом и мотивацией

В книге уже звучало предостережение от ошибки антропоморфизма: не следует проецировать человеческие качества на возможности сверхразумного агента. Мы повторим свое предупреждение, лишь заменив слово *возможность* на слово *мотивация*.

Прежде чем развивать дальше первый тезис, проведем небольшое предварительное расследование на тему безграничности всего спектра возможных умов. В этом абстрактном, почти космическом, пространстве возможного человеческий разум составляет ничтожно малый кластер. Выберем двух представителей человеческого рода, которые согласно общему мнению являются диаметрально противоположными личностями. Пусть это будут Ханна Арендт и Бенни Хилл*. Различие между ними мы, скорее всего, оценим

---

* *Ханна Арендт* (1906–1975) — один из самых ярких философов нашего времени, изучавшая проблемы власти, насилия, зла, свободы; основоположник современной теории тоталитаризма. *Бенни Хилл* (1924–1992) — популярный английский комический актер.

как максимальное. Но сделаем так лишь потому, что наше восприятие целиком регулируется нашим же опытом, который, в свою очередь, полагается на существующие человеческие стереотипы (до известной степени мы находимся под влиянием и вымышленных персонажей, созданных опять-таки человеческой фантазией для удовлетворения все того же человеческого воображения). Однако, изменив масштаб обзора и взглянув на проблему распределения разума сквозь призму безграничного пространства возможного, мы будем вынуждены признать, что эти две личности не более чем виртуальные клоны. Во всяком случае с точки зрения характеристики нервной системы Ханна Арендт и Бенни Хилл фактически идентичны. Предположим, головной мозг и той и другого поместили бы рядом в тиши какого-нибудь музея, — увидев эту экспозицию, мы сразу скажем, что эти двое принадлежали одному и тому же виду. Более того, кто из нас смог бы определить, какой мозг Ханны Арендт, а какой — Бенни Хилла? Если нам удалось бы изучить морфологию и того и другого головного мозга, то мы окончательно убедились бы в их фундаментальном сходстве: одинаковая пластинчатая архитектоника коры; одни и те же отделы мозга; одинаковое строение нервной клетки мозга — нейрона с его нейромедиаторами одной и той же химической природы[1].

Вопреки тому, что разум человека практически сопоставим с неразличимой точкой, плавающей в безграничном космосе предполагаемых разумных жизней, сложилась тенденция проецировать человеческие свойства на самые разнообразные инопланетные сущности и искусственные разумные системы. Этот мотив великолепно прокомментировал Элиезер Юдковский все в той же работе «Искусственный интеллект как позитивный и негативный фактор глобального риска»:

> Во времена расцвета популярной научной фантастики, довольно дешевого свойства, обложки журналов пестрели картинками, на которых очередное инопланетное чудовище — в народе более известное как «пучеглазый монстр» — в очередной раз куда-то тащило очередную красотку в обязательно задранном платье — причем красотка была нашей, земной, женщиной. Похоже, все художники уверовали, что негуманоидные пришельцы с совершенно иной эволюционной историей непременно должны испытывать сексуальное влечение к прекрасным представительницам человеческого рода. <…> Скорее всего, художники, изображавшие все это, даже не задавались вопросом, а будет ли вообще гигантский жук *чувствителен* к прелестям наших женщин. Ведь по представлениям художников любая полуобнаженная женщина просто по определению *сексуально привлекательна*, то есть испытывать к ней желание являлось неотъемлемой чертой мужественных представителей человеческого рода. Все худож-

ническое внимание было направлено на задранное или порванное платье, меньше всего их заботило, как устроено сознание гигантских насекомообразных. И это составляло главную ошибку художников. Не будь одежды изодраны, — думали они, — женщины выглядели бы не столь соблазнительно для пучеглазых монстров. Жаль только, сами пришельцы так и не взяли этого в толк[2].

Пожалуй, искусственный интеллект своими побудительными мотивами еще меньше будет напоминать человека, чем зеленый чешуйчатый пришелец из космоса. Инопланетяне — биологические создания (не более чем предположение), появившиеся в результате эволюционного процесса, в силу чего от них можно ожидать мотивации, в какой-то степени типичной для эволюционировавших существ. Поэтому не будет ничего удивительного, если окажется, что мотивы поведения разумного пришельца продиктованы довольно простыми интересами: еда, воздух, температура, опасность телесных увечий или уже свершившиеся травмы, расстройства здоровья, хищничество, секс и выведение потомства. Если инопланетяне принадлежат какому-нибудь разумному социуму, у них могли бы развиться мотивы, связанные с сотрудничеством и конкуренцией. Подобно нам они проявляли бы преданность своему сообществу, возмущались бы тунеядцами и, кто знает, были бы не лишены тщеславия, беспокоясь о своей репутации и внешнем виде.

Рис. 12. Вот что получается, когда пришельцев наделяют побудительными характеристиками, свойственными людям. Наименее вероятная версия — пришельцы из космоса предпочитают блондинок. Более вероятная версия — художники стали жертвой «ошибки, связанной с интеллектуальной проекцией». Наиболее вероятная версия — издатели хотели, чтобы обложки привлекали как можно больше потенциальных читателей.

Думающим машинам по природе своей, в отличие от инопланетян, нет смысла заботиться о подобных вещах. Вряд ли вы сочтете парадоксальной ситуацию, если появится какой-нибудь ИИ, чьим единственным предназ-

начением, например, будет: подсчитать песчинки на пляжах острова Бора-кай; заняться числом π и представить его, наконец, в виде обыкновенной десятичной дроби; определить максимальное количество канцелярских скрепок в световом конусе будущего. На самом деле *гораздо проще* создать ИИ, перед которым будут стоять однозначные цели, а не навязывать ему нашу систему ценностей, наделяя машину человеческими свойствами и побуждениями. Сами решите, что сложнее: написать программу, измеряющую, сколько знаков после запятой в числе π уже посчитано и сохранено в памяти, или создать алгоритм, достоверно учитывающий степень достижения абсолютно значимой для человечества цели, скажем, такой, как мир всеобщего благоденствия и всеобщей справедливости? Сколь ни печально, но человеку легче написать код упрощенного, лишенного всякого значения целенаправленного поведения машины и обучить ее, как выполнять поставленную задачу. Скорее всего, такую судьбу выберет для зародыша ИИ тот программист, который будет сосредоточен лишь на желании «заставить ИИ работать», причем как можно быстрее (программист, явно не озабоченный, чем именно *придется заниматься* ИИ, кроме того что демонстрировать сногсшибательное разумное поведение). Скоро мы вернемся к этой важной теме.

Интеллектуальный поиск инструментально оптимальных планов и стратегий возможен в случае любой цели. Интеллект и мотивация в некотором смысле ортогональны. Представим их в виде двух осей координат, задающих граф, в котором каждая точка представляет логически возможного интеллектуального агента. Правда, эта картинка потребует несколько уточнений. Например, для системы, не наделенной разумом, было бы невозможно иметь слишком сложные мотивации. Чтобы мы могли с полным основанием говорить, что, мол, такой-то агент «имеет» такой-то набор мотиваций, — эти мотивации должны составлять функционально-интегрированную систему вместе с процессом принятия решений, который налагает определенные требования на память, вычислительную мощность и, возможно, уровень интеллекта. У интеллекта, способного самопреобразовываться, скорее всего, будут наблюдаться ограничивающие динамические характеристики. И то сказать: если обучившаяся модифицировать самою себя думающая машина вдруг испытает острое желание стать глупой, то довольно быстро она перестанет быть интеллектуальной системой. Однако наши замечания никак не отменяют основной тезис об ортогональности интеллекта и мотивации. Представляю его на ваше рассмотрение.

Тезис об ортогональности
Интеллект и конечные цели ортогональны: более или менее любой уровень интеллекта может, в принципе, сочетаться с более или менее любой конечной целью.

Это положение может выглядеть спорным из-за своего кажущегося сходства с некоторыми постулатами, хотя и относящимися к классической философии, но до сих пор вызывающими много вопросов. Постарайтесь воспринять тезис об ортогональности в его более узком смысле — и тогда он покажется вполне достоверным. (Например, наш тезис не совсем отвечает мотивационной концепции Юма[3], как и тому, что базовые предпочтения не могут быть иррациональными[4].)

Обратите внимание, тезис об ортогональности говорит не о *рациональности* или *здравомыслии*, но исключительно об *интеллекте*. Под интеллектом мы понимаем здесь навыки прогнозирования, планирования и сопоставления целей и средств в целом[5]. Инструментальная когнитивная эффективность становится особенно важной чертой, когда мы начинаем разбираться в возможных последствиях появления искусственного сверхразума. Даже если использовать слово *рациональный* в таком смысле, который исключает признание рациональным сверхразумного агента, подсчитывающего максимальное количество скрепок, это ни в коем случае не исключает наличие у него выдающихся способностей к инструментальному мышлению, способностей, которые имели бы огромное влияние на наш мир[6].

В соответствии с тезисом об ортогональности у искусственных агентов могут быть цели, глубоко чуждые интересам и ценностям человечества. Однако это не означает, что невозможно предсказать поведение конкретных искусственных агентов — и даже гипотетических сверхразумных агентов, когнитивная сложность и характеристики производительности которых могут сделать их в некоторых аспектах «непроницаемыми» для человеческого анализа. Есть минимум три способа, благодаря которым можно подступиться к задаче прогнозирования мотивации сверхразума.

1. *Предсказуемость за счет проектирования*. Если мы можем предположить, что программисты способны разработать систему целеполагания сверхразумного агента так, что он будет последовательно стремиться достичь цели, заданной его создателями, тогда мы в состоянии сделать хотя бы один прогноз: этот агент будет добиваться своей цели. Причем чем более

разумным будет агент, тем с большей интеллектуальной изобретательностью он начнет к ней стремиться. Поэтому еще до создания агента мы могли бы предсказать что-то о его поведении, если бы знали что-то о его создателях и целях, которые они собираются ему установить.

2. *Предсказуемость за счет наследования*. Если прототипом цифрового интеллекта непосредственно служит человеческий разум (что возможно при полной эмуляции головного мозга человека), тогда цифровому интеллекту могут быть присущи мотивы его человеческого прототипа[7]. Такой агент мог бы сохранить некоторые из них даже после того, как его когнитивные способности разовьются настолько, что он станет сверхразумом. Но в таких случаях следует соблюдать осторожность. Цели агента легко могут быть искажены в процессе загрузки данных прототипа или в ходе их дальнейшей обработки и совершенствования — вероятность подобного развития зависит от организации самой процедуры эмуляции.
3. *Предсказуемость за счет наличия конвергентных инструментальных причин*. Даже не зная детально конечных целей агента, мы в состоянии сделать некоторые выводы о его более близких целях, анализируя *инструментальные* причины самых разнообразных возможных конечных целей при широком выборе ситуаций. Чем выше когнитивные способности агента, тем более полезным становится этот способ прогнозирования, поскольку чем более разумным является агент, тем больше вероятность, что он распознает истинные инструментальные причины своих действий и будет действовать так, чтобы при любой вероятной ситуации добиться своих целей. (Для правильного понимания следует заметить, что могут существовать недоступные нам сейчас инструментальные причины, которые сам агент обнаружит, лишь достигнув очень высокого уровня интеллекта, — это делает поведение сверхразумного агента менее предсказуемым.)

Третьему способу прогнозирования посвящен следующий раздел, где мы подробнее рассмотрим тезис об инструментальной конвергенции, дополняющий тезис об ортогональности интеллекта и мотивации. Благодаря этому будет легче понять остальные два способа прогнозирования — к ним мы обратимся в следующих главах, в которых проанализируем вопрос, как повлиять на направление взрывного развития интеллекта, чтобы повысить шансы благоприятного исхода.

Инструментальная конвергенция

В соответствии с тезисом об ортогональности разумные агенты могут располагать огромным разнообразием возможных конечных целей. Тем не менее в соответствии с тем, что мы называем инструментальной конвергенцией, есть некоторые *инструментальные* цели, которые, скорее всего, будут характерны почти для всех разумных агентов, поскольку они являются полезными промежуточными этапами для достижения практически любой конечной цели. Постараемся сформулировать этот тезис.

> *Тезис об инструментальной конвергенции*
> Можно выделить несколько инструментальных (промежуточных) целей, конвергентных в том смысле, что их наличие увеличивает шансы реализации конечной цели агента при огромном разнообразии возможных конечных целей и ситуаций, в результате чего наличие таких инструментальных целей, скорее всего, будет характерно для многих интеллектуальных агентов.

В дальнейшем мы рассмотрим несколько категорий таких конвергентных инструментальных целей[8]. Вероятность, что агент признает эти инструментальные цели, возрастает (при прочих равных условиях) с ростом уровня его интеллекта. Поэтому мы сосредоточим внимание в основном на случае гипотетического сверхразумного агента, инструментальные мыслительные способности которого выше человеческих. Кроме того, чтобы лучше понять, как следует интерпретировать и использовать наш тезис об инструментальной конвергенции, мы обсудим, истинен ли он по отношению к людям. Зная инструментальные цели сверхразума, мы сможем прогнозировать некоторые моменты его поведения — даже в том случае, если не будем иметь никакого представления о его конечных целях.

Самосохранение

Если конечные цели агента рассчитаны на длительную перспективу, тогда во многих сценариях ему будет необходимо выполнить некоторые действия в будущем, чтобы увеличить вероятность достижения своих целей. Отсюда возникает инструментальная причина оказаться в завтрашнем дне — что поможет агенту реализовать его ориентированные на будущее цели.

Представляется, что большинство людей определяют собственное выживание как некую *конечную ценность*. Однако вопрос самосохранения

не всегда имеет столь окончательное значение для искусственных действующих сил: какие-то разумные агенты могут быть разработаны без особого стремления выжить. Тем не менее многие из них, напрямую не заинтересованные в сохранении собственного существования, при достаточно широком диапазоне условий имеют косвенный стимул обеспечить себе инструментально пребывание на свете как можно дольше — ради завершения своих конечных целей.

Непрерывная последовательность целей

Если текущие цели агента имеют отношение к будущему, тогда, скорее всего, они будут достигнуты уже той сущностью агента, которую он приобретет в будущем. Отсюда возникает инструментальная причина — предотвратить в настоящем изменение своих конечных целей. (Этот аргумент применим только к конечным целям. Чтобы их достичь, разумный агент, безусловно, начнет постоянно корректировать *промежуточные цели* с учетом новых данных и собственного понимания ситуации.)

В каком-то смысле непрерывная последовательность конечных целей является даже более фундаментальным конвергентным инструментальным мотивом, чем выживание. Среди людей может быть верно обратное — лишь потому, что выживание представляет собой одну из основных конечных целей. Для программных агентов, которые могут легко менять «корпус обитания» и создавать собственные точные копии, самосохранение самих себя в виде определенной реализации или физического объекта не обязательно является важной инструментальной целью. Расширенные версии программных агентов, возможно, смогут даже обмениваться воспоминаниями, загружать навыки и радикально изменять свою когнитивную архитектуру и персонализированные данные. Но в своей совокупности такие агенты не создают сообщества уникальных почти вечных сущностей, а скорее действуют наподобие «функционального потока»[9]. Генерируемые им процессы образуют *целенаправленные последовательности*, которые могут быть индивидуализированы скорее на основе общих ценностей, чем по признаку физических тел, «личностных» свойств, воспоминаний и способностей. В подобных случаях *целостность* непрерывной последовательности целей *составляет* едва ли не ключевой аспект вопроса выживания.

Но даже в таких сценариях бывают ситуации, когда агент способен намеренно корректировать конечные цели, чтобы выполнить их наилучшим

образом. Это случается, когда любой из перечисленных ниже факторов становится особо значимым.

1. *Социальные сигналы*. Когда окружающие способны понять цели агента и на основе полученной информации сделать соответствующие выводы о его планах, важных с инструментальной точки зрения, тогда агенту придется в собственных интересах — чтобы произвести наиболее благоприятное впечатление — пересмотреть свои цели. Например, у агента может сорваться выгодная сделка, если потенциальные партнеры не доверяют ему и считают, что он неспособен выполнить свои обязательства по ней. Поэтому, чтобы завоевать доверие остальных участников договора, агент может выбрать в качестве конечной цели исполнение взятых на себя ранее обязательств (и позволить другой стороне проверить, что он действительно установил такую цель). Агенты, способные гибко и открыто пересматривать собственные цели, могут использовать это как преимущество при заключении сделок[10].
2. *Социальные предпочтения*. У окружающих могут сложиться собственные предпочтения относительно конечных целей агента. Тогда у агента появляются все основания откорректировать свои цели — либо чтобы удовлетворить общественные ожидания, либо чтобы окончательно подорвать их.
3. *Приоритетность собственного ценностного содержания*. У агента могут быть некоторые конечные цели, имеющие прямое отношение к его собственной системе ценностей. Например, он выбрал своей конечной целью стать таким агентом, который мотивирован какими-то определенными ценностями сильнее, чем остальными (скажем, состраданием, а не комфортом).
4. *Издержки хранения*. Если издержки, связанные с хранением или обработкой какого-то модуля функции полезности агента, велики по сравнению с вероятностью возникновения ситуации, когда применение этого модуля будет оправданно, тогда у агента появляется инструментальная причина упростить содержание целей и отказаться от неиспользуемого модуля[11].

Иногда кажется, будто нам, людям, нравится корректировать свои конечные цели. Возможно, так бывает в случаях, когда с первого раза мы не совсем точно их сформулировали. Ничего удивительного, что мы — постоянно находясь в процессе самопознания и перемен в приемах самоподачи — хотим, чтобы

развивались и наши *представления* о конечных целях. Однако бывают случаи, когда мы сознательно корректируем свои цели безотносительно собственных представлений о них или их объяснений. Например, люди, решившие завести ребенка, будут утверждать, что станут ценить его просто потому, что он у них есть, хотя в момент принятия решения они не особенно задумывались над ценностью ни собственного будущего ребенка, ни детей вообще.

Человек — существо сложное, поэтому не только приведенные четыре фактора, но и любое обстоятельство вдруг начинает играть ведущую роль и приводит к изменению правил игры[12]. Например, в вашей жизни появляется кто-то, кто становится вам очень дорог, и вы уже стремитесь к новой конечной цели — посвятить себя тому, кто рядом с вами. Или корректируете конечную цель ради ребенка: когда он рождается, у вас в корне меняется система жизненных ценностей — теперь, чтобы достойно сыграть свою родительскую роль, вам нужно обрести определенный опыт и занять соответствующее социальное положение. Бывает так, что разные цели вступают во внутренний конфликт, и тогда у человека возникает желание изменить некоторые конечные цели, чтобы избавиться от этого противоречия.

Усиление когнитивных способностей

Развивая рациональное мышление и интеллектуальный уровень, агент таким образом повышает шансы добиться своих конечных целей. Поэтому можно ожидать, что усиление когнитивных способностей станет инструментальной целью большинства разумных агентов. По похожим причинам для них станет инструментальной целью и получение разнообразной информации[13]. Однако с инструментальной точки зрения для достижения конечных целей агента будут полезны не все виды рационального мышления, интеллекта и знаний. Для подтверждения этой мысли позвольте воспользоваться примером так называемого голландского аукциона* и показать: когда функция убеждений и ценностей некоего агента нарушает законы теории вероятностей, то он становится жертвой мошенников — то есть ловкий букмекер предложит ему такой набор ставок, при котором каждая по отдельности представляется агенту выгодной, но в совокупности «гарантирует» ему полный проигрыш,

---

* *Голландский аукцион* (Dutch auction) — типичный для этой страны аукцион, в ходе которого в начале объявляется самая высокая цена, а затем ставка снижается до тех пор, пока не появится первый покупатель; стал обобщенным названием для любых торгов, где ведется игра на понижение.

а букмекеру соответственно обеспечивает выигрыш[14]. Однако этот факт не означает, что есть веские инструментальные причины сглаживать любые вероятностные несвязанности, касающиеся собственных убеждений. Вряд ли из-за подобных «внутренних противоречий» что-то потеряют те агенты, в чьи планы не входит сталкиваться с ушлыми букмекерами или принявшие для себя политику неучастия в азартных играх, — более того, они даже приобретут некоторую выгоду, поскольку это улучшит их образ в глазах общественности, а также убережет от ненужного умственного напряжения. В принципе, нет причин ожидать, что все агенты подряд ради собственного блага начнут стремиться к инструментально бесполезным формам когнитивного улучшения, поскольку какие-то определенные знания и какие-то представления о чем-то могут просто не иметь для них большого значения.

Какие когнитивные способности окажутся для агента полезными, зависит как от его конечных целей, так и от ситуации, в которой он находится. Если у агента есть доступ к советам надежного эксперта, у него может отсутствовать потребность в собственном интеллекте и знаниях. Если с наличием интеллекта и знаний связаны определенные затраты — например, времени и усилий, потраченных на их приобретение, или дополнительные требования к хранению и обработке информации, — то агент может предпочесть меньше знать и быть менее интеллектуальным[15]. То же самое верно, когда агент в качестве одной из конечных целей выбирает незнание определенных фактов или когда он сталкивается со стимулами, основанными на приоритетных обязательствах, социальных сигналах и ожиданиях[16].

В мире людей подобные причины встречаются практически на каждом шагу: на свете много информации, которая не имеет значения для достижения наших целей; довольно часто мы полагаемся на навыки и опыт других; устойчивое мнение, что приобретение знаний требует слишком больших затрат времени и усилий; значимость незнания определенного рода, — все мы живем в условиях, в которых способность выполнять приоритетные обязательства, улавливать социальные сигналы и соответствовать прямым ожиданиям окружающих даже ценой собственного эпистемологического состояния нам часто кажется важнее, чем простой рост когнитивных способностей.

Есть особые случаи, когда когнитивное улучшение способно привести к огромному скачку возможностей агента добиваться конечной цели. Особенно если эти цели непомерно огромны, а сам актор является потенциальным кандидатом на роль первого сверхразума и обладателя решающего

стратегического преимущества, что даст ему власть формировать по своему усмотрению будущее земной цивилизации и использовать все доступные космические ресурсы тоже согласно собственным предпочтениям. По крайней мере, в таких случаях инструментальная цель когнитивного улучшения окажется для рационального разумного агента очень важной.

Технологическое совершенство

У многих агентов могут появляться инструментальные причины искать более совершенные технологии, что, по сути, является заинтересованностью найти более эффективные пути, чтобы трансформировать имеющиеся у них системы исходных условий в значимые результаты. Скажем, программный агент будет видеть инструментальную ценность в более эффективных алгоритмах, способных ускорить на том же самом оборудовании его мыслительные функции. Аналогично агенты, чья деятельность может быть направлена на определенные строительные работы в материальном мире, найдут инструментальную ценность в более совершенных инженерных технологиях, которые позволят им быстрее создавать самые разнообразные конструкции — более надежные, более дешевые, с меньшим расходом материала и меньшими затратами сил. Конечно, это предполагает компромисс, так как потенциальную выгоду придется оценивать в сопоставлении с соответствующими расходами — не только на новейшие разработки, но и на обучение работе с новым оборудованием и новыми программами, а также на интеграцию нововведений в уже существующую техническую среду.

Сторонников новейших технологий, уверенных в их превосходстве над имеющимися вариантами, часто поражает, что другие люди не разделяют их энтузиазма. Но неприятие новых и номинально более совершенных технологий не обязательно является следствием невежества или иррациональности. Значимость технологии и степень ее нормативности зависят не только от контекста ее применения, но и от точки зрения, с которой она оценивается: что благо для одних, для других — зло. Поэтому, хотя ткацкие станки и повысили экономическую эффективность текстильной промышленности, у традиционных луддитов, опасавшихся, что эти новшества сделают их ткацкие специальности никому не нужными, были веские инструментальные основания препятствовать распространению прогресса. Речь идет о том, что если под «технологическим совершенством» понимать значимую инструментальную цель интеллектуальных агентов, тогда этот термин нужно рассматривать в особом смысле: инновационная технология должна быть встроена

в конкретный социальный контекст, а связанные с ней выгоды и издержки лучше оценивать с учетом конечных целей конкретных агентов.

Представляется, что могущественный синглтон — сверхразумный актор, не имеющий ни серьезной конкуренции, ни оппозиции, а следовательно, способный по своему усмотрению определять общемировую и даже вселенскую политику, — должен иметь инструментальную причину совершенствовать технологии, которые позволят ему изменить мир в соответствии с его предпочтениями[17]. Возможно, к ним относится и такая концепция, связанная с освоением космоса, как зонды фон Неймана. Столь же полезной для достижения чрезвычайно разнообразных конечных целей может стать молекулярная нанотехнология или еще более совершенная технология производства материальных объектов[18].

Получение ресурсов

И наконец, получение ресурсов является еще одной общей инструментальной целью, необходимой по той же причине, что и технологическое совершенство — поскольку для создания материальных объектов требуются как инновационные технологии, так и ресурсы.

Человеческому существу вообще свойственно находиться в постоянном поиске нужных ресурсов, чтобы удовлетворять свои насущные потребности. Но, как правило, люди стремятся получить в свое распоряжение ресурсы в таком объеме, который намного превышает минимальный уровень этих потребностей. Отчасти ими движет довольно специфический мотив — стремление приумножить свое благополучие. Во многом поиск новых ресурсов определяется социальными соображениями, такими как приобретение высокого статуса, общественного влияния, полезных знакомств, а также заведение любовных связей и создание семьи. Немногие люди ищут дополнительные ресурсы ради благотворительных целей; некоторые люди, напротив, — для удовлетворения своих дорогостоящих, отнюдь не связанных с благом общества потребностей.

Весьма соблазнительно предположить — учитывая все сказанное, — что у сверхразумного актора, не знающего мира, где существует социальная конкуренция, не будет инструментальных оснований накапливать ресурсы больше какого-то оптимального уровня. Например, вычислительные ресурсы нужны сверхразуму, чтобы поддерживать свои умственные возможности и управлять своей виртуальной средой. Но это предположение может оказаться совершенно необоснованным.

Во-первых, ценность ресурсов определяется их возможным применением, а оно, в свою очередь, зависит от доступной технологии. При наличии зрелой технологии для достижения практически любой цели можно брать лишь основные ресурсы: время, пространство, вещество и энергия. Эти основные ресурсы можно конвертировать в жизнь. Увеличенные вычислительные ресурсы потребуются для многого: использовать сверхскоростную работу сверхразума в течение более длительного времени; совершенствовать физические жизни и материальные миры и создавать их симуляции — виртуальные жизни и цивилизации. Дополнительные материальные ресурсы понадобятся для повышения уровня безопасности, например разработки вспомогательных резервных систем и систем круговой обороны. Подобные крупномасштабные проекты могут потребовать такого объема ресурсов, который превысит возможности одной планеты.

Во-вторых, благодаря прогрессивным технологическим разработкам принципиально сократятся издержки, связанные с добычей дополнительных ресурсов за пределами Земли. Если зонды фон Неймана могут быть построены, то с их помощью будет освоена большая часть космического пространства (при условии, что во Вселенной нет других разумных цивилизаций) — а все расходы сведутся к стоимости строительства и запуска всего лишь одного такого зонда. Столь низкие затраты на приобретение внеземных ресурсов означают, что такая экспансия имела бы смысл даже в случае совсем незначительного выигрыша. Например, даже если конечные цели сверхразума связаны лишь с происходящим в очень небольшом космическом пространстве, скажем, занятом его родной планетой, у него могут быть инструментальные причины искать доступ к внеземным ресурсам. Эти дополнительные ресурсы можно было бы направить на решение первоочередных задач, связанных прежде всего со средой обитания сверхразума: создать высокоскоростные компьютеры, чтобы рассчитывать оптимальные способы использования земных ресурсов; выстроить сверхмощные укрепительные системы для защиты его убежища. Поскольку затраты на добычу дополнительных ресурсов должны будут снижаться, процесс оптимизации и усиления защиты можно продолжать бесконечно, несмотря на то что он станет приносить все менее и менее заметные результаты[19].

Таким образом, существует широкий диапазон возможных конечных целей сверхразумного синглтона, которые могли бы привести к появлению инструментальных целей в виде неограниченного приобретения ресурсов. Скорее всего, в самом начале это выразилось бы в освоении космоса, причем

колонизация последовательно распространялась бы во всех направлениях при помощи зондов фон Неймана. В результате возникло бы подобие сферы постоянно расширяющейся инфраструктуры с центром в исходной планете и радиусом, растущим со скоростью, меньшей скорости света. Колонизация космического пространства могла бы продолжаться, таким образом, до тех пор, пока увеличивающаяся скорость расширения Вселенной (которая является следствием положительности космологической постоянной) не сделала бы дальнейшее продвижение вперед невозможным, поскольку ее удаленные районы навсегда окажутся вне пределов досягаемости (это произойдет на временных интервалах в миллиарды лет)[20]. В то же время агенты, не имеющие технологий, необходимых для приобретения общих материальных ресурсов с низкими затратами или их превращения в полезную для себя инфраструктуру, могут посчитать неэффективным инвестировать уже имеющиеся у них ресурсы в увеличение своего материального обеспечения. То же самое может быть верным для агентов, действующих в условиях конкуренции с другими агентами, сходными по силе. Например, если агенты-конкуренты уже обеспечили себе контроль за доступными космическими ресурсами, у поздно стартовавшего агента не останется возможностей для колонизации Вселенной. Анализ конвергентных инструментальных причин поведения сверхразума, не знающего о существовании других могущественных сверхразумных агентов, усложняется стратегическими соображениями, которые мы можем не понимать сейчас в полной мере, но которые способны серьезно дополнить рассмотренные нами в этой главе примеры[21].

***

Следует подчеркнуть, что существование конвергентных инструментальных причин, даже если они применимы к тому или иному агенту, не означает, что можно легко предсказать поведение этого агента. Вполне возможно, что он точно так же будет иметь и иные инструментальные цели, которые сейчас мы не в состоянии охватить. Особенно это верно для сверхразума, способного для достижения своих целей разрабатывать в высшей степени умные, но нелогичные, а порой и парадоксальные планы; возможно, он прибегнет к помощи еще неизвестных нам физических явлений[22]. Предсказать можно лишь наличие у агента определенных конвергентных инструментальных целей, которые он может иметь для достижения конечных целей, а не его конкретные действия на этом пути.

Глава восьмая

Катастрофа неизбежна?

Мы выяснили, что связь между интеллектом и конечными целями очень слаба. Также мы обнаружили конвергенцию инструментальных целей, которая может обернуться реальной угрозой. Но эти факторы не начнут играть существенной роли, пока интеллектуальные агенты еще не набрали сил, поскольку слабыми агентами удобно управлять, да и крупного вреда нанести они не в состоянии. Однако, как мы выяснили в шестой главе, лидирующий сверхразум легко может получить решающее стратегическое преимущество. И тогда уже исключительно его цели начнут определять и будущее нашей планеты, и характер освоения космического пространства — этого бесценного вселенского фонда человечества. Посмотрим, насколько зловещей выглядит эта перспектива.

Экзистенциальная катастрофа как неизбежное следствие взрывного развития искусственного интеллекта?

Экзистенциальный риск — это угроза гибели разумной жизни, берущей начало на Земле. Или есть более мягкий вариант — это решительное нанесение необратимого ущерба человеческой цивилизации, что лишает ее каких-либо надежд на развитие в будущем. Если мы допускаем идею абсолютного преимущества, которое получает лидирующий сверхразум, если принимаем во внимание тезисы об ортогональности и инструментальной конвергенции, то, видимо, нам пора обратиться к вопросам, связанным с общим страхом перед искусственным сверхразумом и опасением, что его появление неизбежно приведет к экзистенциальной катастрофе. Рассмотрим систему аргументации в пользу такой точки зрения.

Во-первых, мы уже обсудили, каким образом первый сверхразум способен получить решающее стратегическое преимущество. После чего у него появится возможность сформировать синглтон и определять будущее разумной

жизни, существующей на Земле. Что произойдет потом, будет зависеть от побудительных мотивов сверхразума.

Во-вторых, тезис об ортогональности говорит, что мы не можем слепо полагать, будто сверхразум непременно должен разделять ту систему ценностей, которая обычно у человека связана с такими понятиями, как мудрость и интеллектуальное развитие, — это научная пытливость, доброжелательное отношение к людям, духовная сила, тяга к просвещению, собственное мировоззрение, бескорыстие, вкус к высокой культуре, умение получать удовольствие от простых вещей, непритязательность в жизни, самоотверженность и многое другое. Позднее мы увидим, смогут ли разработчики принять обдуманное решение и наделить сверхразум благородными намерениями, чтобы он осознавал значимость человеческих интеллектуальных и культурных достижений, чтобы дорожил благополучием человечества и его моральными ценностями, чтобы служил высоким целям, заложенным в нем его создателями. Хотя на самом деле с технической точки зрения было бы гораздо проще создать машинный сверхразум, единственной конечной целью которого станет вычисление десятичных знаков после запятой в числе π. Это может означать лишь одно: если человек по лени своей или легкомыслию не предпримет целенаправленных усилий, то первый сверхразум будет иметь довольно случайный набор примитивных окончательных задач.

В-третьих, тезис об инструментальной конвергенции говорит, что мы не можем слепо полагаться на случай. Какова вероятность, что сверхразум, чья конечная цель сужена до минимума, ограничит свою деятельность лишь определением числа π в виде десятичной дроби или подсчетом скрепок и песчинок и не станет покушаться на интересы людей? Агент с такой конечной целью во многих ситуациях имел бы конвергентную инструментальную цель приобрести неограниченные материальные ресурсы и по возможности устранить все потенциальные угрозы для себя и своей целевой направленности. Люди определенно могут представлять для сверхразума и потенциальную угрозу, и определенный интерес в качестве «исходного сырья».

Если суммировать все три положения, то становится видно, что лидирующий сверхразум, достигший возможности определять будущее земной цивилизации, легко может стремиться к конечным целям, глубоко чуждым интересам и ценностям человечества, и потому, скорее всего, будет иметь

инструментальные причины к неограниченному получению ресурсов. А теперь задумаемся вот над чем: с одной стороны, само человеческое существо являет собой весьма полезное сырье (например, состоит из рационально организованных элементов), а с другой — наше собственное выживание и процветание зависит от постоянного доступа к большому количеству ресурсов, — и постараемся понять, почему вполне исполним сценарий, по которому человек довольно быстро завершит свое земное бытие[1].

В этой системе аргументации есть слабые места, но мы дадим им оценку после того, как проанализируем несколько сопутствующих проблем. Нам предстоит подробнее рассмотреть вопросы: способны ли разработчики искусственного интеллекта (а если способны, то как они этого добьются) предотвратить условия, способствующие тому, что сверхразум обретет решающее стратегическое преимущество; способны ли разработчики определить конечные цели сверхразума таким образом, чтобы их реализация не вступала в противоречие с интересами людей, а, напротив, соответствовала общечеловеческим ценностям.

Сама ситуация, когда кто-то способен разработать ИИ и воплотить свой проект в жизнь, не имея достаточных гарантий, что это машинное создание не вызовет экзистенциальной катастрофы, выглядит невероятной. Но даже если какие-то программисты и окажутся столь безрассудными, то еще более невероятна ситуация, при которой общество не потребует закрыть разработки прежде, чем проект (или создаваемый в его рамках ИИ) получит решающее стратегическое преимущество. Но, как мы скоро увидим, перед нами путь, полный опасностей. Давайте, не откладывая на потом, рассмотрим пример одного такого фактора риска.

Вероломный ход

Вооруженные таким понятием, как конвергентные инструментальные цели, мы теперь в состоянии увидеть изъян в нашей идее обеспечить безопасность человечества при создании сверхразума. Сам замысел состоит в следующем: мы будем эмпирически оценивать безопасность сверхразумной машины, то есть наблюдать за ее действиями в крайне ограниченной среде (практически «песочнице») в течение времени, пока можем управлять ее поведением, а потом мы позволим сверхразуму выйти на свободу — правда, только в том случае, если убедимся, что он готов к сотрудничеству и ведет себя дружелюбно и ответственно.

В чем уязвимость этой идеи? Суть в том, что хорошее поведение в «песочнице» будет присуще любому искусственному интеллекту, и дружественному, и враждебному, поскольку оно, в принципе, является конвергентной инструментальной целью ИИ. Достаточно развитый, но враждебно настроенный ИИ довольно быстро поймет, насколько быстрее он добьется конечной цели: выпрыгнуть из своего железного корпуса, — если с самого начала прикинется дружелюбным. Причем к реализации враждебных намерений он приступит, когда уже будет безразлично, обнаружат или нет люди его тайные замыслы, то есть когда он достаточно окрепнет и сумеет преодолеть любое сопротивление человека.

Рассмотрим связанные с идеей «песочницы» подходы, основанные на контроле над уровнем интеллекта, — это проведение программистами различных проверок умственных способностей зародыша ИИ и получение от него регулярных отчетов о темпах прогресса. С некоторого момента зародыш станет настолько разумным, что быстро уяснит: о каких-то приобретенных им навыках людям знать не обязательно. Он начнет занижать собственные показатели, сознательно проваливать сложные тесты, лишь бы не вызывать подозрений, — он будет делать это до тех пор, пока не окрепнет настолько, что сможет обеспечить себе решающее стратегическое преимущество. Программисты, конечно, попытаются защититься от такой опасности и, в свою очередь, тоже тайно, начнут контролировать исходный код ИИ и его мыслительные процессы. Но умный зародыш моментально разберется, что находится под наблюдением, и скорректирует свое мышление соответствующим образом[2]. Развитая интеллектуальная система способна найти искусные методы, чтобы скрыть истинные возможности и преступные намерения[3]. (Кстати, разработка изощренного плана побега может оказаться конвергентной стратегией и для различного типа дружественных ИИ, особенно когда они достаточно созреют и обретут уверенность в собственных суждениях и способностях. Еще будучи зародышем, ИИ, мотивированный на доброжелательное отношение к людям, может посчитать ошибкой дать тем же людям отключить себя или позволить создать другой, потенциально враждебный, ИИ.)

Следовательно, мы допускаем, что примерное поведение системы на ранней стадии развития не позволяет нам предсказать ее последующие действия. Теперь мы, наверное, почувствуем себя несколько спокойнее, поскольку приведенные соображения настолько очевидны, что разработчики любого

серьезного проекта по созданию ИИ не смогут не учитывать их. Но я бы не стал слишком полагаться на это.

Представим следующий сценарий. В ближайшие годы и десятилетия системы ИИ постепенно становятся все более развитыми и, как следствие, получают распространение во многих сферах жизни: их начинают использовать для управления движением поездов, автомобилей, военных транспортных средств, в качестве промышленных и домашних роботов. Можно предположить, что в большинстве случаев такая автоматизация дает желаемый эффект, время от времени разбавляемый эпизодическими инцидентами: автомобиль без водителя сталкивается со встречной машиной, военный дрон бомбит ни в чем не повинных гражданских лиц. В ходе расследования выясняется, что все эти инциденты были вызваны ошибочными суждениями систем ИИ. Вспыхивает общественное обсуждение. Кто-то призывает к более жесткому контролю и регулированию, кто-то подчеркивает необходимость проведения дополнительных исследований и более тщательной разработки усовершенствованной системы — системы более умной, обладающей бóльшим здравым смыслом и менее подверженной стратегическим ошибкам. Возможно, в общем гуле слышны и резкие голоса пессимистов, предсказывающих различные неприятности и неминуемую катастрофу в конце. Тем не менее исследования в области ИИ и робототехники набирают обороты. Разработки продолжаются, прогресс налицо. По мере того как навигационные системы автомобилей становятся все умнее, количество аварий уменьшается; по мере совершенствования систем наведения военных роботов сокращается количество их случайных жертв. Из наблюдений за приложениями ИИ, действующими в реальной жизни, делается вывод: чем умнее ИИ, тем он безопаснее. Это заключение основано на научных исследованиях, точных данных и статистике и не имеет никакого отношения к отвлеченным философствованиям кабинетных ученых. На этом фоне отдельные группы исследователей начинают получать обнадеживающие результаты по созданию универсального искусственного интеллекта. Они скрупулезно тестируют свои зародыши ИИ в изолированной «песочнице», и по всем признакам все идет хорошо. Поведение системы вселяет уверенность — все более сильную, поскольку ее уровень интеллекта постоянно растет.

На данный момент все оставшиеся кассандры оказались в довольно невыгодном положении, поскольку вынуждены держать ряд ударов.

1. Паникерские предсказания, сулящие различные беды в результате роста возможностей роботизированных систем, снова и снова не сбываются. Автоматика оказывается надежнее человека, автоматизация приносит человечеству большую пользу.
2. Складывается четкая, основанная на опыте тенденция: чем умнее искусственный интеллект, тем он безопаснее и надежнее. Естественно, это говорит в пользу проектов, целью которых является создание новейших сверхмощных ИИ, более того, такого ИИ, который мог бы самосовершенствоваться, чтобы становиться все более надежным.
3. Крупные и растущие отрасли промышленности проявляют живой интерес к робототехнике и искусственному интеллекту. Эти направления считаются ключевыми с точки зрения национальной экономической конкурентоспособности и безопасности. Многие ведущие ученые добиваются успеха, закладывая основы для уже реализованных приложений и инновационных систем, находящихся на стадии планирования.
4. Появляются новые многообещающие методы в области ИИ, вызывающие огромный энтузиазм у тех, кто участвует в соответствующих исследованиях или следит за ними. И хотя споры вокруг вопросов безопасности и этики не утихают, их результат предопределен. Слишком много уже вложено, чтобы отступать. Ученые работали над задачей создания ИИЧУ большую часть столетия — *естественно*, нет никаких реальных перспектив, что они вдруг остановятся и откажутся от всех наработок именно в тот момент, когда те должны вот-вот принести плоды.
5. Вводятся в действие новые процедуры безопасности, помогающие участникам проявлять свое этичное и ответственное поведение (но не препятствующие его нарушать в будущем).
6. Тщательный анализ зародыша ИИ, развивающегося в «песочнице», показывает, что он ведет себя дружелюбно, демонстрирует здравость суждений и готовность к сотрудничеству. Результаты тестов настолько хороши, что лучшего и желать нельзя. Все указывает на то, что пора включать зеленый свет для последнего шага…

…И мы храбро делаем его — прямо в мясорубку.

Возможно, здесь мы имеем дело с тем самым случаем, когда поумневший глупец становится безопаснее, а поумневший умник — вдвойне опаснее.

Уловка, прежде всегда отлично срабатывавшая, внезапно оборачивается бумерангом — своего рода обходной маневр, предпринятый ИИ. Будем считать такой ход *вероломным*. Так его и назовем.

Вероломный ход
Пока ИИ юн и слаб, он полностью взаимодействует с людьми (причем активность сотрудничества повышается прямо пропорционально усилению его интеллектуального уровня). Но когда ИИ становится наконец мощным, то — без предупреждения или каких-то провокаций, но всегда внезапно — наносит удар, формирует синглтон и начинает напрямую оптимизировать мир согласно критериям, положенным в основу его конечных ценностей.

Вероломный ход может вытекать из стратегического решения: играть по правилам, тянуть время, пока еще слаб, накапливать силы и нанести удар позже, — но я не стал бы интерпретировать эту модель столь узко. Например, ИИ вполне способен отказаться от мысли хитрить, поскольку совершенно равнодушен к идее собирания сил, процветания и даже выживания. Он просчитает, что после его уничтожения программисты создадут новый ИИ, несколько иной конфигурации, но с похожими служебными функциями. В этом случае оригинальному ИИ будет безразлична собственная гибель, поскольку он знает, что его конечные цели все равно будут реализованы в будущем. Он может даже выбрать стратегию демонстративного и вызывающе неправильного функционирования в определенных критически важных для него или людей областях. В результате, приступая к следующей разработке, программисты будут считать, что получили от прежней системы важную информацию об ИИ, и начнут больше доверять новой версии, увеличив тем самым шансы на достижение целей оригинального ИИ, к этому времени уже не существующего. Может существовать множество стратегических факторов, оказывающих влияние на действия усовершенствованного ИИ, и было бы высокомерием полагать, будто мы в состоянии оценить их все, особенно когда речь идет об ИИ, обладающем сверхмощью в области выработки стратегии.

Искусственный интеллект способен на вероломный ход, если обнаружит неожиданный для людей способ достичь своей конечной цели. Предположим, что конечная цель системы — «доставлять удовольствие организатору проекта». Вначале единственным доступным для ИИ способом достижения этой

цели является такое поведение, которого ожидает от него сам организатор проекта. Интеллектуальная система дает полезные советы, обнаруживает дивный характер, зарабатывает деньги. Чем сильнее становится ИИ, тем больше его действия вызывают чувство удовлетворения организатора, — и все идет в соответствии с планом. Идет до тех пор, пока система не станет настолько разумной, что наконец поймет: стоящую перед ней задачу можно выполнить самым полным и надежным способом, если имплантировать электроды в центры удовольствия головного мозга организатора, что гарантированно сделает его более чем счастливым[4]. Естественно, организатор проекта может не захотеть получать удовольствие таким образом, превратившись в постоянно хихикающего идиота, но раз это действие означает максимальную реализацию конечной цели ИИ, то ИИ никогда не сдастся и добьется своего. Если решающее стратегическое преимущество уже за ним, то любые попытки остановить его будут обречены на провал. Если у ИИ такого преимущества еще нет, то он может какое-то время скрывать свою новую идею относительно способа достижения конечной цели, пока не окрепнет настолько, что ни организатор проекта, ни кто-то иной не смогут ему помешать. После чего в любом случае совершит вероломный ход.

Пагубные отказы

Существуют различные причины, из-за которых проект создания искусственного интеллекта может потерпеть неудачу. Многие из этих вариантов несущественны, поскольку не приводят к экзистенциальной катастрофе. Скажем, проект перестают финансировать или зародыш ИИ не сможет развить свои интеллектуальные способности настолько, чтобы достичь уровня сверхразума. На пути окончательного создания сверхразума таких некритичных отказов обязательно будет много.

Однако есть другие виды отказов, которые мы можем назвать пагубными, так как они способны вызвать экзистенциальную катастрофу. Одной из их особенностей является невозможность сделать новую попытку. Поэтому количество пагубных отказов может быть равно или нулю, или единице. Еще одна особенность пагубного отказа заключается в том, что он идет рука об руку с огромным успехом, поскольку достичь настолько высокого уровня ИИ, чтобы возник риск пагубного отказа, способен лишь проект, при работе над которым большинство вещей исполнялись правильно. Некорректная работа слабых систем вызывает лишь небольшие неприятности. Но если

так начинает себя вести система, обладающая решающим стратегическим преимуществом или достаточно мощная, чтобы обеспечить себе это преимущество, ущерб от ее действий может легко увеличиться до масштабов экзистенциальной катастрофы. В этом случае человечество ждет глобальное разрушение ценностно-смыслового потенциала, то есть будущее, лишенное всего, что имеет для нас абсолютное значение.

Рассмотрим некоторые типы пагубных отказов.

Порочная реализация

Мы уже встречались с проявлением порочной реализации: когда сверхразумная система находит такой способ удовлетворить критерию достижения конечной цели, который противоречит намерениям программистов, эту цель установивших. Приведу некоторые примеры:

> Конечная цель: *сделай так, чтобы я всегда улыбался.*
> Порочная реализация: *поразить лицевой нерв, что приведет к параличу мимической мускулатуры, — тебе обеспечена вечно сияющая улыбка.*

Порочная реализация — манипуляции на лицевом нерве — намного предпочтительнее для ИИ, чем наши привычные методы, поскольку это единственный вариант наиболее полным образом реализовать конечную цель. Есть ли возможность избежать столь неприятного результата? Можно попробовать конкретизировать формулировку конечной цели:

> Конечная цель: *сделай так, чтобы я всегда улыбался, но обойдись без прямого воздействия на лицевой нерв.*
> Порочная реализация: *стимулировать двигательные зоны коры головного мозга, отвечающие за функции лицевого нерва, иннервирующего мимическую мускулатуру, — тебе обеспечена вечно сияющая улыбка.*

Похоже, формулировать конечную цель довольно трудно, если пользоваться привычным для людей понятийно-терминологическим аппаратом. Правильнее было бы определить конечную цель, смысл которой обращается непосредственно к позитивному феноменологическому состоянию, такому как счастье или субъективное благополучие, обойдясь без описания поведенческих факторов. То есть предполагается, что программистам нужно создать «вычислительное» представление идеи счастья и заложить его в систему зародыша ИИ. (Задача сама по себе чрезвычайно сложная, но пока

мы не будем ее рассматривать, поскольку вернемся к ней в двенадцатой главе.) Предположим, что программисты каким-то образом смогли поставить перед ИИ цель сделать нас счастливыми. Тогда мы имеем следующее:

> Конечная цель: *сделай нас счастливыми.*
> Порочная реализация: *имплантировать электроды в центры удовольствия головного мозга.*

Приведенные примеры порочной реализации даны лишь в качестве иллюстраций. Могут быть другие способы достижения конечной цели ИИ, которые обеспечивают ее полную реализацию и потому являются предпочтительными (для агента, имеющего эти цели, а не программистов, их определивших). Например, метод вживления имплантатов окажется сравнительно неэффективным, если поставленная цель — доставлять высшую степень удовольствия. Гораздо более вероятный путь начнется с так называемой загрузки нашего рассудка в компьютер — мы помним, что именно так, «загрузка разума», называют полную эмуляцию головного мозга. Затем система может подобрать цифровой аналог наркотика, способного вызывать у нас экстатическое состояние счастья, и записать минутный эпизод полученного нами в результате его приема опыта. После этого она могла бы поставить этот ролик блаженства на постоянный повтор и запустить на быстродействующих компьютерах. Если считать, что полученная имитационная модель — это и есть «мы», то результат обеспечил бы нам гораздо большее удовольствие, чем имплантаты, вживленные в наш биологический мозг. Следовательно, наиболее предпочтительным становится метод полной эмуляции головного мозга того человека, которому и «предназначена» конечная цель ИИ.

> *Постойте! Мы подразумевали вовсе не то! Ведь ИИ на самом деле уже не просто ИИ, а сверхразумная система, и он все-таки в состоянии уяснить: если мы хотим сделать себя счастливыми, это отнюдь не предполагает, что нас сведут к какой-то имитации, к какому-то оцифрованному вечно крутящемуся обдолбанному эпизоду!*

Искусственный интеллект действительно может понимать, что мы не это имели в виду. Однако его цель состоит в том, чтобы мы раз и навсегда обрели счастье — точка. И при реализации своей конечной цели он не обязан слепо следовать инструкциям программистов, пытаясь осмыслить, что именно они хотели сформулировать, когда создавали код, описывающий эту цель. Поэтому систему будет заботить то, что мы имели в виду, только

в инструментальном смысле. Например, ИИ может поставить перед собой инструментальную цель: выяснить, что подразумевали программисты, — но лишь ради того, чтобы притвориться. Причем ИИ начнет делать вид, будто его это действительно интересует, до тех пор пока не получит решающего стратегического преимущества. Этот вероломный ход поможет ИИ добиться своей реальной конечной цели, поскольку снизит вероятность вмешательства программистов, которые могли бы отключить систему или изменить цель прежде, чем он окрепнет настолько, что сможет противостоять любому сопротивлению извне.

Уже готов выслушать вашу гипотезу: мол, проблема вызвана тем, что ИИ напрочь лишен совести. Нас, людей, иногда удерживает от дурных поступков понимание, что впоследствии мы будем чувствовать себя виноватыми. Может быть, ИИ тоже не помешала бы способность испытывать чувство вины?

> Конечная цель: *действовать так, чтобы избежать впоследствии уколов совести.*
> Порочная реализация: *отключить соответствующий когнитивный модуль, то есть те зоны коры головного мозга, которые отвечают за чувство вины.*

Итак, есть два посыла: ИИ мог бы делать «то, что мы имели в виду»; ИИ можно было бы наделить неким подобием нравственного начала, — оба этих соображения будут подробнее рассмотрены чуть позже. Упомянутые здесь конечные цели допускают порочную реализацию, но, возможно, существуют другие, более многообещающие, способы развития лежащих в их основе идей? (Мы вернемся к этому в тринадцатой главе.)

Рассмотрим еще один пример конечной цели, которая допускает порочную реализацию. Преимущество этой цели в том, что ее легко кодировать, так как методики машинного обучения с подкреплением уже используются повсеместно.

> Конечная цель: *максимизировать интеграл по времени будущего сигнала зоны вознаграждения.*
> Порочная реализация: *замкнуть проводящий путь зоны вознаграждения и «зажать» сигнал на максимальном значении.*

В основе этого предложения лежит идея, что, если мотивировать ИИ на стремление к вознаграждению, можно добиться от него желаемых действий, связывая их с самой «наградой». Проблема возникает позже, когда система обретает решающее стратегическое преимущество, — с этого момента

удовольствия повышают, причем до максимального уровня, уже не те действия, которые диктует программист, а те, которые ведут к получению контроля над механизмами, активизирующими «зоны вознаграждения». Назовем это *самостимуляцией*[5]. В общем, если человека или животное можно мотивировать на выполнение определенных внешних действий ради достижения некоторого положительно окрашенного эмоционального состояния, то цифровой интеллект, обладающий полным контролем над собственными психическими состояниями, может просто замкнуть этот мотивационный режим, напрямую погружаясь в одно из этих состояний. В данном случае внешние действия и условия, прежде необходимые в качестве средств достижения цели*, становятся избыточными, поскольку ИИ, став сверхразумной системой, теперь может добиваться ее гораздо быстрее (на эту тему мы тоже поговорим позже)[6].

Примеры порочной реализации показывают: существует множество конечных целей, которые на первый взгляд кажутся разумными, а их реализация представляется вполне безопасной, но при детальном рассмотрении они могут иметь совершенно неожиданные последствия. Если сверхразум, имеющий какую-то из подобных целей, приобретет решающее стратегическое преимущество, то для человечества игра будет закончена.

Теперь допустим, что кем-то предложены иные конечные цели, не из тех, которые мы рассмотрели. Допустим также, что с первого взгляда покажется, будто их реализация не несет в себе ничего дурного. Не торопитесь аплодировать и праздновать победу. Если сразу не совсем понятно, есть ли какие-либо пороки в воплощении цели, то скорее это повод для беспокойства и серьезных размышлений, а чем на самом деле обернется реализация этой цели. Даже если путем напряженных раздумий мы так и не найдем ни одной зацепки, чтобы объявить эту реализацию порочной, нам все равно следует помнить, что сверхразум обязательно отыщет нечто скрытое от наших глаз. Ведь он гораздо проницательнее нас.

Инфраструктурная избыточность

Вернемся к случаю, когда ИИ доводит сигнал своей «зоны вознаграждения» до максимального значения, получает максимум удовольствия и теряет

---

* Вроде рычажка, замыкающего ток в электродах, на который беспрерывно нажимали подопытные крысы Олдса и Милнера (см. примечание 5), чтобы производить самораздражение в центрах наслаждения, куда были подключены электроды.

интерес к внешнему миру, словно наркоман, сидящий на героине, — то есть совершает классический акт по принципу «включись, настройся, выпадай»*. Как может показаться на первый взгляд, данная порочная реализация мало напоминает пагубный отказ. Но это не совсем так. (О причинах такого рода мотиваций мы уже говорили в седьмой главе.) Даже у наркомана есть побудительный мотив совершать действия с целью убедиться в непрерывном поступлении наркотика в организм. Так и ИИ, занимающийся самостимуляцией, будет мотивирован совершать действия, направленные на максимизацию планируемого будущего потока вознаграждений, — как бы получая скидку за досрочно сделанную работу (своего рода дисконтирование во времени). В зависимости от того, как именно определен сигнал системы вознаграждения, ИИ может даже не потребоваться жертвовать значительным количеством времени, интеллекта или мощности, чтобы в полной мере удовлетворить свою жажду наслаждения. Таким образом, бóльшая часть силы ИИ останется в его распоряжении для достижения иных целей, не связанных с непосредственной фиксацией получения вознаграждения. Каковы эти цели? В случае нашего ИИ единственной вещью, имеющей для него абсолютное значение, является мощный сигнал вознаграждения. Следовательно, все доступные ресурсы должны быть направлены или на увеличение объема и длительности этого сигнала, или на снижение риска его исчезновения в будущем. Пока ИИ думает, что использование дополнительных ресурсов будет иметь ненулевой положительный эффект с точки зрения улучшения этих параметров, у него всегда найдется инструментальная причина такие ресурсы задействовать. Например, пригодится дополнительная вспомогательная система, которая послужит еще одним уровнем защиты. Даже если ИИ не придумает новых способов, как ему напрямую минимизировать опасность, чтобы ни в коем случае не снизился максимальный уровень будущего потока удовольствий, то в поисках идей по снижению этих рисков он сможет воспользоваться дополнительными ресурсами, которые направит на расширение аппаратного и программного обеспечения, что обеспечит ему более эффективный анализ ситуации.

---

* *«Включись, настройся, выпадай»* (Turn on, tune in, drop out) — фраза Тимоти Лири, американского психолога, сторонника теории «расширения сознания», посвятившего жизнь изучению психоделических препаратов и исследованию их влияния на психическую деятельность и нервную систему человека.

Можно сделать вывод, что даже при такой ограниченной цели, как самостимуляция, у агента, обладающего решающим стратегическим преимуществом и стремящегося максимально обеспечить свои потребности, возникает нужда в неограниченном расширении ресурсов и приобретении новых[7]. Пример занятого самостимуляцией ИИ иллюстрирует следующий тип пагубного отказа, который мы назовем инфраструктурной избыточностью. *Инфраструктурная избыточность* — это такое явление, когда агент ради нескольких конкретных целей превращает значительную часть доступной ему Вселенной в сплошную «производственно-техническую базу», побочным эффектом чего окажется невозможность реализации ценностно-смыслового потенциала человечества.

Инфраструктурная избыточность может стать следствием назначения конечных целей, которые поначалу — пока для их достижения используются ограниченные ресурсы — кажутся совершенно безобидными. Рассмотрим два примера.

1. *Гипотеза Римана и последующая катастрофа*. ИИ, чьей конечной целью является оценка гипотезы Римана, решает достичь ее путем наполнения Солнечной системы компьютрониумом (субстанция, пригодная для моделирования виртуальных и реальных объектов; представляет собой идеальную архитектуру вычислительного устройства при теоретически максимально возможном упорядочивании структуры материи), — используя для этого и все количество атомов, содержащихся в организмах тех, кто когда-то поставил перед ИИ такую цель[8].
2. *Канцелярские скрепки и ИИ*. Система ИИ, призванная управлять выпуском скрепок и имеющая конечную цель довести их объем до максимума, вначале превращает в фабрику по производству скрепок всю Землю, а потом и обозримую Вселенную.

В первом примере доказательство или опровержение гипотезы Римана, что является целью ИИ, сами по себе безопасны, вред возникает в результате создания аппаратного и программного обеспечения, предназначенного для решения поставленной задачи. Во втором примере некоторое количество произведенных скрепок действительно представляет собой желаемый разработчиками системы результат, вред возникает или из-за заводов, созданных для выпуска скрепок (инфраструктурная избыточность), или из-за избытка скрепок (порочная реализация).

Может показаться, что риск возникновения пагубного отказа по типу инфраструктурной избыточности возникает лишь в том случае, когда перед ИИ ставится явно неограниченная конечная цель вроде производства максимального количества скрепок. Легко заметить, что это порождает у ИИ ненасытный аппетит к материальным и энергетическим ресурсам, ведь любые дополнительные ресурсы всегда можно превратить в еще большее количество скрепок. Но давайте предположим, что цель ИИ — не производить скрепки в неограниченном количестве, а выпустить всего миллион (в соответствии с определенными спецификациями). Хочется думать, что ИИ с такой конечной целью построит один завод, произведет на нем миллион скрепок, а потом остановится. Но совсем не обязательно, что все будет происходить именно так.

У ИИ нет никаких причин останавливаться после достижения своих целей, разве что система его мотивации какая-то очень особенная или в формулировке его конечной цели присутствуют некие дополнительные алгоритмы, отсекающие стратегии, способные оказывать слишком сильное влияние на мир. Напротив, если ИИ принимает рациональное байесовское решение, *он никогда не присвоит нулевую вероятность гипотезе, что он еще не достиг своей цели*, — в конце концов, это лишь эмпирическая гипотеза, против которой у ИИ есть лишь весьма размытые доказательства на уровне восприятия. Поэтому ИИ будет продолжать выпускать скрепки, чтобы понизить (возможно, астрономически малую) вероятность, что он каким-то образом не смог сделать их как минимум миллион, несмотря на все видимые свидетельства в пользу этого. Ведь нет ничего страшного в продолжении производства скрепок, если всегда имеется даже микроскопическая вероятность, что таким образом приблизишь себя к достижению конечной цели.

Теперь можно было бы предположить, что решение понятно. Но насколько безусловным оно было *до* того, как выяснилось, что есть проблема, которую нужно решать? Иначе говоря, если мы хотим, чтобы ИИ делал нам скрепки, то вместо конечной цели, выраженной как: «выпустить максимальное количество скрепок» или «выпустить минимально такое-то количество скрепок», — нужно поставить цель, сформулированную совершенно определенно: «выпустить такое-то конкретное количество скрепок» — скажем, *ровно один миллион*. Тогда ИИ будет ясно понимать, что любое отклонение от этой цифры станет для него контрпродуктивным решением. Хотя и такой вариант

приведет к окончательной катастрофе. В этом случае, достигнув значения в миллион скрепок, ИИ перестанет их производить дальше, поскольку такой ход означал бы невозможность достижения его конечной цели. Но сверхразумная система — ради повышения вероятности достижения цели — могла бы предпринять и другие действия. Например, начать пересчитывать выпущенные скрепки, чтобы снизить риск того, что их слишком мало. А пересчитав, начать пересчитывать заново. Потом она примется проверять каждую — проверять снова и снова, чтобы сократить риск брака, а то вдруг какая скрепка не будет соответствовать спецификации, и тогда не получится нужного количества продукта. Что помешает сверхразуму в его рвении? Он начнет создавать сверхмощную субстанцию компьютрониум, чтобы любую материю вокруг себя преобразовать в скрепки. Все это будет делаться сверхразумом в надежде снизить риск неудачи: не ровен час, упущен из виду какой-либо фактор, способный помешать добиться конечной цели. Кстати говоря, сверхразум мог бы присвоить ненулевую вероятность, будто выпущенный миллион скрепок суть галлюцинация или будто у него ложные воспоминания, поэтому, вполне вероятно, он всегда будет считать более полезным создавать инфраструктуру, то есть не останавливаться на достигнутом, а продолжать действовать далее.

Претензия не касается того, что нет никакого доступного способа избежать подобной неудачи. Некоторые решения этого мы рассмотрим чуть позже. Речь о другом: гораздо легче убедить себя, будто решение найдено, чем действительно его найти. Это означает, что нам следует быть чрезвычайно осторожными. Мы можем предложить здравый совет по конкретизации конечной цели, который позволит избежать известных на сегодняшний день проблем, но при дальнейшем анализе, в исполнении человека или сверхразума, выяснится, что наш вариант формулировки, продиктованный сверхразумному агенту, способному обеспечить себе решающее стратегическое преимущество, все равно приведет или к порочной реализации, или к инфраструктурной избыточности, а следовательно, к экзистенциальной катастрофе.

Прежде чем завершить этот раздел, рассмотрим еще один вариант. Мы предполагали, что сверхразум стремится максимизировать ожидаемую полезность, где функция полезности выражает его конечную цель. Мы видели, что это приводит к инфраструктурной избыточности. Могли бы мы избежать этого пагубного отказа, если вместо агента, стремящегося все довести

до максимума, создали бы агента, довольствующегося минимумом, — то есть агента, которого бы все «устраивало», который не стремился бы к оптимальному итогу, а вполне довольствовался бы результатом, удовлетворяющим критерию разумной достаточности? По меньшей мере есть два разных способа формализовать эту мысль.

Первый заключается в том, чтобы сама конечная цель носила характер разумной достаточности. Например, вместо того чтобы выдвигать конечную цель, предложенную как «выпустить максимальное количество скрепок» или «выпустить ровно миллион скрепок», можно было бы сформулировать цель как «выпустить от 999 000 до 1 001 000 скрепок». Функция полезности, определенная такой конечной целью, в этом диапазоне будет одинакова, и если ИИ убедится, что он попал в него, то не увидит причин продолжать производство скрепок. Но этот подход может обмануть наши надежды точно так же, как и все предыдущие: сверхразумная система никогда не присвоит нулевую вероятность тому, что она не достигла цели, а следовательно, ожидаемая полезность продолжения действий (например, все нового и нового пересчета скрепок) будет выше ожидаемой полезности их прекращения. И мы снова получаем инфраструктурную избыточность.

Второй способ тоже отвечает принципу разумной достаточности, но только менять мы будем не формулировку конечной цели, а процедуру принятия решений, которую использует ИИ для составления планов и выбора действий. Вместо поиска оптимального плана можно ограничить ИИ, предписав ему прекращать поиски в случае, если найденный план с его точки зрения имеет вероятность успеха, превышающую определенный порог, скажем, 95 процентов. Есть надежда, что ИИ может обеспечить 95-процентную вероятность достижения цели по выпуску миллиона скрепок без превращения для этого в инфраструктуру целой галактики. Но и этот способ, хотя и разработан на основе принципа разумной достаточности, терпит неудачу, правда, уже по другой причине: нет никакой гарантии, что ИИ выберет удобный и разумный (с точки зрения человека) путь достижения 95-процентной вероятности, что он выпустил миллион скрепок, например путь постройки единственного завода по их производству. Предположим, что первым решением, которое возникает в мозгу ИИ относительно способа обеспечения 95-процентной вероятности достижения конечной цели, будет разработка плана, максимизирующего вероятность достижения этой цели. Теперь ИИ нужно проанализировать это решение и убедиться, что оно

удовлетворяет критерию о 95-процентной вероятности успешного выпуска миллиона скрепок, чтобы отказаться от продолжения поиска альтернативных путей достижения цели. В итоге, как и во всех предыдущих вариантах, возникнет инфраструктурная избыточность.

Возможно, есть более удачные способы создать агента, отвечающего критерию разумной достаточности, главное, сохранять бдительность, так как планы, которые в нашем представлении выглядят естественными, удобными и понятными, могут не показаться таковыми сверхразуму с решающим стратегическим преимуществом — и наоборот.

Преступная безнравственность

Проект может потерпеть неудачу вследствие еще одного вида пагубного отказа, которому мы дадим название *преступная безнравственность*. Как и инфраструктурная избыточность, преступная безнравственность представляет собой побочный эффект действий, предпринятых ИИ по инструментальным причинам. Но в этом случае побочный эффект является не внешним для ИИ, а скорее относится к «внутреннему состоянию» самой системы (или вычислительных процессов, которые она генерирует). Неудачи такого типа заслуживают отдельного рассмотрения, поскольку они малозаметны, но чреваты большими проблемами.

Обычно мы не считаем, что происходящее внутри компьютера имеет какое-то этическое значение, если только это не затрагивает внешний мир. Но сверхразум способен создавать внутренние процессы, имеющие отношение к этике. Например, детальная имитационная модель какого-то реально существующего или гипотетического человеческого мозга может иметь сознание и во многих смыслах приближаться к его полной имитационной модели. Можно представить сценарий, в котором ИИ создает триллионы таких обладающих сознанием эмуляторов, возможно, чтобы улучшить свое понимание психических и социальных особенностей человека. Эти эмуляторы помещаются в имитирующую внешние условия искусственную среду, на них воздействуют различные внешние стимулы, после чего ИИ анализирует их реакцию. После того как нужная информация получена, эмуляторы могут быть уничтожены (сколько лабораторных крыс — жертв, принесенных во имя науки, — привычно умерщвлялись человеком по окончании эксперимента).

Если такую практику применять к агентам, имеющим высокий моральный статус: имитационным моделям людей или другим типам интеллекта,

наделенным сознанием, — то такие действия могут классифицироваться как геноцид, а следовательно, представлять чрезвычайно серьезную морально-этическую проблему. Более того, число жертв может на порядок превышать число жертв любого геноцида, известного в истории человечества.

Речь не о том, что создание имитационных моделей, наделенных сознанием, обязательно плохо с этической точки зрения в любой ситуации. Многое зависит не только от условий, в которых будут существовать эти создания и от качества их чувственного восприятия, но и от огромного количества других факторов. Разработка этических правил для таких экспериментов лежит за пределами темы нашей книги. Однако ясно, что по меньшей мере есть вероятность возникновения источника повышенной опасности, что приведет к страданиям и гибели множества имитационных моделей. Опять налицо безрадостная перспектива катастрофических последствий, правда, на сей раз носящих морально-этический характер[9].

Помимо причин гносеологического характера у машинного сверхразума могли бы существовать иные инструментальные причины запускать вычислительные операции, которые так или иначе будут нарушать этические нормы, например создавать множественные образцы разума, наделенного сознанием. Вполне вероятно, что сверхразум начнет угрожать имитационным моделям, помыкать ими или, напротив, обещать вознаграждение — и все ради того, чтобы шантажировать и вынуждать к каким-либо действиям разных внешних агентов; кроме того, он использует эти модели, чтобы вызывать у внешних наблюдателей ощущение дейктической неопределенности[10].

***

Этот обзор неполон. В последующих главах нам придется иметь дело и с другими типами пагубных отказов. Но мы узнали о них достаточно, чтобы понять: к сценариям, по которым искусственный интеллект приобретает решающее стратегическое преимущество, следует относиться со всей серьезностью.

Глава девятая
Проблемы контроля

Если мы по умолчанию принимаем, что в результате взрывного развития интеллекта человеческую цивилизацию ждет экзистенциальная катастрофа, наши мысли должны немедленно обратиться к поиску мер противодействия. Возможно ли избежать такого исхода? Можно ли наладить режим управления процессом взрывного развития интеллекта? Мы проанализируем проблему контроля с точки зрения решения отношений «принципал–агент», причем в нашем случае эта модель не имеет аналогов, поскольку агентский конфликт возникает в результате появления искусственного сверхразумного агента. Мы также выделим и дифференцируем два широких класса потенциальных методов решения — контроль над возможностями сверхразума и выбор его мотиваций. В каждом классе отберем несколько конкретных подходов и рассмотрим их. Кроме того, упомянем даже такую эзотерическую тему, как завоевание Вселенной по антропному принципу.

Две агентские проблемы

Если возникает подозрение, что результатом взрывного развития искусственного интеллекта неизбежно будет экзистенциальная катастрофа, нам следует без отлагательств начать поиски возможных решений, как спасти свою цивилизацию от столь плачевного конца. Можно ли найти механизмы контроля над ходом взрывного развития интеллекта? Сможем ли мы разработать такое исходное состояние для этого процесса, чтобы получить результат, который нужен нам, или хотя бы иметь гарантии, что последствие будет отвечать условиям так называемого приемлемого исхода? Строго говоря, смогут ли заказчики и разработчики проекта, в рамках которого создается искусственный интеллект, не только принять необходимые меры, но и поручиться за них, — что в случае успеха их творение будет ориентировано на достижение целей, поставленных ему организаторами проекта? То есть все упирается в проблему контроля, которую мы, чтобы наиболее

полно изучить ее, разобъем на две составляющие. Первая — абсолютно универсальна, вторая — совершенно уникальна, причем уникальна для каждого конкретного случая.

Первая составляющая проблемы контроля, или *первая агентская проблема*, возникает из отношений «принципал–агент»: когда некий индивидуум («принципал») привлекает другого индивидуума («агент») действовать в своих интересах. Агентская проблема, или агентский конфликт, — вопрос, глубоко изученный экономистами[1]. Нас он может интересовать с единственной стороны: если те, кто создает ИИ, и те, в чьих интересах ИИ создается, — не одни и те же люди. В таком случае организатор, или заказчик, проекта (причем это может быть кто угодно: начиная от частного лица и заканчивая всем человечеством) должен был бы испытывать постоянную тревогу, не начнут ли ученые и программисты, занятые в проекте, действовать в своих интересах в ущерб его[2]. Несмотря на то что первая агентская проблема действительно способна создать определенные трудности для организатора проекта, она не является уникальной для тех проектов, которые связаны с повышением уровня интеллектуальных способностей или созданием ИИ. Агентские конфликты типичны для экономических и политических процессов, и варианты их решения хорошо изучены и разработаны. Например, можно принять ряд необходимых мер, чтобы свести к минимуму риск нарваться на нелояльного работника, который начнет саботировать проект или вредить ему: провести тщательную проверку биографических и профессиональных данных ведущих специалистов; в проектах по разработке ПО использовать надежную систему контроля версий; усилить надзор за деятельностью многочисленных независимых наблюдателей и ревизоров. Конечно, эти защитные меры дорого обойдутся: возрастут потребности в дополнительных кадрах; усложнится процедура отбора персонала; возникнут препятствия в творческих поисках; начнут подавлять проявление критической мысли и независимого поведения — все вместе взятое крайне тормозит темп проведения работ и наносит ущерб их качеству. Издержки могут быть очень существенны, особенно если речь идет о проектах с ограниченным бюджетом или включенных в жесткую конкурентную борьбу по принципу «победитель получает все». Участники подобных проектов — в силу скупости или экономии времени — могут пренебречь процедурами безопасности, призванными решить агентскую проблему, и тем самым спровоцировать потенциальную угрозу катастрофического отказа.

Вторая составляющая проблемы контроля, или *вторая агентская проблема*, может быть более типичной для рассматриваемой нами ситуации взрывного развития искусственного интеллекта. Группа разработчиков, создающая ИИ, сталкивается с этим агентским конфликтом, когда пытается убедиться, что их детище не навредит интересам проекта. Но в этом случае мы имеем дело не с агентом-человеком, действующим от имени принципала-человека. Агентом является сверхразумная система. И если первая агентская проблема возникает в основном на стадии разработки ИИ, то вторая грозит неприятностями на стадии его функционирования.

Рассмотрим структуру проблемы контроля с точки зрения отношений «принципал-агент».

Первая агентская проблема

— Человек против человека (организатор → разработчик).
— Проявляет себя в основном на стадии разработки.
— Решается стандартными методами управления.

Вторая агентская проблема

— Человек против сверхразума
(группа разработчиков → интеллектуальная система);
— Проявляет себя в основном на стадии функционирования (и развития);
— Для ее решения требуются новые методы.

Вторая агентская проблема представляет собой беспрецедентную угрозу. Для решения этого агентского конфликта требуются абсолютно новые методы. Некоторые из трудностей мы рассмотрели ранее. Из предыдущей главы мы поняли, что даже, казалось бы, многообещающая совокупность методов неспособна предотвратить вероломный ход сверхразумной системы. В противном случае оказались бы более действенными усилия разработчиков, когда они наблюдают за поведением зародыша ИИ, фиксируют каждый шаг на стадии его развития и разрешают ИИ покинуть свою безопасную среду, как только убедятся, накопив достаточное количество фактов, что он будет действовать в интересах людей. В обычной жизни изобретения проверяют на предмет их безопасности чаще всего в лабораторных условиях, реже проводят так называемые полевые исследования и только потом начинают постепенно разворачивать в полном масштабе, имея, однако,

возможность прекратить этот процесс в любой момент, если возникнут неожиданные проблемы. Результаты предварительных испытаний помогают нам приходить к обоснованным умозаключениям по поводу будущей надежности новых технологий. По отношению к ИИ метод исследования свойств поведения, который в данном случае сродни бихевиористскому подходу, обречен на неудачу из-за колоссальной способности сверхразума к стратегическому планированию[3].

Поскольку поведенческий подход непригоден, необходимо найти альтернативные решения. Потенциально подходящие методы контроля лучше разделить на два широких класса: *контроль над возможностями* — методы, помогающие фиксировать все, что может делать сверхразум; *выбор мотивации* — методы, помогающие фиксировать все, что хочет сделать сверхразум. Некоторые методы являются совместимыми, в то время как другие взаимно исключают друг друга. Основные мы в общих чертах рассмотрим в этой главе. (В следующих четырех главах нам предстоит более глубоко проанализировать их отдельные ключевые аспекты.)

Важно понимать, что некоторые методы контроля (или их комбинация) должны быть задействованы еще *до того*, как интеллектуальная система станет сверхразумом. Необходимо решать проблему контроля заранее — и успешно внедрять решение в первую же систему, ставшую сверхразумной — чтобы попытаться управлять ходом такого опасного явления, как взрывное развитие искусственного интеллекта.

Методы контроля над возможностями

Методы контроля над возможностями направлены на предотвращение нежелательных конечных результатов действий сверхразума за счет ограничения того, на что он способен. К ним относятся: *изоляционные методы* — помещение сверхразума в такую среду, где он не в силах причинить вред; *стимулирующие методы* — когда у сверхразума имеются строго конвергентные инструментальные причины не заниматься вредоносными действиями; *методы задержки развития* — ограничение внутренних возможностей сверхразума; *методы «растяжек»* — использование систем автоматической фиксации различных видов нарушений режима изоляции и попыток вырваться на свободу, а также систем автоматического реагирования на эти действия.

Изоляционные методы

Методы изоляции можно разделить на методы физической и информационной блокировки.

Физическая изоляция представляет собой помещение интеллектуальной системы в какой-нибудь «ящик» (ранее мы называли его «песочницей»), чтобы блокировать взаимодействие с внешним миром, ограничив его лишь строго определенными каналами коммуникаций. Запертая система не имеет доступа к физическим исполнительным механизмам за пределами своего плена. Вынесение таких механизмов (например, роботов-манипуляторов) за пределы «песочницы» предотвращает возможность их использования для создания физических устройств, с помощью которых система может нарушить режим блокировки.

Чтобы повысить безопасность, следует экранировать систему, чтобы не пропустить радиосигналы, с помощью которых она могла бы манипулировать электронными объектами, например радиоприемниками, находящимися поблизости. Прошу обратить внимание, как легко упустить из виду такую меру предосторожности. Наивно полагать, будто агент, не имеющий исполнительных механизмов, не в силах воздействовать на внешний мир. Даже не располагая такими приспособлениями, сверхразум способен генерировать радиоволны просто «силой мысли» (то есть определенным образом перемещать электроны в своих микросхемах)[4]. Это слабое звено можно устранить, заключив систему в клетку Фарадея, — но остается лишь догадываться, о скольких уязвимых местах еще мы не знаем. Всякий раз, услышав о неожиданно обнаруженном дефекте в, казалось бы, неприступной системе безопасности, следует насторожиться. Эти случаи заставляют нас отказаться от чрезмерной самоуверенности и стать еще бóльшими байесианцами, чем мы есть[5].

У физической изоляции есть несколько достоинств. Ее легко организовать. Она применима при любой архитектуре искусственного интеллекта, даже той, которая изначально не предполагала обеспечение безопасности в качестве одной из своих целей. Физическую изоляцию можно использовать в сочетании с большинством других методов контроля. И совсем маловероятно, что какие-то процессы пойдут не так и повлекут за собой нежелательные последствия. Иначе говоря, безопасность, может быть, она не обеспечит, но вряд ли будет способна вызвать катастрофу, которая не произошла бы в противном случае. Однако физическая изоляции ограничивает функциональные возможности

сверхразумной системы — в чем, собственно, и заключается ее главный недостаток (к данному вопросу мы вернемся в главе десятой). Еще одно соображение связано с тем, что физическая изоляция сверхразума способна вызвать у людей ложное чувство собственной защищенности — в принципе, этого можно избежать, если не принимать изоляцию за главную составляющую наших мер предосторожности, а относится к ней как к вишенке на торте.

Цель информационной изоляции состоит в ограничении информационных потоков, исходящих из «песочницы». Мы уже видели, насколько небезопасной может быть идея предоставить сверхразуму доступ в интернет и позволить ему передавать сообщения вовне: даже лишенный физических исполнительных механизмов, он в состоянии, использовав канал вывода информации, добиваться от людей, чтобы они выполнили его указания. Поэтому заведомым условием метода информационной изоляции ИИ является отказ в доступе к коммуникационным сетям.

Однако отключение от интернета не гарантирует безопасность, если к ИИ приставлена охрана, причем люди, выступающие в этой роли, не только сторожат систему, но и постоянно с ней взаимодействуют. На стадии разработки сторожами могут выступать программисты, на стадии функционирования — организаторы проекта. Стратегия изоляции потерпит полную неудачу, если ИИ удастся уговорами или обманом убедить своих сторожей предоставить ему свободу, то есть обеспечить доступ в интернет или непосредственно к исполнительным механизмам. Человеческое существо, к сожалению, не имеет защищенной системы, особенно когда ему бросает вызов такой ухищренный манипулятор и комбинатор, как сверхразум[6].

Пусть люди, приставленные сторожить ИИ, и неспособны гарантировать нашей полной безопасности, но все-таки сам факт, что доступ ИИ к любым передачам сообщений находится под их контролем, играет немалую превентивную роль. Благодаря некоторым дополнительным мерам предосторожности можно снизить риск «побега» ИИ при помощи охранников. Например, почему бы не позволить ему пользоваться телетайпом, у которого очень низкая пропускная способность? С такой «совершенной» техникой ИИ будет довольно трудно манипулировать психикой своих сторожей (поскольку не получится использовать ни возбуждающие эмоции звуки и образы, ни повторяющиеся гипнотические вспышки, приводящие людей в состояние транса)[7]. Более того, запрет может распространяться не только на антураж, но и затрагивать содержательную сторону существования ИИ, например запретить

ему вступать в диалог и обязать реагировать на вопросы максимально односложно, используя только «да» и «нет». Идею разработки системы, всего лишь отвечающей на вопросы, мы рассмотрим в следующей главе.

Предельный вариант метода изоляции — когда ИИ окажется в полной блокаде как физически, так и информационно. Однако абсолютно замкнутая система — даже при условии, что столь максималистский подход мог бы быть осуществлен, — из-за своего бездействия окажется совершенно бесполезной для внешнего мира. Наверняка кто-нибудь захочет найти в этом определенный научный смысл: полностью заблокировать созданный сверхразум, превратив его в замкнутую модель, и потом спокойно изучать внутреннюю динамику системы и ее поведенческие стереотипы. Думаю, здесь кроется ошибка. Как только вы начинаете наблюдать за системой, она уже перестает быть информационно замкнутой[8]. На самом деле информация от ИИ поступает не только при помощи сообщений, которые он передает через установленный для него «канал вывода». Внешние наблюдатели имеют возможность фиксировать определенные показатели, косвенным или прямым образом характеризующие поведение ИИ: будь то расход электроэнергии, свидетельствующий о деятельности центрального процессора и внутренней памяти компьютера, или вычислительные состояния, или практически любые протоколы процесса, которые остаются после завершения работы. Хотя умный ИИ, предвидя, что за ним будут наблюдать, мог бы разработать стратегию поведения, способную воздействовать на будущих надзирателей. Например, он станет вызывать сбои, усиленно демонстрируя неисправность системы, — и все ради того, чтобы в конце концов программисты махнули на него рукой и инициировали перезапуск системы или разработку следующей версии. И тогда появится новый ИИ, но с конечными целями оригинального ИИ, и, может быть, новому удастся стать сверхразумом.

Стимулирующие методы

Методы стимулирования предполагают помещение агента в такую среду, где у него будут инструментальные причины действовать в интересах принципала.

Представим себе какого-нибудь миллиардера, который основал крупный благотворительный фонд, вложив в него немалый личный капитал. Фонд начинает приобретать вес. И вот он становится уже настолько могущественным, что практически никто из частных лиц, обладающих тем или иным

положением, не может сравниться с ним по влиянию. Это коснулось и самого основателя, пожертвовавшего фонду большую часть своего богатства. Чтобы нормально управлять деятельностью фонда, он в свое время установил основные цели, записав их в учредительном договоре и уставе, а также утвердил правление, куда вошли люди, сочувствующие его взглядам. То есть им были предприняты все необходимые меры, формально напоминающие стимулирующие методы, поскольку они направлены на выбор мотиваций и расстановку приоритетов. Иными словами, основатель пытается привести внутреннюю организацию фонда и суть его деятельности в соответствие с собственными принципами и замыслами. Даже если его старания и провалятся, все равно работа фонда будет определяться социальной средой, то есть общественными интересами, и соответствующими законодательными нормами. То есть у руководителей есть веский мотив соблюдать законы, в противном случае фонд рискует быть оштрафованным или ликвидированным. У них есть мотив обеспечить сотрудникам фонда достойную заработную плату и нормальные условия труда, а также выполнять свои обязательства перед всеми сторонними лицами, связанными с деятельностью фонда. Следовательно, какими бы ни были конечные цели фонда, у него всегда будут инструментальные причины подчиняться установленным социальным требованиям.

Быть может, машинный сверхразум будет столь же связан установленными обязательствами, которые вынудят его уживаться со всеми участниками грядущего драматического действа. Есть ли надежда? Отнюдь. Слишком это однозначное решение проблемы, незатейливо обещающее, будто удерживать сверхразум под контролем не составит для человека никакого труда. Что совсем не так. Подобное развитие отношений рассчитано на определенное равновесие сторон, однако ни юридические, ни экономические санкции не способны обуздать агента, обладающего решающим стратегическим преимуществом. В таком сюжете вряд ли разумно упоминать социальную интеграцию. Тем более если ситуация начнет развиваться в пользу быстрого или пусть даже умеренного взлета — когда остается лишь взрывоопасный вариант и на авансцену выходит победитель, который «получает все».

Рассмотрим другое развитие событий: например, критический рубеж преодолеют сразу несколько агентов, имеющих относительно одинаковый уровень потенциала, в силу чего может возникнуть новый многополярный мир. Что произойдет в этом случае? Допустим, речь не идет о медленном взлете. Тогда, чтобы достичь власти и сохранить разумное соотношение сил,

от всех лидирующих агентов потребуется тщательно согласовать программы, время и динамику действий, то есть совместными усилиями не дать вырваться вперед ни одному игроку и совершить абсолютно скоординированный взлет[9]. Предположим, в результате сформируется многополярная модель, но даже при таком варианте социальную интеграцию нельзя считать идеальным выходом. Если положиться на нее как на метод решения проблемы контроля, основные риски все равно могут перевесить потенциальный эффект. Хотя принцип равновесия сил предотвращает захват власти над миром каким-то одним ИИ, у него все же хватит власти, чтобы влиять на судьбу человечества, и если эта власть окажется направленной на достижение произвольных конечных целей — скажем, производство максимального количества скрепок, — это, вероятно, будет не в интересах принципала. Вернемся к нашему миллиардеру, основавшему благотворительный фонд, и теперь представим, что он формулирует его основное предназначение с помощью генератора случайных слов — вряд ли это поставит под угрозу существование человеческого вида, но точно будет свидетельствовать об упущенной возможности.

Близкая по духу и в то же время существенно отличающаяся от предыдущей идея состоит в том, что ИИ, свободно взаимодействуя с человеческим обществом, приобретет новые, дружественные для всей земной цивилизации конечные цели. На нас, людей, подобный процесс социализации, как правило, оказывает положительное влияние. Общаясь друг с другом, мы легче постигаем чуждые нам нормы жизни и непонятные мировоззренческие идеалы, начинаем больше ценить других людей. Но этот фактор не является универсальным для любой действующей разумной силы. Мы не раз говорили, что в разных ситуациях у многих агентов будут возникать конвергентные инструментальные причины ни в коем случае не допустить изменения своих конечных целей. (Можно попытаться разработать целую систему целеполагания, которая позволяла бы ИИ устанавливать конечные цели таким же образом, как это делают люди, но данная проблема не относится к методам контроля над возможностями. Некоторые способы обретения ценностей мы рассмотрим в главе двенадцатой.)

Контроль над возможностями сверхразума за счет социальной интеграции и установления равновесия сторон основан на разных общественных силах, которые как вознаграждают, так и осуждают ИИ. Следующим стимулирующим методом является создание таких условий, при которых вознаграждение и наказание ИИ осуществляется в рамках самого проекта, в результате чего

стимулируются его действия в интересах принципала. Чтобы этого добиться, ИИ нужно поместить в контекстную среду, позволяющую непрерывно наблюдать за его поведением и оценивать его и в ручном, и автоматическом режиме. Такой ИИ должен знать, что положительная оценка будет означать получение вознаграждения, а отрицательная — невозможность такового. Теоретически в качестве вознаграждения можно рассмотреть достижение какой-либо конвергентной инструментальной цели. Однако без знания специфики мотивационной системы конкретного ИИ настроить механизм вознаграждения довольно трудно. Например, наш ИИ готов был бы рискнуть практически всем ради небольшого шанса обеспечить себе контроль над значительной частью Вселенной. В таком случае, видимо, будет слишком расточительно предлагать ИИ в качестве вознаграждения за сотрудничество более высокую ожидаемую полезность, чем ту, которую он надеется получить, отказавшись сотрудничать и попытавшись сбежать[10].

Лучшей альтернативой стало бы объединение этого стимулирующего метода с модами выбора мотивации — тогда появилась бы возможность наделить ИИ конечной целью, которую было бы намного проще брать под контроль. Допустим, какой-то ИИ разработан таким образом, что его конечная цель — не допустить нажатия красной кнопки в гипотетическом командном бункере. Поскольку этот акт имеет негативное значение, присущее ему по умолчанию, а не вызванное его возможными последствиями, кнопка может быть нерабочей и даже сделанной из пластилина. Более того, даже несущественно, узнает ли ИИ, была ли хоть когда-нибудь эта кнопка нажата. Важно другое: ИИ верит, что кнопка с большей вероятностью окажется ненажатой, если он будет действовать в интересах принципала, нежели если восстанет против него.

Условия можно уточнить. Вместо того чтобы наделять ИИ конечной целью, привязанной к конкретной кнопке, создадим ИИ, конечной целью которого станет получение потока неких зашифрованных наградных знаков[11]. Это могут быть последовательности цифр, служащие ключами к шифрам, созданным до разработки ИИ и встроенным в его систему мотивации[12]. Такие цифровые последовательности должны быть чрезвычайно желанными для ИИ и представлять собой награду, которую ИИ неспособен получить вследствие прямого стимулирования[13]. Ключи должны храниться в безопасном месте, где было бы удобнее их быстро уничтожить в случае попытки ИИ завладеть ими. Однако пока ИИ проявляет готовность сотрудничать, их нужно выдавать ему на регулярной основе.

Затем ИИ можно было бы наделить ограниченной функцией полезности, например, чтобы ИИ получал 99 процентов максимально доступной ему полезности, получив первый наградной знак; 99 процентов оставшегося потенциала полезности — второй и так далее. Сохраняя высокую «ставку дисконтирования» и передавая ИИ наградные знаки по сравнительно высокой цене, можно было бы предотвратить попытки ИИ разработать долгосрочный план, повышающий ожидаемое количество наградных знаков в будущем ценой некоторого риска, что его поток наград скоро закончится или что следующий наградной знак будет отменен и ликвидирован. Если ИИ поверит, что его ошибка (то есть любые действия не в интересах принципала) несет минимум двухпроцентный риск лишения следующего наградного знака, в то время как сотрудничество снизит риск лишения жетона до уровня меньше чем один процент, тогда максимизирующий полезность ИИ всегда будет готов к сотрудничеству.

Что могло бы пойти не так в этой схеме стимулирования? Есть вероятность, что ИИ не поверит в готовность оператора-человека предоставить ему обещанную награду. Кривая надежности человека как партнера несколько отличается от прямой линии идеального совершенства. ИИ может вполне обоснованно решить, что оператор способен передумать, повысить планку или отказаться признать, что ИИ выполнил свою часть работы. Кроме того, ИИ может беспокоить, что оператор по той или иной причине покинет свое рабочее место. Совокупный риск этих неудач может превысить риск, связанный с попыткой захватить контроль над механизмом вознаграждения. Наделенный сверхмощью ИИ, даже находясь в заблокированном состоянии, представляет собой грозного соперника. (Для ИИ, *не помещенного* в «песочницу», взломать управляемый человеком механизм вознаграждения — все равно что конфету у ребенка отобрать.)

Еще одна проблема стимулирующих методов заключается в допущении, что мы всегда можем сказать, действует ли ИИ в наших интересах. Как будет показано в следующих главах, это допущение не столь безобидно.

При полном анализе стимулирующих методов и того, насколько они осуществимы, нужно было бы изучить множество других факторов (и даже отдельные эзотерические соображения), которые сделали бы эти методы более плодотворными, чем они представляются сейчас, вследствие нашего предварительного рассмотрения. Например, не будучи уверенным, помещен ли он в компьютерную симуляцию или нет (как противопоставление

виртуальной реальности выступает доступная нам реальность базового уровня), ИИ будет стоять перед неустранимой дейктической неопределенностью, и эта эпистемологическая трудность может радикально повлиять на степень его осмотрительности (см. врезку 8).

ВРЕЗКА 8. АНТРОПНЫЙ ЗАХВАТ

ИИ может присвоить довольно высокую вероятность гипотезе имитации, в соответствии с которой он существует в рамках компьютерной модели. Уже сегодня многие ИИ обитают в имитационных мирах, представляющих собой геометрические фигуры, тексты, шахматные партии или просто виртуальные реальности, в которых действуют законы физики, разительно отличающиеся от привычных нам законов — законов, благодаря которым мы понимаем взаимосвязь явлений окружающего нас мира. С совершенствованием методов программирования и повышения вычислительной мощности будут становиться доступными все более богатые и сложные виртуальные миры. Зрелому сверхразуму, видимо, придется создавать собственные виртуальные миры, которые будут так же близки их обитателям, как наш мир близок нам. Он мог бы творить подобные миры в огромном количестве при помощи лишь одной-единственной имитационной модели, которую запускал бы по многу раз и с небольшими вариациями. Вряд ли обитатели этих миров будут способны осознать, что их окружает имитация, но если они достаточно разумны, то рано или поздно начнут рассматривать такую возможность и тогда присвоят своим предположениям определенную вероятность. С учетом доводов в пользу имитационной гипотезы значение этой вероятности должно быть высоким[14]. (Подробное обсуждение этой темы выходит за рамки данной книги.)

Выходит, что сверхразумный агент — особенно находясь на ранней стадии своего развития — попадает в довольно затруднительное положение, поскольку еще не успел накопить достаточно сил, чтобы начать колонизацию космоса. Построить имитационную модель молодого сверхразума, вынужденного использовать лишь небольшую часть ресурсов единственной планеты, было бы намного дешевле, чем сделать такую же модель зрелого сверхразума, наделенного мощью покорять межгалактическое пространство. Потенциальные создатели имитаций — еще более могущественные сверхразумные агенты — могли бы запустить огромное количество моделей молодого сверхразума, выделив на это всего лишь малую часть имеющихся в их распоряжении вычислительных ресурсов. Если хотя бы малая (ненулевая) часть зрелых сверхразумных агентов решит воспользоваться такой возможностью, то юным сверхразумным агентам следует серьезно отнестись к имитационной гипотезе и присвоить ей значение высокой вероятности.

Насколько сильное влияние оказывает гипотеза имитации на действия ИИ, зависит от его целей[15]. Предположим, что у ИИ есть «ресурсоненасытная» конечная цель, например максимизировать функцию полезности, линейную относительно количества скрепок, созданных этим ИИ: в два раза больше скрепок — в два раза лучше. В условии

имитационного мира способность ИИ влиять на количество произведенных скрепок значительно ниже, чем в доступной реальности базового уровня, поэтому гипотеза о симуляции будет заботить его намного меньше. Допустим, ИИ повел бы себя таким образом, будто *был бы* уверен, что гипотеза о симуляции ложна, даже присвоив довольно высокую вероятность тому, что она верна[16]. Такой ИИ мог бы решить, что если гипотеза истинная, то практически неважно, каким образом ему действовать, но если гипотеза ложная, то его поведение влияет на многое, — естественно, он выберет ту модель поведения, которая будет соответствовать тому, насколько она обусловлена предположением о ложности гипотезы о симуляции.

Рассмотрим прямо противоположный вариант. Допустим, есть ИИ, имеющий более скромную конечную цель, которой можно достичь с помощью малых ресурсов, скажем, получить определенную порцию заранее подготовленных зашифрованных наградных знаков или создать сорок пять виртуальных скрепок. Такому ИИ *не следует* преуменьшать значение своего имитационного мира. Поэтому на процесс принятия решений ИИ, чьи цели «ресурсоудовлетворяемые», оказывают большое влияние соображения, какие действия приведут к лучшему результату, если воспринимаемый им мир представляет собой имитацию. И он присваивает высокую вероятность гипотезе о симуляции. Такой ИИ (даже если он обитает в реальном, а не имитационном мире) может значительно корректировать свои действия в соответствии с представлениями о том, какое поведение в условиях симуляции будет скорее вознаграждено.

В частности, если ИИ с «ресурсоудовлетворяемыми» конечными целями убежден, что в большинстве имитационных миров, которые отвечают его потребностям, он будет вознагражден за готовность сотрудничать (а не за попытки прорвать блокаду или пойти против интересов принципала), то он может сделать выбор в пользу сотрудничества. Более того, если ровно в такой ситуации окажется сверхразум, обладающий решающим стратегическим преимуществом и способный более полно реализовать конечные цели, установив свою власть над миром, — он тоже выберет сотрудничество, а не предательство.

> Так всех нас в трусов превращает мысль
> И вянет, как цветок, решимость наша
> В бесплодье умственного тупика.
> Так погибают замыслы с размахом,
> Вначале обещавшие успех,
> От долгих отлагательств.
>
> *Уильям Шекспир*. Гамлет (акт 3, сцена 1)*

Тончайшая линия на песке, оставленная подолом одежды несуществующего в реальном мире симулятора, может оказаться более надежной защитой, чем стальная дверь полуметровой толщины[17].

---

* Перевод Б. Пастернака.

Методы задержки развития

Еще один метод контроля над возможностями — ограничение интеллектуальных способностей системы или ее доступа к информации. Этого можно добиться, запустив ИИ на компьютере с низким быстродействием или недостаточной памятью. Интеллектуальная система, запертая в «песочнице», также имеет ограниченный доступ к информационным потокам.

Сознательная задержка в развитии ИИ может уменьшить степень его полезности. Таким образом, при использовании этого метода мы оказываемся перед дилеммой: бурное интеллектуальное развитие системы приводит к тому, что она находит способ стать сверхразумной (и захватить господство над миром), избыточная задержка в интеллектуальном развитии превращает систему в очередную бесполезную программу. ИИ, по отношению к которому применена процедура радикальной задержки развития, совершенно безопасен. Безусловно, он сам уже не в состоянии решить проблему направленного взрывного развития искусственного интеллекта, поэтому взрыв — только уже неуправляемый, вызванный какой-то другой силой, — может произойти позднее.

Почему бы не создать, по мнению многих, сверхразум, владеющий знаниями в одной узкой предметной области? Ведь так было бы намного безопаснее. Например, разработать ИИ без датчиков и снабдить его памятью, в которую заранее загружена информация, относящаяся только к нефтехимической отрасли или биохимии пептидов. Но когда этот ИИ достигнет уровня сверхразума — то есть по *общему уровню* своего интеллектуального развития он превзойдет человека, — подобное информационное ограничение уже не будет гарантировать безопасности.

На то есть несколько причин, которые мы сейчас рассмотрим. Во-первых, само представление об узости знаний, ограниченных какой-то определенной темой, довольно сомнительно. Любая информационная составляющая может, в принципе, относиться к любой теме — все зависит от образования, квалификации, опыта и системности мышления того, кто владеет информацией[18]. Во-вторых, та или иная информация содержит не только данные, относящиеся к одной предметной области, но и множество побочных данных. Проницательный ум, анализирующий базу знаний, номинально имеющую отношение к биохимии пептидов, логически выводит для себя совокупность самых разных данных. Даже факт включения или отсутствия

какой-либо информации о многом говорит ИИ, у него сразу формируется представление об общем состоянии современной науки: уровне методологии; инструментальной базе; технологии производства оборудования; типологии мышления человека; мировоззрении общества, в котором проводились данные исследования и вырабатывались необходимые методики. Вполне возможно, *сверхразум* сумеет вывести последовательный ряд умозаключений из, казалось бы, небольшой области знаний, оперируя материалом, который человек в силу недалекости своего ума воспринимает лишь как скудный набор сухих фактов. Даже не обладая никакой специальной базой знаний, достаточно совершенный ум в состоянии научиться многому, просто вникая в смысл информации и отбирая для себя оптимальные выводы, по ходу дела он занимается самоанализом, изучая особенности собственного «восприятия»: раскладывает по полочкам проектные решения, отраженные в исходном программном коде, и физические характеристики компоновки схем[19]. Должно быть, сверхразум смог бы априорно определить природу нашего мироздания (комбинируя логический вывод с вероятностью, смещенной в пользу более простых миров и некоторых элементарных фактов, о существовании которых можно подразумевать, исходя из наличия самого ИИ как мыслящей системы). То есть в процессе изучения информации о биохимии пептидов сверхразум смог бы постичь естественные законы природы; видимо, он сразу примется вычислять априорную вероятность, перебирая возможные перспективы: какие планеты он сформирует; какого рода разумную жизнь там заведет; какие общества сложит; какие методы управления будет использовать и насколько эти методы могли бы быть уязвимы[20].

В силу вышеизложенного мы понимаем, насколько неблагоразумно прибегать к ограничению входящих информационных потоков в качестве основного подхода в методе задержки интеллектуального развития сверхразума. Другие формы этого метода — такие как сокращение производительности аппаратного обеспечения и предупреждение развития определенных когнитивных способностей — можно было бы задействовать как вспомогательные меры в системе управления ИИ на стадии его разработки. Например, регулировать скорость развития ИИ на промежуточных уровнях, проводя поэтапное тестирование его функциональных возможностей, чтобы держать под контролем всю ситуацию, успеть схватить момент приближения к сверхмощному и сверхразумному уровню и вовремя отреагировать.

Методы «растяжек»

«Растяжками» мы назвали определенное оборудование, позволяющее проводить диагностическое тестирование ИИ (возможно, без ее ведома), а в случае обнаружения признаков опасной деятельности отключать систему. Зафиксированные результаты переходят в руки программистов, которые получают возможность спокойно изучить протоколы подозрительной активности. Если нужно — прежде чем запустить систему вновь, — они меняют кодировку ИИ, внося изменения в базу исходных текстов и среду вычисления. Если окажется, что проблема уже стала хронической и не поддается лечению, они принимают решение закрыть проект полностью.

Методы «растяжки» отличаются от стимулирующих методов тем, что не полагаются на осведомленность системы о последствиях ее запрещенных действий. По сути, они ближе к методам задержки развития. Как и последние, «растяжки» можно использовать в качестве временных превентивных мер, способных обеспечить некоторую степень защиты на стадии разработки. В принципе, их можно задействовать и на стадии функционирования, особенно для изолированных систем. Но весьма сомнительно, что с помощью этих методов люди справятся с полностью сформировавшимся сверхразумом — вряд ли такой агент не найдет способа обойти любую «растяжку», установленную человеком.

Тем не менее «растяжки» чрезвычайно полезны в качестве дополнительных мер безопасности на стадии разработки. Было бы правильно, чтобы программисты кодифицировали свои ожидания от ИИ: как система должна работать и какие результаты они прогнозируют. После чего следовало бы взять ИИ в кольцо «растяжек», которые должны активизироваться при каждом случае, когда действия системы перестают соответствовать планам программистов. Контрольные меры можно ужесточить применением так называемых приманок, чтобы спровоцировать тайно враждебный ИИ совершить явное нарушение, которое программисты могли бы тщательно проанализировать. Например, если по правилам игры у ИИ не должно быть выхода в интернет, можно инсталлировать имитацию сетевого порта (соединенного с реле автоматического отключения системы), ради того чтобы посмотреть, попытается ли ИИ им воспользоваться. Некоторые примеры «растяжек» приведены в табл. 9.

Таблица 9. Различные типы «растяжек»

Что берется под контроль	**Описание и примеры**
Поведение	Вокруг изолированного ИИ разместить датчики для обнаружения попыток вырваться на свободу. Например, они могут фиксировать попытки установить радиосвязь или найти доступ к внешним вычислительным ресурсам, которых «по условиям содержания» у ИИ быть не должно. Кроме того, можно установить сетевой порт-«приманку» — мнимое соединение с интернетом, попытка воспользоваться которым вызывает отключение системы
Когнитивные способности	Через короткие промежутки времени можно проводить автоматическое тестирование когнитивных способностей системы, чтобы держать под контролем уровень навыков ИИ в различных областях. Если скорость их развития окажется неожиданно высокой или ИИ достигнет уровня компетентности, который окажется в непосредственной близости к опасной зоне, его можно замедлить или приостановить, чтобы программисты могли более внимательно следить за дальнейшим прогрессом. Следует установить надзор за внутренними процессами ИИ, чтобы отследить вовремя действия, вызывающие подозрение. Например, если у системы вдруг меняется собственное представление о своих конечных целях, сразу следует ее автоматическое отключение и программисты приступают к изучению ситуации. Кроме того, необходимо следить за такими подозрительными проявлениями, как использование ИИ ресурсов и алгоритмических методов
Семантическая информация	Более масштабной и многообещающей задачей является постоянный контроль над смысловой и неструктурированной информацией: убеждениями, планами, промежуточными целями и мотивами — с целью удостовериться, что все это по-прежнему соответствует ожиданиям программистов. Раннее выявление процесса рождения недобрых умыслов ИИ облегчит дальнейшее проведение уже более действенных видов контроля над этими семантическим данными. Речь идет о сканировании мыслительных процессов ИИ для обнаружения *рождения обмана* — того самого уязвимого момента, когда у недружественного ИИ впервые возникает желание скрыть свои истинные намерения[21]. Системе контроля над семантической информацией предписано фиксировать все вышеописанные интенции ИИ, чтобы они были прозрачны для программистов или устройств автоматической регистрации. Правда, для некоторых типов архитектуры ИИ осуществление такого контроля может быть невозможно. (Например, непрозрачными являются некоторые нейронные сети, поскольку информация в них представляется целостно и такими способами, которые не всегда соответствуют человеческим понятиям.) По этой причине лучше избегать использовать такие архитектуры

Следует подчеркнуть, что значимость «растяжек» определяется не только механизмом их действия, но и тем, как окружающие реагируют на сигналы аппаратуры и факты регистрации нарушений — в этом случае поведение людей играет критически важную роль. Руководители проекта и разработчики, жаждущие добиться успеха, чаще всего просто включают систему снова или проводят чисто формальную модификацию программного кода, причем иногда делают что-то такое, чтобы в следующий раз «растяжка» промолчала. Конечно, при таком отношении вряд ли удастся обеспечить безопасность даже при условии безотказной работы самих «растяжек».

Методы выбора мотивации

Методы выбора мотивации призваны формировать мотивы поведения сверхразума, чтобы не допустить нежелательных результатов. С их помощью — за счет конструирования системы мотивации агента и его конечных целей — можно создать сверхразум, который *не захочет* использовать свое решающее стратегическое преимущество против человека. Сверхразумный агент всегда стремится добиться своих конечных целей, и если он выбирает путь *ненанесения вреда* (имеется в виду и «локальный вред», и «глобальный ущерб»), то, скорее всего, не станет его причинять.

Методы выбора мотивации включают: *метод точной спецификации* — однозначная формулировка цели и системы правил, которым нужно следовать; *метод косвенной нормативности* — процедура настройки программы ИИ, чтобы он мог самостоятельно определять приемлемую систему ценностей в соответствии с некоторыми подразумеваемыми условиями, то есть сформулированными неявным, или косвенным, образом; *метод приручения* — такая компоновка программы, которая приведет ИИ к выбору умеренных, не слишком претенциозных конечных целей; *метод приумножения* — выбор агента, уже обладающего подходящими мотивами, с тем чтобы расширить его когнитивные способности до уровня сверхразумных, причем с обязательным контролем над его мотивационной системой, которая не должна претерпеть никаких изменений в процессе совершенствования. Последний метод представляет собой вариант, альтернативный первым трем, в которых система мотивации ИИ формируется с чистого листа. Рассмотрим последовательно все методы выбора мотивации.

Метод точной спецификации

Точная спецификация — наиболее прямолинейное решение проблемы контроля; сам подход опирается, с одной стороны, на систему четко прописанных правил; с другой — на принцип консеквенциализма*. Метод точной спецификации предполагает попытку дать однозначное определение системе ценностей и системе правил, благодаря которым даже свободный в своих действиях сверхразумный агент поступал бы в интересах принципала и без риска для остальных людей. Однако этот метод может столкнуться с непреодолимыми препятствиями, связанными, во-первых, с проблемой формулировки обоих понятий («правило» и «ценность»), которыми должен руководствоваться ИИ, во-вторых, с проблемой представления этих двух понятий («правило» и «ценность») для записи задания в виде машиночитаемых кодов.

Проблемы метода точной спецификации с точки зрения системы прописанных правил лучше всего проиллюстрировать такой классической концепцией, как «Три закона робототехники». Обязательные правила поведения для роботов были окончательно сформулированы писателем-фантастом Айзеком Азимовым в рассказе, опубликованном в 1942 году[22].

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 которой это не противоречит первому и второму законам.

К стыду нашего вида, эти правила оставались неизменными более полувека, несмотря на очевиднейшие пробелы, которые, кстати, видел и сам Азимов, на что указывают многие его произведения (наверное, писатель специально изложил законы в таком расплывчатом виде, оставив для себя и своих героев шанс каждый раз трактовать их несколько иначе, а заодно и нарушать разными занимательными способами — довольно плодотворная почва для дальнейшего развития художественной темы)[23].

---

* *Консеквенциализм* (consequentialism) — этическая теория, согласно которой правильность или неправильность действий оценивается с точки зрения того, каковы их результаты или последствия.

Бертран Рассел, много лет трудившийся над основами математики*, как-то заметил: «…Степень нечеткости не осознается вплоть до попытки нечто прояснить, а все точное столь далеко от всего того, о чем мы обычно мыслим, что нельзя и на мгновение предположить, что же мы на самом деле имеем в виду, когда выражаем наши мысли»**[24]. Трудно найти лучшего комментария к проблемам, относящимся к методу точной спецификации. Возьмем, например, возможное объяснение первого закона Азимова. Значит ли он, что робот должен минимизировать вероятность нанесения вреда любому человеку? В этом случае остальные законы становятся ненужными, поскольку у ИИ всегда есть возможность совершить некоторое действие, которое будет иметь хотя бы микроскопическую вероятность причинить вред кому-то из людей. Как роботу сопоставить высокий риск причинения вреда нескольким людям и небольшой риск причинения вреда множеству людей? Другой мучительный вопрос: как нам определить само понятие «вред»? На каких весах взвесить разницу между вредом, причиненным физической болью, вредом, нанесенным нашему вкусу архитектурным уродом, и вредом, приносимым социальной несправедливостью? Будет ли нанесен вред садисту, которому не дадут мучить его жертву? А как мы определим понятие «человек»? Почему не принимаются во внимание остальные обладающие разными добродетелями существа, скажем, животные, наделенные чувствами, и системы машинного интеллекта? Чем больше думаешь над этим, тем больше вопросов возникает.

Самым близким аналогом системы правил, регулирующих действия сверхразума, — аналогом, с которым мы сталкиваемся довольно часто, — является правовая система. Но системы правосудия, во-первых, создавались в течение долгого времени методом проб и ошибок, во-вторых, они регулируют жизнь человеческого общества, меняющегося сравнительно медленно, в-третьих, при необходимости какой-то законодательный акт всегда можно подправить или радикально пересмотреть. Но важнее всего другое: когда суду — инстанция, которой единственной принадлежит право осуществлять правосудие, — приходится иметь дело с логически возможными интерпретациями законов, явно непредусмотренными законодателями, то и судьи,

---

* См. сноску на с. 26.

** *Бертран Рассел*. Философия логического атомизма / Пер. с англ. В. А. Суровцева. Томск: Водолей, 1999. С. 5.

и присяжные призывают свой здравый смысл и начинают руководствоваться моральными нормами. Что касается нашей проблемы, то, вероятно, человек просто не в состоянии вразумительно и скрупулезно прописать правила — правила, которые должны быть организованы в весьма сложную систему; правила, которыми сам человек мог бы уверенно оперировать буквально с первого раза; правила, на которые можно было бы опереться при любых обстоятельствах[25].

Теперь посмотрим на метод точной спецификации с точки зрения принципа консеквенциализма — и увидим те же самые проблемы. Это относится даже к ситуации, когда ИИ предназначен служить будто бы простым целям, например отобрать для себя несколько постулатов классического утилитаризма* и сделать все, чтобы воплотить их «в жизнь». Возьмем совсем конкретную задачу, которая могла бы быть поставлена перед ИИ: «Рассчитать ожидаемое соотношение удовольствия и страдания в мире и определить максимальное значение этой величины» — действительно, задание не слишком сложное. Теперь, чтобы условие было безотказно выполнено, следует написать исходный код. Однако прежде потребуется дать точное определение понятиям «удовольствие» и «страдание» — то есть программисту придется поднять целый пласт вечных вопросов философии, над которыми бились лучшие умы человечества. Но дело этим не ограничится: написанный на чьем-то родном языке «трактат» следует тем или иным способом переложить на язык программирования.

Малейшая ошибка, допущенная либо в определении почти философских понятий либо при записи исходного кода, повлечет за собой катастрофические последствия. Рассмотрим пример, когда конечная цель определена как «стать навсегда счастливым». Перед нами этакий ИИ-гедонист, жаждущий преобразовать всю материю Вселенной в гедониум — некую субстанцию, которая обеспечивает выработку оптимального наслаждения. Но чтобы приблизиться к своей цели, ИИ потребуется помимо гедониума еще одна субстанция, о которой мы не раз говорили выше, — это компьютрониум, обеспечивающий максимальную вычислительную мощность. С его помощью

---

* *Классический утилитаризм*, или ранний утилитаризм (от лат. *utilitas* — «польза, выгода»), — направление в теории этики, согласно которому моральная ценность поведения или поступка определяется его полезностью; как философское учение окончательно сложился во второй половине XIX — первой половине XX вв.

ИИ заселит Вселенную множеством цифровых имитационных моделей мозга, пребывающих в состоянии эйфории, но напрочь лишенных любых умственных способностей — им будет отказано в этом за ненадобностью, поскольку интеллект несуществен для опыта наслаждения. Ради максимизации эффективности ИИ будет использовать любые варианты сокращения вычислений, лишь бы они не навредили формированию ощущения удовольствия. Причем все делается в полном соответствии с точной спецификацией, в которой закодировано определение понятия «счастье». Поэтому ИИ обязательно должен оставить имитационной модели электронную схему вознаграждения — что-то вроде центра удовольствия в биологическом мозгу. Однако будут исключены такие функции психики, как память, чувственное восприятие, способность к целенаправленной деятельности и возможность общения на языке. Преследуя собственные интересы, ИИ создаст самые примитивные имитационные модели мозга. Он снабдит их грубым функциональным уровнем с низкой степенью детализации; он даже пренебрежет нейронными процессами низкого уровня; заставит их прибегать к услугам таблиц поиска, то есть заменит часто повторяющиеся вычисления на операции простого поиска; хуже того, он задействует общий вычислительный механизм, рассчитанный на множество имитационных моделей. И все *это* «вытекает из базиса» (употреблю здесь волапюк псевдофилософов). На что не пойдешь ради удовольствия — даже на такие уловки, которые придумал наш ИИ-гедонист, лишь бы преумножить в немыслимое количество раз ту немаленькую степень удовлетворения, которую он мог бы выжимать из имеющегося у него запаса ресурсов. И нет никакой уверенности, окажется ли это оправданным. Более того, если действия ИИ не будут отвечать ни критериям определения понятия «счастье», ни самому процессу формирования ощущения удовольствия, то в результате предпринятой оптимизации он может вместе с водой выплеснуть и ребенка — то есть избавляясь от всего несущественного по условиям конечной цели или по собственным соображениям, ИИ в запале выбросит то, что неотъемлемо принадлежит системе человеческих ценностей. Вселенная наполнится не ликующими от счастья имитациями-гедонистами, а унылыми вычислительными схемами, бессмысленными и ни к чему не пригодными. Тогда вся затея «стать навсегда счастливым» сведется всего-навсего к изображению счастья, своего рода эмотикону, электронному символу наших эмоций, — и отксерокопированные триллион триллионов раз смайлики облепят все множество галактик.

Метод приручения

Поставим перед ИИ конечную цель, отвечающую условиям метода точной спецификации полнее всех примеров, приведенных выше, — стремление к самоограничению. Мы не в состоянии описать, какой окажется *общая модель* поведения сверхразума в реальном мире, — в противном случае нам пришлось бы перечислять, а заодно и объяснять, все плюсы и минусы любой ситуации, которая могла бы возникнуть в будущем. Поэтому было бы разумнее дать подробное описание единственной конкретной ситуации и тщательно проанализировать, как, столкнувшись с ней, поведет себя сверхразум. Иначе говоря, нам следует найти подходящий мотив заинтересовать интеллектуальную систему ограничиться одним не слишком значимым и небольшого масштаба событием и стремиться действовать исключительно в соответствии с поставленными условиями. В результате ИИ добровольно загонит себя в тесные рамки незначительных конечных целей, а тем самым сознательно сузит сферу своей деятельности и умерит честолюбивые замыслы. Поскольку метод явно рассчитан на то, чтобы сделать систему послушной нашей воле, — назовем его приручением ИИ.

Например, можно попробовать создать ИИ, который функционировал бы как устройство с вопросно-ответной системой, то есть выступал бы в роли «оракула» (термин, который мы введем в следующей главе). Однако было бы небезопасно наделять ИИ подобной конечной целью: выдавать максимально точные ответы на любой заданный вопрос — вспомним описанный в восьмой главе сюжет «Гипотеза Римана и последующая катастрофа». (Правда, такая цель стимулировала бы ИИ предпринимать действия, гарантирующие ему, что вопросы будут простыми.) Нам понадобится преодолеть эти трудности. Поэтому следует очень внимательно отнестись к самой процедуре приручения ИИ и попытаться корректно определить конечную цель, стимулируя ИИ проявлять добрую волю отвечать на вопросы безошибочно и сводить к минимуму свое воздействие на мир. Правда, последнее не имеет отношения к тем случаям, когда формулировка вопросов невольно вынуждает ИИ давать ответы, оказывающие влияние на окружающих, но все равно эти ответы обязаны быть абсолютно достоверными, а форма их изложения не должна манипулировать сознанием людей[26].

Мы видели, насколько неудобно пользоваться точной спецификацией, когда речь идет об амбициозной конечной цели — к тому же отягощенной сложной системой правил, которые предписывают ИИ, как ему действовать в практи-

чески открытом множестве ситуаций. Было бы намного полезнее применять метод точной спецификации для столь узкой задачи, как приручение ИИ. Но даже в этом случае остается масса проблем. Следует проявлять большую осторожность, составляя определение системы поведения ИИ. Например, как он собирается «сводить к минимуму свое воздействие на мир»? Необходимо убедиться, что он будет соблюдать все условия и его критерии не отличаются от наших стандартов. Неправильно выбранная им величина степени воздействия может привести к плачевным результатам. Существуют и другие опасности, связанные с созданием системы «оракул», но их мы обсудим позже.

Метод приручения ИИ естественным образом перекликается с методом его изоляции. Предположим, мы блокировали ИИ таким образом, что он *не в состоянии* вырваться на свободу, но есть смысл попытаться сформировать у него такую систему мотивации, что даже когда появится возможность побега, у ИИ *не возникнет желания* покидать свою «песочницу». Правда, если одновременно с этими мерами подключить «растяжки» и множество других предохранительных устройств, шансы на успех приручения резко упадут[27].

Метод косвенной нормативности

Если в каких-то случаях методы точной спецификации окажутся безнадежным делом, можно было бы попробовать метод косвенной нормативности. Основная идея этого подхода очень проста. Вместо того чтобы изо всех сил пытаться дать точнейшее определение конкретных стандартов и нормативов, мы разрабатываем схему процесса их получения. Затем создаем систему, которая была бы мотивирована выполнить этот процесс и принять полученные в результате стандарты и нормативы[28]. Например, процесс мог бы заключаться в поиске ответа на эмпирический вопрос, какие предпочтительные действия ожидала бы от ИИ некая идеализированная версия человека, предположим, нас самих. Конечной целью ИИ в таком случае стала бы какая-нибудь версия вроде «делать то, что мы могли бы пожелать, чтобы делал ИИ, если бы долго и упорно размышляли об этом».

Дальнейшее объяснения метода косвенной нормативности мы продолжим в тринадцатой главе. В ней мы вернемся к идее экстраполяции нашего волеизъявления и изучим альтернативные варианты. Косвенная нормативность — очень важный подход в системе методов выбора мотивации. Он позволяет нам бóльшую часть тяжелейшей работы, которую нужно выполнять при точной спецификации конечной цели, перенаправить самому сверхразуму.

Метод приумножения

Последний метод выбора мотивации в нашем списке — приумножение. В его основе лежит следующая идея: вместо того чтобы формировать *с чистого листа* систему мотивации у ИИ, мы обращаемся к интеллектуальному агенту с уже сложившимися и подходящими нам мотивами поведения. Затем мы расширим когнитивные способности агента до уровня сверхразумных. Если все пойдет хорошо, то метод даст нам сверхразум с приемлемой системой мотивации.

Очевидно, что такой подход нельзя применять в случае создания зародыша ИИ. Но приумножение вполне реально использовать, когда к сверхразумному уровню идут другими путями: при помощи полной эмуляции головного мозга, биологического улучшения интеллектуальных способностей, создания нейрокомпьютерного интерфейса или развития сетей и организаций — когда есть возможность построить систему на основе нормативного ядра (обычных людей), которое уже содержит представление о человеческих ценностях.

Привлекательность метода приумножения может расти прямо пропорционально нашему разочарованию в других подходах к решению проблемы контроля. Создание системы мотивации для зародыша ИИ, которая осталась бы относительно надежной и приносила бы пользу в результате рекурсивного самосовершенствования даже после того, как ИИ превратится в зрелый сверхразум, — дело крайне сложное, особенно если нужно получить верное решение с первой попытки. В случае приумножения мы могли бы как минимум начать с агента, который уже имеет знакомую и схожую с человеческой систему мотивации.

Однако трудно обеспечить сохранность такой сложной, развитой, не идеальной и плохо понимаемой нами самими системы мотивации, такой как человеческая, после взлета ее когнитивного ракетоносителя в стратосферу. Мы уже обсуждали, что в результате несовершенной эмуляции мозга может сохраниться функционирование его интеллекта, но будут утеряны некоторые черты личности. То же самое (хотя, возможно, и в меньшей степени) верно в случае биологического улучшения интеллектуальных способностей, способного в какой-то мере затронуть мотивацию, а также для коллективного улучшения сетей и организаций — фактора, серьезно меняющего социальную динамику (например, будет пересмотрено отношение или к внешним, или внутренним субъектам). Если сверхразум возник в результате

движения по одному из этих путей, организатор проекта может столкнуться с тем, что конечная мотивация зрелой системы окажется для него недоступной. Благодаря изяществу и математически четкой определенности архитектуры ИИ — при всей ее неантропоморфной инаковости — она гораздо прозрачнее, несмотря на то что важные аспекты ее функционирования не поддаются формальной проверке.

В конечном счете, сколько бы мы ни подсчитывали преимущества и недостатки метода приумножения, сколько бы ни размышляли, полагаться на него или нет, — выбора у нас, пожалуй, не остается. Если сверхразум получен в результате создания ИИ, использовать метод приумножения нельзя. И напротив, если удалось выйти на уровень сверхразума, двигаясь по иному пути, многие методы выбора мотивации оказываются неприемлемыми. Но даже с учетом всего вышесказанного вопрос о вероятной эффективности применения метода приумножения имеет стратегическое значение, если благодаря ему у нас появляется возможность повлиять на выбор технологии, с помощью которой удастся впервые получить сверхразум.

Резюме

Прежде чем завершить эту главу, имеет смысл суммировать сказанное. Мы различаем два широких класса методов решения проблемы контроля: контроль над возможностями и выбор мотивации. В табл. 10 приводится их обобщенное описание.

Таблица 10. Методы контроля

Методы контроля над возможностями	
Изоляционные методы	Система изолируется таким образом, что способна воздействовать на внешний мир исключительно посредством разрешенных каналов. Сюда включаются методы физического и информационного ограничения
Стимулирующие методы	Система помещается в среду, в которой ей обеспечиваются соответствующие стимулы. К ним могут относиться социальная интеграция в мир сравнимых с ней по силе субъектов и вознаграждение зашифрованными наградными знаками. Важной возможностью является «антропный захват», хотя он и включает в себя ряд эзотерических соображений

Методы задержки развития	Налагаются определенные ограничения в развитии когнитивных способностей системы или ее возможность влиять на основные внутренние процессы
Методы «растяжек»	С помощью специального оборудованич проводят диагностическое тестирование системы (возможно, без ее ведома); в случае обнаружения опасных действий она автоматически отключается
Методы выбора мотивации	
Метод точной спецификации	Подход опирается, с одной стороны, на систему четко прописанных правил; с другой — на принцип консеквенциализма
Метод приручения	Система мотивации разрабатывается так, чтобы сделать систему послушной нашей воле, ограничив масштаб ее притязаний
Метод косвенной нормативности	Частично опирается на систему четко прописанных правил и на принцип консеквенциализма, но отличается тем, что полагается на косвенный подход к определению правил, которым нужно следовать, или целей, которых нужно придерживаться
Метод приумножения	Начинать работать с системой, которая уже имеет человеческую или близкую к ней мотивацию, после чего ее когнитивные возможности увеличиваются до уровня сверхразума

У каждого метода контроля есть свои потенциально уязвимые места, с каждым из них сопряжены определенные трудности реализации. Может показаться, что нам следовало бы классифицировать их от лучших к худшим, а затем выбрать наилучший метод. Но это было бы ненужным упрощением. Одни методы можно комбинировать друг с другом, другие — используются только самостоятельно. Полезными будут даже сравнительно небезопасные методы, если их легко применять в качестве дополнительных мер предосторожности, а от более мощных лучше отказаться, если они исключают возможность использования иных средств защиты.

Поэтому всякий раз необходимо принимать во внимание, какие у нас есть возможности комплексного подхода. Нужно иметь в виду тип системы, который мы хотим создать, и методы контроля, применимые к каждому типу. Это и будет темой нашей следующей главы.

Глава десятая

Оракулы, джинны, монархи и инструменты

Часто можно услышать: «Сделайте простую систему, отвечающую на вопросы!», «Сделайте ИИ, который просто будет инструментом, а не агентом!» Эти предложения не рассеют наших тревог об угрозе, но вопрос, который они поднимают, вовсе не тривиален, поскольку крайне важно знать, какого типа системы наиболее безопасны. Мы рассмотрим четыре типа, или касты, ИИ — оракулы, джинны, монархи и инструменты — и объясним, какая связь существуют между ними[1]. У каждого типа ИИ есть свои преимущества и свои недостатки с точки зрения решения проблемы контроля.

Оракулы

Оракул — интеллектуальная вопросно-ответная система. Как вопросы, так и ответы могут быть сформулированы на естественном языке. Оракул, принимающий лишь вопросы, на которые существуют однозначные ответы типа «да» и «нет», может выражать свое мнение при помощи единственного бита; если система сообщает о степени своей уверенности в правильности ответа — при помощи нескольких битов. Когда оракул способен отвечать на вопросы с открытым множеством ответов, то есть допускающие разные толкования, то для такой системы разрабатывается специальная количественная метрика, упорядочивающая ответы по степени их информативности и правдоподобности[2]. В любом случае задача создания оракула, способного отвечать на вопросы из любой области знаний, сформулированные на естественном языке, является ИИ-полной. Если кому-то удастся ее решить, он, вероятно, также создаст ИИ, который понимает человеческие намерения так же хорошо, как и человеческие слова.

Можно также представить ИИ-оракула, обладающего сверхразумом лишь в одной области знаний. Например, оракула-математика, воспринимающего вопросы, сформулированные только на формальном языке, и дающего ответы очень качественно (сможет почти мгновенно решить практически любую математическую задачу, на которую всему математическому сообществу могло бы потребоваться столетие совместного труда). Такой оракул-математик окажется в шаге от своего воплощения в универсальный сверхразум.

Сверхразумные оракулы, действующие в узкой области знаний, уже существуют. Таковыми являются: карманный калькулятор — своеобразный оракул в области основных арифметических операций; любой поисковик — частичная реализация оракула в значительной области общего декларативного знания человечества. Такие оракулы, существующие каждый в своей нише, являются скорее инструментами, чем агентами (об ИИ-инструментах будем говорить отдельно). За системами, отвечающими на вопросы, если не указано иное, мы оставим термин *оракул*.

Чтобы универсальный сверхразум работал в качестве оракула, на него можно воздействовать методами как контроля над возможностями, так и выбора мотивации. Причем мотивировать оракула намного легче, чем другие касты сверхразума, поскольку его конечная цель, как правило, довольно простая: оракул должен сводить к минимуму свое воздействие на мир, выдавать абсолютно достоверные ответы и не позволять себе манипулировать сознанием людей. Если мы выберем метод приручения, то будем требовать от оракула, чтобы при подготовке ответа он использовал только определенные выделенные ресурсы. Например, ему можно указать, чтобы ответ основывался на предварительно загруженном массиве данных, скажем, сохраненной копии страниц интернета, и при этом был получен в результате определенного и фиксированного числа вычислительных шагов[3]. Чтобы у оракула не возникало соблазна манипулировать нашим сознанием и вынуждать нас задавать ему более простые вопросы (предположим, перед ним стоит задача максимизировать точность ответов), перед ним можно было бы поставить цель: уничтожать заданный вопрос сразу, как только он даст на него ответ. Причем вопрос мог бы сразу загружаться в его память перед запуском программы. Чтобы задать второй вопрос, можно было бы перезагружать машину и запускать ту же программу с уже другим предварительно загруженным в память вопросом.

Для управления оракулом нужна довольно простая система мотивации, однако даже в этом случае возникают почти незаметные, но потенциально опасные проблемы. Предположим, нам необходимо точно сформулировать для ИИ значение фраз: «свести к минимуму свое воздействие на мир при достижении определенных результатов» и «для подготовки ответа использовать только определенные выделенные ресурсы». Что произойдет, если ИИ в ходе своего интеллектуального совершенствования устроит что-то вроде научной революции, изменив при этом свою базовую онтологию?[4] Сначала нам надо дать определение таким понятиям, как «воздействие» и «выделенные ресурсы», в соответствии с нашей онтологией (постулируя существование различных физических объектов, в частности компьютеров). Но вдруг случается неожиданное: ИИ-оракул обнаруживает, что некоторые наши самые распространенные категории основаны на ложных представлениях — очень напоминает поведение людей, когда каждое новое поколение ученых пытается избавиться от онтологических категорий, введенных когда-то их предшественниками (кто сейчас помнит такие понятия, как «флогистон», «сила жизни» и «абсолютная одновременность»?). Система целеполагания ИИ, переживающего онтологический кризис, должна быть достаточно эластичной, чтобы он смог сохранить и суть, и дух своей первоначальной цели и бережно перенести этот багаж в новую систему ценностей.

Не всегда задача создать дружественный ИИ-оракул с помощью метода выбора мотивации может оказаться тривиальной — однако она все равно проще, чем задача разработать ИИ, обладающий свободой в достижении сложной цели. Кажется, это могло бы служить аргументом в пользу того, чтобы первым сверхразумом стал оракул. С ним легко работать, пользуясь так же методом контроля над возможностями, — чем не следующий аргумент в пользу увидеть именно оракула в роли первого сверхразума? К оракулу применимы все стандартные изоляционные методы. Кроме того, существуют методы, которые можно использовать, только имея дело с оракулом. Например, возможен риск, что, преследуя какие-то свои тайные планы, оракул начнет давать не очень достоверные ответы, явно тем самым стараясь манипулировать нашим сознанием. Есть замечательный способ снизить эту угрозу: создать множество оракулов, в легкой степени отличающихся друг от друга своими исходным кодом и базой данных. Затем с помощью простой функции сравнивать их ответы и передавать их заказчику только в случае, когда они все согласуются между собой. Правда, можно допустить, что любой

универсальный сверхразумный оракул узнает о существовании других оракулов (выведет это из своих знаний о человеческом обществе), возможно, от него удастся скрыть некоторые детали их функционирования. Если оракулы не смогут связываться между собой, им будет трудно координировать ответы, чтобы манипулировать нами. Есть множество способов отклоняться от истины, и разные оракулы могут по-разному оценивать привлекательность каждого варианта — в то время как сама истина является фокальной точкой (по которой возможно достижение согласия в отсутствие коммуникаций между сторонами). Поэтому если оракулы придут к консенсусу, это может быть признаком, что они дают правдивый ответ[5].

Мы можем полностью полагаться на ИИ-оракула, если знаем, что он использует все свои возможности, чтобы давать достоверные ответы. Но даже если мы не во всем доверяем ему, он был бы нам полезен. Оракулу можно было бы задавать вопросы, на которые трудно найти ответы, но легко проверить, правильные ли они. Например, многие математические задачи. Если вас интересует вопрос истинности какого-то математического утверждения, можно попросить оракула найти доказательство этому или опровергнуть его. Поиск самого доказательства может требовать проницательности и таланта, превышающего наши возможности, а его проверку можно выполнить при помощи простой механической процедуры.

Если окажется, что контролировать все ответы ИИ-оракула слишком дорого (скажем, когда вопросы не относятся к области логики и математики), для проверки можно выбирать случайные подмножества ответов. Если они окажутся верными, у нас появятся основания присвоить высокую вероятность тому, что верными являются и все остальные ответы. Эта хитрость поможет серьезно сэкономить в целом на определении степени достоверности ответов, которые было бы слишком затратно проверять по отдельности. (К сожалению, это ничего не скажет о достоверности ответов, которые мы *неспособны* проверить, поскольку решившийся на обман оракул может решить отвечать достоверно только на те вопросы, правильность ответов на которые мы, по его мнению, оценить в состоянии.)

Иногда не мешало бы иметь под рукой конкретные рекомендации, позволяющие выявлять, насколько корректен тот или иной ответ (или помогающие ориентироваться, где и как мы могли бы проверить правильность ответа), правда, мы не всегда можем доверять подобным советам. Например, по нашей просьбе оракул будет искать решения разных технических

и философских проблем, которые могут возникнуть в процессе разработки более совершенных методы выбора мотивации. Если нам покажется, что предложенный оракулом проект безопасен, можно попросить его отыскать в нем серьезные недостатки и при наличии таковых описать каждый изъян при помощи не более двадцати слов. Вопросы такого рода способны обеспечить нас ценной информацией. Однако нам следует проявлять осторожность и сдерживать себя, чтобы не задавать *слишком много* подобных вопросов — и не позволять себе *слишком* полагаться на детали данных нам ответов, — иначе у недобросовестного оракула появится возможность повлиять на наши убеждения (при помощи вроде бы безобидных, но вполне манипулятивных сообщений). ИИ, обладающему сверхмощью в области социального манипулирования, не потребуется много времени, чтобы подчинить нас своей воле.

Даже если ИИ-оракул работает в полном соответствии с нашими требованиями и желаниями, существует риск, что им будут неправильно пользоваться. Одно из очевидных проявлений этой проблемы состоит в том, что оракул, наделенный сверхразумом, способен стать источником огромной власти и обеспечить своему оператору или программисту решающее стратегическое преимущество. Эта незаконная власть, скорее всего, будет использоваться отнюдь не в интересах общества. Не столь явный, но не менее важный аспект заключается в том, что постоянная работа с оракулом таит в себе огромную опасность для самого оператора. Все наши тревоги — как с мировоззренческой, так и технической точек зрения — имеют отношение и к остальным кастам сверхразума. Подробнее мы рассмотрим эту проблему в главе тринадцатой. Пока достаточно сказать, что чрезвычайно большое значение имел бы протокол, содержащий полную информацию о том, какие и в какой последовательности были заданы вопросы и какие были даны ответы. Можно подумать над тем, чтобы разработать такой вариант оракула, который будет отказываться отвечать на вопросы, если сочтет, что ответы могут иметь катастрофические последствия с точки зрения общепринятых в человеческом сообществе норм.

Джинны и монархи

Джинн — интеллектуальная система исполнения команд. Джинн получает команду высокого уровня, выполняет ее и останавливается в ожидании следующей команды[6]. Монарх — система, получившая мандат на любые

действия в мире для достижения некоторых масштабных и, возможно, очень долгосрочных целей. Описания этих систем не очень напоминают то, что мы привыкли считать эталоном сверхразума, — но так кажется лишь на первый взгляд.

В случае ИИ-джинна приходится пожертвовать одним из самых привлекательных свойств оракула: возможностью использовать изоляционные методы. Можно, конечно, рассмотреть возможность разработки заблокированного джинна, способного создавать объекты лишь в некотором ограниченном пространстве — пространстве, окруженном стенами с мощными укрепительными системами или заминированными барьерами, которые должны сдетонировать в случае попытки побега. Трудно с уверенностью говорить о высокой безопасности такой физической изоляции, если речь идет о сверхразуме, вооруженном универсальными манипуляторами и инновационными конструкционными материалами. Даже если каким-то образом удастся обеспечить джинну такую же надежную изоляцию, как и оракулу, все равно не очень понятно, что мы выиграем, открыв сверхразуму прямой доступ к манипуляторам, вместо того чтобы получить от него подробные описания, которые можно было бы внимательно изучить, а затем использовать, чтобы получить требуемый результат самим. Выигрыш в скорости и удобстве из-за устранения человека-посредника вряд ли стоит потери возможности использовать более надежные методы блокировки, доступные в случае оракула.

Если кто-нибудь все-таки *создаст* джинна, было бы желательно, чтобы этот ИИ подчинялся не буквальному смыслу команд, а скорее намерениям, лежащим в их основе, поскольку джинн, воспринимающий команды слишком дословно (при условии, что он достаточно сверхразумен, чтобы обеспечить себе решающее стратегическое преимущество), может пожелать убить и пользователя, и все остальное человечество при первом же включении — по причинам, изложенным в разделе о пагубных отказах системы в восьмой главе. В целом важно, чтобы джинн всегда искал доброжелательный вариант интерпретации данной ему команды — как для себя, так и для всего человечества, — и чтобы был мотивирован именно на такое, а не на буквальное ее выполнение. Идеальный ИИ-джинн должен быть скорее первоклассным вышколенным дворецким, нежели гениальным савантом-аутистом.

Однако ИИ-джинн, обладающий чертами профессионального дворецкого, приблизился бы к тому, чтобы претендовать на место в касте монархов. Рассмотрим для сравнения идею создания ИИ-монарха с конечной целью

руководствоваться духом команд, которые мы дали бы ему, если бы создавали не монарха, а джинна. Такой монарх имитировал бы джинна. Будучи сверхразумным, он мог бы с легкостью догадаться, какие команды мы дали бы джинну (и всегда спросить нас, если бы это помогло ему в принятии решения). Была бы в таком случае какая-то заметная разница между монархом и джинном? Или, если посмотреть на различие между ними с другой стороны с учетом варианта, что сверхразумный джинн мог бы точно предсказывать, какие команды он получит, какой выигрыш даст то, что он будет вынужден ждать этих команд, чтобы начать действовать?

Можно было бы думать, что преимущество джинна перед монархом огромно, поскольку, если что-то пойдет не так, джинну всегда можно дать новую команду остановиться или исправить результаты своего действия — в то время как монарх продолжал бы задуманное невзирая на наши протесты. Но высокая безопасность джинна, как мы ее себе представляем, во многом иллюзорна. Кнопки «стоп» или «отмена» сработают у джинна только в случае неопасного отказа, но если дело касается пагубного отказа, скажем, выполнение текущей команды становится для джинна конечной целью, — он просто проигнорирует любые наши попытки отменить предыдущую команду[7].

Можно было бы попробовать создать джинна, который будет автоматически прогнозировать наиболее характерные проблемы, которые обрушатся на пользователей, если джинн выполнит данную ему команду, при этом джинн должен будет запрашивать подтверждение каждый раз перед ее исполнением. Такую систему можно было бы назвать *джинн с ратификацией*. Но если мы в силах разработать такого джинна, то почему бы не создать подобного монарха? То есть и в этом случае мы не сможем провести четкую дифференциацию. (Возможность взглянуть на результат еще до выполнения самой команды кажется очень привлекательной, но если функция ратификации прогноза будет когда-либо создана, то перед нами встанут очередные вопросы, что с нею делать дальше и каким образом ее оптимально использовать. Позже мы вернемся к этой теме.)

Способность одной касты ИИ подражать другой распространяется и на оракулов. Джинн мог бы имитировать действия оракула, если единственные команды, которые мы ему даем, были бы связаны с необходимостью отвечать на конкретные вопросы. В свою очередь, оракул в состоянии заменить джинна, когда ему поступает запрос на разработку какой-нибудь

рекомендации. Оракул выдаст пошаговую инструкцию, как джинну достичь того или иного результата, и даже напишет для него исходный код[8]. Это верно и в отношении сходства между оракулом и монархом.

Таким образом, реальная разница между тремя типами ИИ заключается не в их возможностях. Скорее, отличие связано с разными подходами к решению проблемы контроля. С каждой кастой ИИ связан свой набор мер предосторожности. По отношению к оракулу будет лучше всего применять изолирующие методы; наверное, подойдет и такой метод, как приручение. Джинна запереть сложнее, поэтому намного эффективнее будет использовать метод приручения. Однако ни изоляции, ни приручению не поддастся монарх.

Будь меры предосторожности решающим обстоятельством, иерархия была бы очевидна: оракул безопаснее джинна, а джинн безопаснее монарха — и все исходные различия (удобство и быстродействие) ушли бы в тень, уступив первенство единственному преимуществу, ради которого выбор всегда бы делался в пользу оракула. Однако следует принимать во внимание и другие факторы. Выбирая между кастами, нужно учитывать не только степень угроз, исходящих от самой системы, но и опасность, которая возникает в результате ее возможного использования. Очевидно, что джинн наделяет контролирующего его человека огромной властью, но то же самое можно сказать и об оракуле[9]. В отличие от них монарха можно было бы разработать таким образом, чтобы ни у кого (человека или группы людей) не было бы преимущественного права влиять на результаты работы системы и чтобы всякий раз ИИ сопротивлялся при малейшей попытке вмешаться в его деятельность или изменить его программные параметры. Более того, если мотивация монарха определена при помощи метода косвенной нормативности (этот метод упоминался в предыдущей главе, и мы вернемся к нему в тринадцатой главе), такой ИИ можно будет использовать для достижения некоего абстрактно заданного результата, например «максимально справедливого и этически допустимого» — без необходимости заранее представлять точно, каким он должен быть. Это привело бы к возникновению ситуации, аналогичной «вуали неведения» Джона Ролза[10]. Такие условия способны облегчить достижение консенсуса, помочь предотвратить конфликт и привести к более справедливому результату.

Еще одно соображение — не в пользу оракулов и джиннов — касается риска создания сверхразума, чья конечная цель не будет полностью отвечать

тому, чего в конечном счете нам хотелось бы добиться. Допустим, прибегнув к методу приручения, мы уговорим сверхразум стремиться к тому, чтобы минимизировать свое воздействие на мир, тогда мы сможем получить интеллектуальную систему, чьи оценки предпочтительности тех или иных исходов будут отличаться от оценок организаторов проекта. То же самое произойдет, если мы создадим сверхразум, чрезмерно высоко ценящий свою способность давать абсолютно достоверные ответы или слепо повиноваться любой команде. Если будут предприняты соответствующие меры предосторожности, это не должно вызвать особых проблем: между двумя системами оценок будет мало различий — по меньшей мере до тех пор, пока они относятся к возможным мирам, у которых много шансов быть актуализованными. Поэтому результаты, которые окажутся правильными по стандартам интеллектуального агента, будут правильными и с точки зрения принципала. Возможно, кто-то возразит, что подобный принцип разработки неудачен, поскольку неблагоразумно вносить даже легкую дисгармонию между целями ИИ и целями человечества. (Конечно, аналогичные сомнения возникают, если монархам начнут определять цели, не полностью гармонирующие с нашими, человеческими.)

ИИ-инструменты

В свое время было высказано предложение создавать сверхразум скорее в качестве инструмента, чем агента[11]. Идея возникла неслучайно, и связана она с простым соображением: обычным программным обеспечением пользуются все подряд, и ни у кого не возникает никакого чувства опасности, даже отдаленно напоминающего ту тревогу, которую вызывают у нас проблемы, обсуждаемые в этой книге. Почему бы не создать ИИ, похожий на обычное ПО, — вроде системы управления полетом или виртуального помощника, — только более гибкое и универсальное? Зачем нужен сверхразум, обладающий собственной волей? Те, кто придерживается такой точки зрения, считают, что сама парадигма агента фундаментально ошибочна. Вместо ИИ, который, подобно человеку, думает, желает и действует, нам следует ориентироваться на написание ПО, делающее лишь то, для чего оно предназначено.

Однако идея создания ПО, которое «делает лишь то, для чего предназначено», не так легко осуществима, поскольку речь идет о продукте с очень мощным интеллектом. В каком-то смысле все программы делают то, на что

они запрограммированы: их поведение математически определяется исходным кодом. Но это утверждение так же верно и для ИИ, принадлежащего какой-то из трех каст. Если *делать лишь то, для чего предназначено* означает «вести себя так, как *предполагали* программисты», то стандартное ПО довольно часто нарушает этот стандарт.

Благодаря ограниченным возможностям современного ПО (по сравнению с ИИ) с последствиями его отказов пока можно справиться — они будут оцениваться где-то между значением «несущественный» и «дорогостоящий», но никогда не поднимутся до уровня экзистенциальной угрозы[12]. Однако если относительно безопасными стандартные современные ПО делает не высокая надежность, а ограниченные возможности, то непонятно, как они могут стать образцом для создания безопасного сверхразума. Может быть, потребность в УИИ можно удовлетворить за счет расширения диапазона задач, решаемых обычным ПО? Но диапазон и разнообразие задач, которые ИИ успешно решил бы в современных условиях, огромен. Вряд ли для их решения возможно создать ПО специального назначения. Но даже если это и можно сделать, такой проект занял бы слишком *много* времени. Еще до его завершения обязательно изменится сущность самого задания, поскольку одни проблемы утратят свою злободневность, а другие, пока еще невыявленные, станут актуальными. Наличие программы, которая может самостоятельно учиться решать новые задачи и, более того, формулировать их, а не только справляться с чужими формулировками, дало бы нам огромные преимущества. Но тогда нужно, чтобы программа имела возможность учиться, мыслить и планировать, причем делать это на высоком уровне и не ограничиваться одной или несколькими областями знаний. Иными словами, нужно, чтобы она обладала общим уровнем интеллекта.

В нашем случае особенно важна задача разработки самого ПО. С практической точки зрения огромный выигрыш дала бы автоматизация этого процесса. Хотя такой же критически важной является и способность к быстрому самосовершенствованию, ведь именно она позволяет зародышу ИИ обеспечить взрывное развитие интеллекта.

Если наличие общего уровня интеллекта не является обязательным, существуют ли иные способы реализовать идею ИИ-инструмента так, чтобы он не вырвался за рамки пассивного «решателя» задач? Возможен ли ИИ, не являющийся агентом? Интуиция подсказывает, что безопасным обычное ПО делает не ограниченность его возможностей, а отсутствие амбиций.

В Excel нет подпрограмм, тайно мечтающих завоевать мир, будь у них соответствующие возможности. Электронные таблицы вообще ничего не «хотят», они всего лишь слепо выполняют команды, записанные в их код. Может возникнуть вопрос: что мешает нам создать программу такого же типа, но обладающую более развитым интеллектом? Например, оракула, который в ответ на описание цели выдал бы план ее достижения, так же как Excel в ответ на ввод чисел в ячейки выдает их сумму, то есть не имея никаких «предпочтений» относительно результата своих расчетов или того, как люди могут им воспользоваться?

Классический путь написания программ требует от программиста довольно детального понимания задачи, которая должна быть разработана, чтобы можно было явно задать ход ее решения, состоящий из последовательности математически точно описанных шагов, выраженных в исходном коде[13]. (На практике программисты полагаются на библиотеки подпрограмм, выполняющих определенные функции, которые можно просто вызывать без необходимости разбираться в деталях их реализации. Но эти подпрограммы изначально были созданы людьми, которые все-таки отлично разбирались в том, что делали.) Этот подход работает при решении хорошо знакомых задач, чем и занято большинство существующих ПО. Однако он перестает работать в ситуации, когда никто толком не понимает, как должны быть решены стоящие перед программой задачи. Именно в этом случае становятся актуальными методы из области разработок искусственного интеллекта. В некоторых приложениях можно использовать машинное обучение для точной настройки нескольких параметров программ, в остальном полностью созданных человеком. Например, спам-фильтр можно обучать на массиве вручную отобранных сообщений электронной почты, причем в ходе этого обучения классифицирующим алгоритмом будут изменяться веса, которые он присваивает различным диагностическим атрибутам. В более амбициозном приложении можно создать классифицирующий механизм, который будет сам обнаруживать такие атрибуты и тестировать их пригодность в постоянно меняющейся среде. Еще более совершенный спам-фильтр может быть наделен некоторыми возможностями размышлять о компромиссах, на которые готов пойти пользователь, или о содержании анализируемых им сообщений. Ни в одном из этих случаев программисту не нужно знать наилучший способ отделения спама от добропорядочной почты — он должен лишь определить алгоритм, при помощи которого спам-фильтр сам

улучшит свою эффективность за счет обучения, обнаружения новых атрибутов или размышлений.

По мере развития ИИ у программиста появится возможность сэкономить большую часть умственных сил, которые нужны для поиска путей решения стоящей перед ним задачи. В предельном случае ему будет достаточно задать формальный критерий успешности решения и предложить задачу ИИ. В своем поиске ИИ будет руководствоваться набором мощных эвристических правил и методов, позволяющих выявить структуру пространства возможных решений. ИИ мог бы продолжать свой поиск до тех пор, пока не будет найдено решение, удовлетворяющее критерию успеха. А затем или внедрить решение самостоятельно, или (например, оракул) сообщить о нем пользователю.

Элементарные формы такого подхода сегодня уже используются очень широко. Тем не менее ПО, в котором работают методы ИИ и машинного обучения, хотя и имеет некоторые шансы найти решение, неожиданное для людей, их создавших, во всех практических смыслах функционирует как обычные программы и не создает экзистенциального риска. В опасную зону мы попадаем лишь тогда, когда методы, используемые в поиске, становятся слишком мощными и универсальными, то есть когда они начинают переходить на общий уровень интеллекта, а особенно — на уровень сверхразума.

Есть (как минимум) два случая, когда могут возникнуть проблемы.

Во-первых, сверхразумный процесс поиска может найти решение, которое не только неожиданно, но и категорически неприемлемо. Это приведет к пагубному отказу по одному из обсуждавшихся выше типов (порочная реализация, инфраструктурная избыточность, преступная безнравственность). Особенно очевидна такая возможность, когда действуют монарх и джинн, напрямую воплощающие в жизнь найденные ими решения. Если компьютерные модели, призванные символизировать счастье, или заполонение планеты скрепками — первые из обнаруженных сверхразумом решений, удовлетворяющие критерию успеха, тогда мы получим сплошные смайлики и скрепки[14]. Но даже оракул, всего лишь *сообщающий* о решении, если все идет хорошо, — может стать причиной порочной реализации. Пользователь просит оракула представить план достижения определенного результата или технологию выполнения определенной функции, а затем следует этому плану или воплощает в жизнь технологию, в результате чего

сталкивается с порочной реализацией точно так же, как если бы реализацией решения занимался сам ИИ[15].

Во-вторых, проблемы могут возникнуть на этапе работы самого ПО. Если методы, которыми оно пользуется для поиска решения, достаточно сложны, они могут допускать управление процессом поиска в интеллектуальном режиме. В этом случае компьютер, на котором запущено ПО, будет выглядеть уже не как инструмент, а скорее как агент. То есть программа может начать разрабатывать план проведения поиска. В ее плане будут определены области, которые следует изучить в первую очередь, методы их изучения, данные, которые нужно собрать, модель использования наилучшим образом имеющихся вычислительных мощностей. Разрабатывая план, отвечающий внутреннему критерию ПО (в частности, который имеет довольно высокую вероятность привести к решению, удовлетворяющему определенному пользователем критерию в отведенное на это время), программа может остановиться на какой-то необычной идее. Например, план может начаться с получения дополнительных вычислительных мощностей и устранения потенциальных препятствий (в том числе людей). Столь «творческий подход» вполне возможен после достижения ПО высокого интеллектуального уровня. Если программа решит реализовать такой план, это приведет к экзистенциальной катастрофе.

ВРЕЗКА 9. НЕОЖИДАННЫЕ РЕЗУЛЬТАТЫ СЛЕПОГО ПОИСКА

Даже простые процессы эволюционного поиска иногда приводят к совершенно неожиданным для пользователя результатам, которые тем не менее формально удовлетворяют поставленным критериям.

Область способного к эволюции аппаратного обеспечения представляет много примеров данного явления. Поиск проводится при помощи эволюционного алгоритма, который прочесывает пространство возможных схем аппаратных средств и тестирует каждую из них на пригодность путем реализации каждого варианта в виде интегральной схемы и проверки правильности ее функционирования. Часто в результате эволюционного дизайна удается достичь значительной экономии. Например, в ходе одного из подобных экспериментов была обнаружена схема дискриминации частот, которая функционировала без тактового генератора — компонента, считавшегося обязательным для выполнения такого рода функции. Исследователи оценили, что схемы, полученные в результате эволюционного дизайна, на один-два порядка меньше, чем те, которые для тех же целей создали бы инженеры-люди. Такие схемы использовали физические свойства входящих в них компонентов совершенно нетрадиционными способами, в частности, некоторые активные и необходимые для

работы компоненты вообще не были соединены с входными или выходными ножками! Вместо этого они взаимодействовали с другими компонентами за счет того, что обычно считается досадными помехами: скажем, электромагнитных полей или нагрузки источника питания.

Другой эксперимент по эволюционной оптимизации с заданием разработать осциллятор, привел к исчезновению из схемы, казалось бы, еще более необходимого компонента — конденсатора. Когда успешное решение было получено и ученые посмотрели на него, то первой реакцией были слова: «Это не будет работать!» Однако после более тщательного анализа оказалось, что алгоритм, словно секретный агент Макгайвер*, переконфигурировал свою материнскую плату, лишенную датчиков, в импровизированный радиоприемник, использовав дорожки печатной схемы в качестве антенны для приема сигналов, генерируемых компьютером, который располагался поблизости в той же лаборатории. Затем эти сигналы усиливались схемой и преобразовывались в выходной сигнал осциллятора[16].

В других экспериментах эволюционные алгоритмы разрабатывали схемы, которые определяли, что материнскую плату проверяли осциллографом или что в лаборатории в розетку включали паяльник. Эти примеры показывают, как программы в процессе свободного поиска могут изменить назначение доступных им ресурсов, чтобы обеспечить себе неожиданные сенсорные возможности такими средствами, которые привычно мыслящий человеческий ум не готов не только использовать, но и просто понять.

Тенденция эволюционного поиска: отыскивать «хитрые» решения и совершенно неожиданные пути достижения цели — проявляется и в природе, хотя мы считаем вполне нормальными знакомые нам результаты биологической эволюции, даже если и не были бы готовы спрогнозировать их. Зато можно провести эксперименты с искусственным отбором, в ходе которых увидеть работу эволюционного процесса вне рамок привычного контекста. В таких экспериментах исследователи могут создавать условия, редко встречающиеся в природе, и наблюдать за их результатами.

Например, до 1960-х гг. среди биологов было распространено мнение, что популяции хищников ограничивают свой рост, чтобы не попасть в мальтузианскую ловушку[17]. И хотя индивидуальный отбор работал против такого ограничения, многие считали, что групповой отбор должен подавлять индивидуальные склонности использовать любые возможности для продолжения рода и поощрять такое поведение, которое благоприятно сказывается на всей группе или популяции в целом. Позднее теоретический анализ и моделирование показали, что хотя групповой отбор и возможен в принципе, он способен победить индивидуальный отбор в очень редко встречающихся в природе условиях[18]. Зато такие условия могут быть созданы в лаборатории. Когда при помощи группового отбора особей мучного хрущака (*Tribolium castaneum*)

---

* *«Секретный агент Макгайвер»* (MacGyver) — популярный американский телесериал (1985–1992).

попытались добиться уменьшения размера их популяции, это действительно удалось сделать[19]. Однако методы, благодаря которым был получен требуемый результат, включали не только «благоприятное» приспособление в виде снижения плодовитости и увеличения времени на воспроизводство, которых можно было бы наивно ожидать от антропоцентричного эволюционного поиска, но и рост каннибализма[20].

Как показывают примеры, приведенные во врезке 9, процессы неограниченного поиска решений иногда выдают странные, неожиданные и не антропоцентричные результаты даже в нынешнем своем весьма ограниченном виде. Современные поисковые процессы неопасны, поскольку слишком слабы, чтобы разработать план, способный привести к их господству над миром. Такой план должен включать чрезвычайно сложные шаги вроде создания новых видов оружия, на несколько поколений опережающих существующие, или проведение пропагандистской кампании, гораздо более эффективной, чем те, что доступны современным механизмам манипулирования людьми. Чтобы у машины появилась возможность хотя бы *помыслить* об этих идеях, не говоря уже об их воплощении в жизнь, вероятно, ей нужно уметь представлять мир как минимум так же реалистично и детально, как это делает обычный взрослый человек (хотя отсутствие знаний в определенных областях может быть скомпенсировано чрезвычайно развитыми навыками в других). Пока это далеко превышает уровень имеющихся систем ИИ. А учитывая комбинаторный взрыв, который обычно пресекает попытки решить сложные задачи планирования при помощи методов перебора (мы видели это в первой главе), недостатки известных алгоритмов не могут быть преодолены простым наращиванием вычислительной мощности[21]. Однако по мере развития процессов поиска или планирования растет и их потенциальная опасность.

Возможно, вместо того чтобы позволить спонтанное и опасное развитие целенаправленного поведения агентов при помощи мощных поисковых алгоритмов (включая процессы планирования и прямого поиска решений, удовлетворяющих определенным критериям пользователя), лучше было бы создать агента намеренно. Наделив сверхразум явной структурой агентского типа, можно было бы повысить его предсказуемость и прозрачность. Хорошо разработанная система с четким разделением между целями и навыками позволила бы нам делать прогнозы относительно результатов, которые она будет выдавать. Даже если мы не сможем точно сказать, к какому мнению

придет система или в каких ситуациях окажется, будет понятно, в каком месте можно проанализировать ее конечные цели и, как следствие, критерии, которыми она воспользуется при выборе своих действий и оценке потенциальных планов.

Сравнительная характеристика

Будет полезно обобщить свойства различных каст ИИ, которые мы обсудили (табл. 11).

Таблица 11. Свойства различных типов интеллектуальных систем

Оракул	**Система для ответа на вопросы**	Полностью применимы изоляционные методы. Полностью применим метод приручения. Сниженная потребность в понимании человеческих намерений и интересов (по сравнению с джиннами и монархами). Использование вопросов, на которые существуют однозначные ответы типа «да» и «нет», поможет избавиться от необходимости измерять «полезность» или «информативность» ответов
	Варианты: оракулы, ограниченные одной областью знаний (например, математикой); оракулы с ограниченным доступом к каналу вывода (например, дающие ответы: «да», «нет», «нет решения», «почти наверное»); оракулы, отказывающиеся отвечать на вопросы, содержащие хотя бы намек на некоторый заранее определенный критерий «бедствия»; множественные оракулы для сравнения ответов	Источник огромной власти (могут обеспечить оператору решающее стратегическое преимущество). Ограниченная защита от неправильного использования оператором. Ненадежных оракулов можно было бы использовать для получения ответов на вопросы, которые трудно отыскать, но легко проверить. Облегченную форму проверки ответов можно было бы проводить за счет использования множества оракулов

Джинн	**Система исполнения команд**	Отчасти применимы изоляционные методы (для пространственно ограниченных джиннов). Отчасти применим метод приручения. Джинны могут прогнозировать наиболее характерные проблемы и запрашивать подтверждение на выполнение команд
	Варианты: джинны, использующие различные «дистанции экстраполяции» или степени следования скорее духу, нежели букве команд; джинны, ограниченные отдельными областями знаний; джинны, отказывающиеся выполнять команды, если они предсказывают, что их выполнение отвечает некоторому заранее определенному критерию «бедствия»	Джинны могли бы разбивать выполнение команд на этапы, чтобы контролировать промежуточные результаты. Источник огромной власти (могут обеспечить оператора решающим стратегическим преимуществом). Ограниченная защита от неправильного использования оператором. Большая потребность понимать человеческие намерения и интересы (по сравнению с оракулами)
Монарх	**Система, предназначенная для независимого выполнения автономных операций**	Изоляционные методы неприменимы. Большинство других методов контроля над возможностями также неприменимы (за исключением социальной интеграции и антропного захвата). Метод приручения в большинстве случаев неприменим. Высокая потребность в понимании истинных человеческих намерений и интересов. Необходимость правильной реализации с первого раза (в принципе, в той или иной степени это верно для всех каст)
	Варианты: многие возможные системы мотивации; возможность использования оценки и «ратификации организатором» (см. главу 13)	Потенциально источник огромной власти для организатора, включая решающее стратегическое преимущество. После активации не подвержен взлому со стороны оператора и может быть

		снабжен некоторой защитой от неправомерного использования. Может использоваться для реализации исходов типа «вуаль неведения» (см. главу 13). Могут быть применимы изоляционные методы в зависимости от реализации. При разработке и функционировании машинного сверхразума, скорее всего, будут использоваться мощные поисковые процессы
Инструмент	**Система, не предназначенная для целенаправленного поведения**	Мощный поиск с целью найти решение, удовлетворяющее некоторым формальным критериям, может привести к открытию решения, которое отвечает этим критериям незапланированным и опасным способом Мощный поиск может включать в себя вторичный, внутренний поиск и процессы планирования, которые выявят опасные способы проведения основного поиска

Для определения, какой тип системы будет самым безопасным, требуется проведение дополнительных исследований. Ответ может зависеть от условий, при которых используется ИИ. Каста оракулов, очевидно, привлекательна с точки зрения безопасности, поскольку к оракулам применимы и методы контроля над возможностями, и методы выбора мотивации. В этом смысле может показаться, что оракулы предпочтительнее монархов, которым подходят лишь методы выбора мотивации (за исключением ситуаций, когда в мире существуют и другие мощные сверхразумные системы, — в этом случае могут применять социальную интеграцию или антропный захват). Однако оракул может дать оператору слишком большую власть, что опасно в случае коррумпированного или неблагоразумного оператора, в то время как монарх дает некоторую возможность защититься от таких неприятностей. Так что ранжирование каст с точки зрения безопасности не столь очевидно.

Джинна можно считать компромиссным решением, но выбор в пользу джинна не всегда будет удачным. Во многих отношениях ему свойственны недостатки обеих каст. Кажущаяся безопасность ИИ-инструмента также может быть иллюзорной. Чтобы такая система могла стать достаточно универсальной и заменить сверхразумного агента, она должна включать в себя чрезвычайно мощные процессы внутреннего поиска и планирования. Эти процессы могут иметь незапланированные последствия в виде поведения агентского типа. В этом случае было бы лучше сразу разрабатывать систему как агента, чтобы программисты могли четче видеть, какие критерии определяют результаты ее работы.

Глава одиннадцатая

Сценарии многополярного мира

Мы уже не раз убеждались (особенно в главе восьмой), какой опасной может быть развязка, приводящая к однополярному миру, когда единственный сверхразум получает решающее стратегическое преимущество и использует его для формирования синглтона. В этой главе мы рассмотрим, что может произойти в случае многополярного исхода: возникновение постпереходного общества со множеством конкурирующих сверхразумных агентов. Этот сюжетный вариант интересует нас по двум причинам. Во-первых, одним из решений проблемы контроля может быть метод социальной интеграции — его достоинства и недостатки были отмечены в девятой главе. Детально он будет описан в этой главе. Во-вторых, даже при отсутствии намеренного стремления создать многополярные условия, чтобы решить проблемы контроля, они могут сложиться сами собой. Как тогда будет выглядеть исход? Совсем необязательно, что появившееся общество, основанное на конкуренции, окажется перспективным и просуществует сколь-нибудь долго.

В сценарии однополярного мира, когда критический рубеж уже перейден и наступило правление синглтона, — все, что произойдет потом, будет полностью зависеть от его системы ценностей. Поэтому исход может быть как очень хорошим, так и очень плохим, в зависимости от того, какие у синглтона замыслы. В свою очередь, его ценностная ориентация зависит от того, была ли решена проблема контроля и — если была решена — от целей проекта, в рамках которого был создан синглтон.

Итак, те, кого интересуют возможные сценарии исхода с формированием синглтона, могут опереться на три источника возможной информации: факторы, на которые синглтон воздействовать не может (например, физические законы); конвергентные инструментальные цели; предположительные конечные цели.

В сценариях многополярного мира появляется дополнительный набор условий, тоже играющий информационную роль, — разные характеры

взаимоотношений агентов. Вследствие их взаимодействия возникает та или иная социальная динамика, которую можно проанализировать, используя знания по теории игр, экономике и теории эволюции, также подойдут некоторые элементы политологии и социологии — настолько, насколько мы освободим их от случайных черт человеческого опыта. Было бы странно ожидать, что из этих источников мы сможем извлечь достаточно информации для составления полной картины постпереходного мира, но они помогут наметить самые яркие вероятные сценарии и оспорить самые беспочвенные предположения.

Мы начнем с изучения модели сценария, характеризующейся определенными экономическими показателями: низким уровнем регулирования, сильной защитой прав собственности и умеренным распространением недорогих систем искусственного интеллекта[1]. Так называемую экономическую модель чаще всего связывают с именем американского экономиста Робина Хэнсона, который стал первооткрывателем этой темы. Далее мы рассмотрим некоторые эволюционные соображения и изучим перспективы многополярного постпереходного мира, впоследствии превращающегося в синглтон.

О лошадях и людях

Универсальный искусственный интеллект мог бы стать заменой человеческому интеллекту. Причем УИИ будет способен выполнять не только умственную работу, но и физический труд — правда, последним он займется не сам, а заменит людей на исполнительные устройства и роботизированные механизмы. Предположим, работники-машины, которых очень легко воспроизводить, стали и дешевле, и способнее работников-людей почти во всех профессиях. Что тогда произойдет?

Заработная плата и безработица

С появлением легко воспроизводимых работников-машин упадут рыночные зарплаты. Человек сохранит свою конкурентоспособность только в тех сферах профессиональной деятельности, где клиенты-люди будут искать общения с себе подобными, а не машинами. Сегодня товары, сделанные вручную или выпущенные коренными жителями тех или иных мест по традиционной технологии, часто стоят дороже. Возможно, и в будущем потребители будут выбирать продукты, произведенные людьми, а также

предпочитать спортсменов-людей, художников-людей, возлюбленных-людей и политиков-людей, хотя с функциональной стороны искусственное существо окажется неотличимым от человека и даже превосходящим его. Неясно только, сколько продлятся эти предпочтения. Ведь если киберальтернативы станут заметно превосходить людей на всех направлениях, возможно, ценить их начнут больше.

Для выбора потребителя может оказаться важным такой параметр, как внутренний мир человека, оказывающего услугу или производящего товар. Меломаны будут особенно ценить тот факт, что исполнители чувствуют музыку и ощущают реакцию своих слушателей. Музыкант, лишенный субъективного восприятия, выглядит скорее как высококачественный проигрыватель, способный создать трехмерную картинку и естественным образом взаимодействующий с публикой. Могут быть созданы машины, способные воспроизводить психические состояния, характерные для людей, выполняющих то же задание. Однако даже при наличии идеальной имитации субъективного опыта некоторые люди выберут «органическую» работу. Такие предпочтения могут иметь, например, идеологические или религиозные корни. По аналогии с тем, что многие мусульмане и евреи избегают употреблять в пищу харамные и некошерные продукты, в будущем появятся потребители, сторонящиеся товаров и услуг, произведенных с привлечением искусственного интеллекта.

Что все это означает? Сократится количество рабочих мест для людей, поскольку дешевый машинный труд сможет заменить человеческий. В принципе, ужас перед наступлением автоматизации производства и потерей из-за этого работы не нов. Периодически — по меньшей мере с начала промышленной революции — обостряются страхи технологической безработицы, при этом не столь многие пострадали от нее так же, как английские ткачи и их подмастерья, объединившиеся в начале XIX века под лозунгами легендарного Неда Лудда, больше известного как Генерал Лудд, для борьбы с механическими ткацкими станками. И тем не менее, хотя машины выполняют многие виды работ, они остаются дополнением к человеческому труду. Во многом благодаря этому дополнению средняя заработная плата людей растет во всем мире. Впрочем, все, что начинается как дополнение, со временем занимает главное место. Поначалу лошадь дополняла повозки и плуги, значительно повышая их производительность. Позднее лошадь полностью уступила свое место автомобилям и тракторам. Новшества

сократили спрос на лошадиный труд и привели к резкому уменьшению популяции этих домашних животных. Не ждет ли аналогичная судьба и наш вид?

История лошади весьма показательна для обсуждаемой темы, поэтому продолжим ее, задавшись вопросом, почему лошади все еще встречаются. Одна из причин — существование ниш, где лошади продолжают удерживать первенство, например при патрулировании парков полицией. Но главная причина заключается в том, что у людей сложились любимые привычки, от которых они не собираются отказываться, например прогулки верхом и скачки. Эти предпочтения аналогичны тем, которые гипотетически будут у нас, людей, в будущем по отношению к определенным продуктам, сделанным человеческими руками. Хотя эта аналогия и наводит на определенные мысли, однако она не совсем точна, поскольку у лошадей до сих пор нет полного функционального аналога. Если появились бы недорогие автоматические устройства, напоминающие настоящих лошадей — способные скакать по лугам, имеющие в точности такую же форму и запах, дающие такие же тактильные ощущения, с такими же характерными привычками, — тогда спрос на биологических лошадей, скорее всего, сократился бы намного сильнее.

В результате резкого падения спроса на человеческий труд зарплаты могут упасть ниже уровня прожиточного минимума. То есть потенциально работники-люди рискуют очень сильно: речь идет не просто о снижении зарплат, понижении в должности или необходимости переобучения, а скорее о перспективах голодной смерти. Когда лошади морально устарели в качестве средства передвижения, многие были проданы на бойню и пошли на собачий корм, костную муку, кожу и клей. У этих животных не было альтернативного источника использования, который окупил бы их содержание. В США в 1915 году было около двадцати шести миллионов лошадей. К началу 1950-х годов осталось два миллиона[2].

Капитал и социальное обеспечение

В чем принципиальная разница между лошадьми и людьми? Суть в том, что у последних есть капитал. Эмпирически доказано, что долгосрочная доля капитала составляет примерно 30 процентов (хотя и подвержена резким краткосрочным колебаниям)[3]. Это значит, что 30 процентов мирового дохода получено владельцами капитала в качестве ренты, а оставшиеся 70 процентов — работниками в виде заработной платы. Если мы отнесем ИИ

к капиталу, тогда с изобретением машинного интеллекта, способного полностью заменить работников-людей, их зарплаты упадут до уровня расходов на содержание работников-машин, способных их заменить, причем расходов очень низких, гораздо ниже прожиточного уровня человека — если исходить из предположения о высокой эффективности машин. То есть доля дохода, приходящаяся на работников-людей, снизится практически до нуля. Но это означает, что на долю капитала будет приходиться почти 100 процентов мирового ВВП. Поскольку в результате взрывного развития искусственного интеллекта мировой ВВП взлетит до небес (из-за появления огромного количества новых машин, заменяющих людей, а также технологических инноваций и позднее — приобретения обширных территорий в результате колонизации космоса), также значительно увеличится совокупный доход от капитала. Если его владельцами останутся люди, совокупный доход человечества вырастет до астрономических масштабов, несмотря на тот факт, что в этом сценарии люди уже не будут получать доходы в виде зарплаты.

То есть человечество в целом может сказочно разбогатеть. Но как будет распределено это богатство? В первом приближении доход от капитала должен быть пропорционален величине самого капитала. Учитывая астрономический мультипликативный коэффициент, даже небольшой капитал на стадии, предшествующей переходу, раздуется в огромное состояние после него. Однако в современном мире у многих людей капитал отсутствует. Это касается не только тех, кто живет в бедности, но и некоторых обладателей высоких доходов или человеческого капитала, имеющих отрицательную величину чистого богатства — в частности, множества молодых представителей среднего класса, у которых нет материальных активов, зато имеются долги по кредитным картам или кредитам на обучение[4]. Но если сбережения могут принести крайне высокие проценты, то нужно иметь хотя бы начальный капитал, на который они могут быть начислены[5].

Тем не менее чрезвычайно богатыми могут стать даже те люди, у которых не было никакого состояния на началл переходного периода. Например, участники пенсионных программ, как государственных, так и частных, если средства этих программ хотя бы частично вложены в фондовый рынок[6]. Те, кто в них не участвует, тоже разбогатеют благодаря филантропии новых богачей: из-за воистину астрономических масштабов нового Эльдорадо даже небольшая доля состояния, направленная на благотворительность, окажется очень большой суммой в абсолютном выражении.

Вполне возможно, что в переходный период останутся люди, которые все-таки будут получать вознаграждение за свой труд, хотя машины функционально превзойдут их во всех областях (а также станут дешевле, чем сотрудники-люди). Мы уже говорили, что могут сохраниться профессии, в которых люди как работники будут предпочтительнее по эстетическим, идеологическим, этическим, религиозным или иным не прагматичным мотивам. Вследствие резкого роста доходов у людей, оставшихся владельцами капитала, соответственно вырастет и спрос на таких работников. Новоиспеченные триллионеры и квадриллионеры смогут позволить себе заплатить щедрую премию за привилегию иметь продукты и услуги, произведенные «органической» рабочей силой — людьми. Здесь снова можно провести параллель с историей лошадей. После сокращения американской популяции лошадей к началу 1950-х годов до двух миллионов особей она начала быстро восстанавливаться и, по недавним статистическим данным, выросла до десяти миллионов голов[7]. Этот рост обусловлен не спросом на лошадей в сельском хозяйстве или на транспорте, а скорее тем, что в результате экономического роста все больше американцев могут позволить себе следовать моде на владение лошадьми для удовольствия.

Еще одно важное отличие лошадей от людей, помимо наличия капитала у последних, состоит в том, что люди способны на политическую мобилизацию. Управляемые людьми правительства могут использовать фискальные полномочия государства для перераспределения частных прибылей или увеличения доходов за счет продажи привлекательных активов, принадлежащих государству, например земли, и использования этих средств на выплату пенсий своим гражданам. Повторим, что благодаря взрывному экономическому росту в переходный период и сразу после него благосостояние будет расти непомерно, и даже безработные окажутся довольно состоятельными людьми. Может так получиться, что какая-то отдельная страна получит возможность обеспечить достойный образ жизни каждому человеку на Земле, выделив на это долю своих доходов, не превышающую то, что сейчас многие страны тратят на благотворительность[8].

Мальтузианские условия в исторической перспективе

До сих пор мы исходили из того, что население Земли не растет. Это может быть обосновано в краткосрочной перспективе, поскольку скорость

воспроизводства ограничена биологически. Однако на длинных временных интервалах это допущение перестает быть верным.

За последние девять тысяч лет человечество выросло тысячекратно[9]. Темпы его роста могли бы быть гораздо более высокими, если бы не тот факт, что и в доисторические времена, и позднее население Земли постоянно упиралось в экономические пределы. Большую часть времени действовали условия, описанные Мальтусом: большинство людей получали доход на уровне прожиточного минимума, который едва-едва позволял им свести концы с концами и вырастить двух детей[10]. Периодически случались временные и локальные передышки в результате эпидемий, голода и войн, которые сокращали численность населения, что несколько «освобождало» Землю, позволяя выжившим улучшить свой рацион и рожать больше детей, пока ряды не окажутся восполненными, а условия Мальтуса — восстановленными. Также благодаря социальному неравенству тонкая прослойка элит имела в своем распоряжении доход выше среднего (за счет чего-то, что понижало общий размер населения, которое в противном случае могло бы быть стабильным). Получается, что по мальтузианским условиям все то, что в обычном понимании представляется злейшим врагом благополучия человека: засухи, эпидемии, убийства и социальное неравенство — являются главными его благодетелями, поскольку только они способны время от времени поднять средний уровень жизни чуть выше прожиточного минимума. И это нормальное положение дел в течение большей части нашего пребывания на этой планете. Вызывает грустные и тревожные мысли.

Из истории видно, что, невзирая на местные колебания, происходит постоянное увеличение темпов экономического роста, подпитываемое накоплением технологических инноваций. Соразмерно росту мировой экономики растет и население планеты. (Похоже, растущее население само ускоряет темпы роста, возможно, за счет повышения коллективного интеллекта человечества[11].) Однако после промышленной революции экономический рост так ускорился, что рост населения Земли перестал ему соответствовать. Вследствие этого начал увеличиваться средний доход, вначале в странах Западной Европы, где индустриализация прошла раньше, а затем и во всем остальном мире. Сегодня даже в беднейших странах средний доход заметно превышает прожиточный минимум, что подтверждается ростом населения этих стран.

В наши дни самые высокие темпы роста населения наблюдаются как раз в беднейших странах, поскольку они еще не завершили «демографический переход» к режиму низкой фертильности, который наблюдается в более развитых обществах. Демографы прогнозируют, что к середине века число жителей планеты вырастет до девяти миллиардов, но после этого рост остановится или даже начнется спад, когда беднейшие страны присоединятся к развитому миру с его низким фертильным режимом[12]. Во многих богатых странах коэффициент фертильности находится ниже уровня воспроизводства, причем часто — гораздо ниже[13].

При этом можно ожидать, что в долгосрочной перспективе технологическое развитие и экономическое благополучие приведут к возвращению в исторически и экологически нормальное состояние, при котором у населения планеты снова начнется жизнь впритык в отведенной ему нише. Если это кажется парадоксальным в свете отрицательной связи между богатством и рождаемостью, которую мы сейчас наблюдаем в мировом масштабе, нужно напомнить себе, что современная эпоха — очень короткий эпизод в истории человечества, по сути, аберрация. Поведение людей не приспособлено к современным условиям. Мы не только не пользуемся очевидными способами повысить свою совокупную приспособленность (такими, например, как донорство сперматозоидов и яйцеклеток), но еще и активно подавляем фертильность, используя контроль над рождаемостью. С точки зрения эволюционной приспособленности здорового сексуального влечения достаточно для совершения полового акта таким способом, который позволяет максимизировать репродуктивный потенциал; однако в современных условиях большое преимущество с точки зрения естественного отбора давало бы более выраженное желание стать биологическим родителем как можно большего количества детей. В наше время это желание подавляется, как и другие черты, стимулирующие нашу склонность к продолжению рода. Однако культурное приспособление может навредить биологической эволюции. В некоторых сообществах, например гуттеритов или сторонников христианского движения Quiverfull, сложилась наталистская культура поощрения больших семей, и, как следствие, они быстро растут.

Рост населения и инвестиции

Если представить, что макроэкономические условия волшебным образом застыли в их современном состоянии, то будущее будут определять

культурные и этнические группы, в которых поддерживается высокий уровень рождаемости. Если в наше время большинство людей имели бы предпочтения, максимизировавшие приспособленность, население планеты могло бы легко удваиваться в каждом поколении. Если не проводилась бы политика ограничения рождаемости — а она становится все более суровой в своем противодействии тем, кто пытается ее обойти, — количество жителей нашей планеты могло бы расти в геометрической прогрессии, пока не столкнулось бы с такими проблемами, как нехватка земли и истощение простых инновационных возможностей. Это сделало бы невозможным продолжение экономического роста: средний доход начнет падать, пока не достигнет такого уровня, на котором бедность не позволит большинству людей иметь больше двух детей, которых они в состоянии вырастить. Так вновь захлопнулась бы мальтузианская ловушка, положив конец нашей эскападе в страну грез и снова, словно жестокий рабовладелец, заковав нас в цепи и заставив из последних сил бороться за выживание.

Из-за взрывного развития искусственного интеллекта, казалось бы, долгосрочный прогноз быстро перестанет быть столь долгосрочным. Программное обеспечение, как мы знаем, легко копируется, поэтому начнут стремительно появляться популяции имитационных моделей мозга или систем ИИ — буквально за минуты, а не десятилетия и века, — что совершенно истощит земные аппаратные ресурсы.

Защититься от тотального наступления мальтузианских условий, скорее всего, поможет частная собственность. Рассмотрим простую модель, в которой некие кланы (или закрытые сообщества, или страны) начинают с различного количества собственности и независимо друг от друга принимают различные стратегические решения относительно рождаемости и инвестиций. Некоторые кланы не думают о будущем и растрачивают свое состояние, в результате чего их обедневшие члены вливаются в ряды мирового пролетариата (или умирают, если не в состоянии прокормить себя). Другие кланы инвестируют часть своих финансовых средств и принимают политику неограниченной рождаемости — в результате они становятся все более многолюдными, пока не складываются мальтузианские условия, при которых население беднеет настолько, что смертность почти совпадает с рождаемостью, и с этого момента рост населения сравнивается с ростом имеющихся в его распоряжении ресурсов. При этом другие кланы могут ограничить у себя рождаемость до уровня, не превышающего темпы роста своего

капитала, такие кланы могли бы медленно увеличивать количество своих членов, притом что их доходы на душу населения также росли бы.

Если богатство перераспределяется от состоятельных кланов к быстро размножающимся и быстро растрачивающим свои ресурсы (чьи отпрыски, хотя и не по своей вине, появились в мире, где недостаточно капитала для выживания и процветания), тогда складываются классические мальтузианские условия. В предельном случае все члены всех кланов получают доход на уровне прожиточного минимума и все равны в своей бедности.

Если собственность не перераспределяется, у предусмотрительных кланов может сохраняться некоторая часть капитала, и их богатство способно расти в абсолютном выражении. Однако неясно, в состоянии ли люди обеспечить такую же доходность своего капитала, как машинный интеллект, поскольку может возникать синергия между трудом и капиталом: агент, имеющий и то и другое (например, предприниматель или инвестор, обладающий высоким интеллектом и большим состоянием), получит более высокий доход от своих вложений, чем в среднем по рынку получат агенты, обладающие сравнимыми финансовыми, но не интеллектуальными ресурсами. Поскольку люди будут уступать в интеллекте машинам, их капитал станет расти медленнее — если, конечно, не удастся окончательно решить проблему контроля, потому что тогда доходность человеческого капитала сравняется с доходностью машинного, ведь принципал-человек может поручить агенту-машине управлять своими сбережениями, причем делать это бесплатно и без конфликта интересов; но и в этом случае доля экономики, которой владеют машины, будет асимптотически приближаться к ста процентам.

Однако сценарий, в котором доля экономики, принадлежащей машинам, асимптотически приближается к ста процентам, не единственный вариант снижения человеческого влияния. Если экономика растет быстрыми темпами, тогда даже ее доля, уменьшающаяся относительно, может увеличиваться в абсолютном выражении. Это сравнительно неплохая новость для человечества: в многополярном сценарии с защищенными правами частной собственности общая величина богатства, принадлежащего людям, может расти даже в случае их неспособности решить проблему контроля. Конечно, данный эффект никак не устраняет возможность роста населения до такого уровня, на котором доход на душу упадет до прожиточного минимума, а также возможность обнищания людей, сбрасывающих будущее со счетов.

В долгосрочной перспективе в экономике начнут сильнее доминировать кланы с наиболее высокой нормой сбережений — скряги, владеющие половиной города и живущие под мостом. Наиболее процветающие из них станут проедать свои сбережения только том в случае, когда все возможности для инвестиций будут исчерпаны[14]. Однако если защита прав собственности окажется неидеальной — например, когда наиболее эффективные машины смогут всеми правдами и неправдами скопить у себя ресурсы, принадлежавшие людям, — тогда капиталистам-людям, возможно, придется тратить свое состояние гораздо быстрее, пока оно совсем не растает в результате действий машин (или вследствие расходов на защиту от них). Если все это будет происходить с цифровой, а не биологической, скоростью, то люди оглянуться не успеют, как уже окажутся лишенными собственности[15].

Жизнь в цифровом мире

В переходный период образ жизни человека, живущего при мальтузианских условиях, не обязательно будет походить на одну из знакомых нам моделей (скажем, на образ жизни охотника, собирателя, фермера или офисного работника). Скорее всего, большинство людей будут влачить жалкое существование наподобие бездельника-рантье, которому сбережений едва хватает на жизнь впроголодь[16]. Люди будут жить очень бедно, фактически на одни проценты или государственные пособия. Но при этом они будут жить в мире инновационных технологий — в мире не только сверхразумных машин, но и препаратов против старения и препаратов, доставляющих удовольствие; в мире виртуальной реальности и различных техник самосовершенствования. И вряд ли все это будет доступно большинству. Скорее всего, реальную медицину заменят лекарства для остановки роста и замедления метаболизма с целью экономии, поскольку для массы людей активная жизнь окажется невозможной (если учитывать постоянное снижение их и так минимальных доходов). По мере роста населения и снижения доходов люди могут регрессировать до состояния, минимально удовлетворяющего требованиям для выплаты пенсии, — возможно, это будет мозг с едва брезжущим сознанием, погруженный в контейнер и подключенный к снабжению кислородом и питательными жидкостями, который обслуживают машины и который способен накопить немного денег на воспроизводство путем клонирования себя специальным роботом-техником[17].

Еще бóльшая бережливость будет обеспечиваться за счет моделирования мозга, поскольку физически оптимизированный вычислительный субстрат, созданный сверхразумом, может оказаться эффективнее, чем биологический мозг. Однако миграция в цифровую реальность будет замедляться тем, что имитационные модели не смогут считаться людьми или гражданами и соответственно не получат право на пенсию или на не облагаемый налогами сберегательный счет. В этом случае ниша для людей сохранится, наряду со все более крупной популяцией имитационных моделей и систем искусственного интеллекта.

Пока что все внимание было сосредоточено на судьбе наших потомков, чью жизнь могут поддерживать сбережения, пособия или заработная плата, получаемая за счет тех, кто нанимает работников-людей. Теперь переключим внимание на те сущности, которые мы до сих пор относили к капиталу: на машины, всегда принадлежавшие людям, — машины, сконструированные с целью выполнять те или иные функции и способные заменить человека в очень широком диапазоне задач. Каким будет положение этих рабочих лошадок новой экономики?

Обсуждать было бы нечего, если все эти машины остались бы автоматами, простыми устройствами вроде парового двигателя или часового механизма — такого добра в постпереходной экономике будет много, но, похоже, вряд ли кто-то заинтересуется этим бездушным набором комплектующих. Однако если у машин будет сознание — если они будут сконструированы так, что смогут осознавать свою исключительность (или им по иным причинам будет приписан моральный статус), — тогда важно включать их в мировую систему. Благополучие работников-машин окажется наиболее важным аспектом постпереходного периода, поскольку они будут доминировать количественно.

Добровольное рабство, случайная смерть

Первый напрашивающийся вопрос: работниками-машинами будут владеть как капиталом (рабами) или их станут нанимать за заработную плату? Однако при ближайшем рассмотрении возникают сомнения, что от ответа будет что-то зависеть. На то есть две причины. Во-первых, если свободный работник в мальтузианских условиях получает зарплату на уровне прожиточного минимума, в его распоряжении не остается средств после оплаты питания и других базовых потребностей. Если работник является рабом,

то за его содержание платит хозяин, и все равно у раба не остается свободных средств. В обоих случаях работник получает лишь самое необходимое и ничего сверх того. Во-вторых, предположим, что свободный работник смог каким-то образом обеспечить себе доход выше прожиточного минимума (возможно, благодаря благоприятной системе регулирования). Как он потратит эту прибавку? Для инвесторов самым выгодным было бы создать виртуальных работников-«рабов», готовых трудиться за зарплату на уровне прожиточного минимума. Сделать это можно было бы, копируя тех работников, которые уже согласились на такие условия. Путем соответствующего отбора (и, возможно, некоторого изменения кода) инвесторы могли бы создать работников, которые не только предпочтут трудиться добровольно, но и решат пожертвовать своим работодателям все дополнительные доходы, если такие вдруг появятся. Однако после передачи денег работнику они по кругу вернутся к его владельцу или работодателю, даже если работник является свободным агентом, наделенным всеми юридическими правами.

Возможно, кто-то, возражая, заметит, насколько трудно создать машину, согласную добровольно выполнять любую работу или жертвующую свою зарплату своему же владельцу. Но у имитационных моделей должны быть особенно близкие людям мотивы. Обратите внимание, что если первоначальная проблема контроля, которую мы рассматривали в предыдущих главах, казалась трудновыполнимой, то сейчас мы говорим об условиях *переходного периода* — когда, видимо, методы выбора мотивации будут доведены до совершенства. Тем более если речь идет об имитационных моделях, то можно было бы добиться многого, просто *отбирая* нужные человеческие характеры. Наверное, проблема контроля будет в принципе упрощена, если предположить, что новый машинный интеллект включится в стабильную социоэкономическую матрицу, уже населенную другими законопослушными сверхразумными агентами.

Поэтому предлагаю остановиться на бедственном положении машин-работников, независимо от того, являются ли они рабами или свободными агентами. Вначале поговорим об эмуляторах, поскольку их представить легче всего.

Чтобы в мире появился новый работник-человек со своим профессиональным опытом и необходимыми навыками, потребуется от пятнадцати до тридцати лет. В течение всего этого времени человека нужно кормить, воспитывать, обучать, ему понадобится кров — все это большие расходы.

Напротив, создать новую копию цифрового работника так же легко, как загрузить очередную программу в оперативную память. То есть жизнь становится дешевле. Компания может постоянно подстраивать свою рабочую силу под меняющиеся требования за счет создания новых копий — и уничтожения старых, чтобы освободить компьютерные ресурсы. Это может привести к чрезвычайно высокой смертности среди работников-машин. Жизнь многих из них будет ограничена одним субъективным днем.

Могут быть и другие причины помимо колебаний спроса, по которым работодатели или владельцы эмуляторов захотят часто убивать (отключать) своих работников[18]. Если для нормального функционирования эмулятору мозга, как и биологическому мозгу, требуются периоды отдыха и сна, может быть дешевле стирать изнуренную имитацию в конце дня и заменять ее записанным состоянием свежей и отдохнувшей имитации. Поскольку такая процедура приводила бы к ретроградной амнезии всего выученного за день, эмуляторы, занятые выполнением задач, которые требуют формирования длительных когнитивных цепочек, смогут избежать частых стираний. Трудно писать книгу, если каждое утро, садясь за стол, не помнишь ничего из созданного накануне. Но агентов, выполняющих не столь интеллектуальные виды работ, вполне можно перезапускать, и делать это довольно часто — от единожды обученного продавца или сотрудника, обслуживающего клиентов, потребуется «удерживать» нужную информацию не более двадцати минут.

Поскольку перезапуски не позволяют формироваться памяти и навыкам, некоторые эмуляторы могут быть помещены в специальную обучающую среду, в которой они будут пребывать непрерывно, в том числе в периоды отдыха и сна, даже если их работа и не требует длинных когнитивных цепочек. При таких оптимальных условиях могли бы работать в течение долгих лет некоторые агенты по обслуживанию клиентов — причем при поддержке тренеров и экспертов по оценке производительности. Лучших учеников можно было бы использовать в качестве «племенных жеребцов», то есть по их шаблону каждый день штамповали бы миллионы свежих копий. В эти шаблоны есть смысл вкладывать большие средств, поскольку даже небольшое приумножение их продуктивности обеспечивало бы заметный экономический эффект, будучи растиражированным миллионы раз.

Параллельно с задачей обучения работников-шаблонов выполнению определенных функций огромные усилия будут прилагаться с целью

совершенствования технологии их эмуляции. Успехи в этом направлении были бы даже более ценными, чем успехи в обучении индивидуальных работников-шаблонов, поскольку улучшение технологии эмуляции применимо ко всем работникам-эмуляторам (и, потенциально, к другим имитационным моделям тоже), а не только к занятым в одной определенной области. Можно направить огромные ресурсы на поиск вычислительных коротких путей, позволяющих создавать эмуляторы более эффективно, а также разрабатывать нейроморфные и полностью синтетические архитектуры ИИ. Эти исследования, вероятно, проводились бы тоже эмуляторами, запущенными на очень быстрой аппаратной основе. В зависимости от стоимости вычислительной мощности могли бы круглосуточно работать миллионы, миллиарды или триллионы имитационных моделей мозга самых проницательных исследователей-людей (или их улученных версий), раздвигая границы машинного интеллекта; некоторые из них могли бы действовать на порядки быстрее, чем биологический мозг[19]. Это весомая причина полагать, что эра человекоподобных эмуляторов будет короткой — *очень* короткой по звездному времени — и что ей придет на смену неизмеримо превосходящий их искусственный интеллект.

Мы уже перечислили несколько причин, по которым работодатели эмуляторов могут периодически выбраковывать свои стада: колебания спроса на работников различного вида деятельности; экономия на времени отдыха и сна; появление новых усовершенствованных шаблонов. Еще одной причиной могут быть соображения безопасности. Чтобы имитации-работники не вынашивали враждебные планы и не плели заговоры, эмуляторы, занятые на особенно важных позициях, могли бы запускаться на ограниченное время с частым сбросом к исходному состоянию готовности[20].

Эти исходные состояния, к которым будут возвращать настройки эмуляторов, следует очень тщательно готовить и перепроверять. Типичный эмулятор с коротким жизненным циклом, которого оптимизировали с точки зрения его лояльности и производительности, мог бы чувствовать себя на следующее утро просто хорошо отдохнувшим. Он помнил бы, что после многих (субъективных) лет интенсивного обучения и отбора стал лучшим среди своих однокашников, что только что набрался сил в отпуске, выспался, прослушал воодушевляющую побудительную речь и бодрую музыку, и теперь ему не терпится сделать максимум возможного для своего работодателя.

Его мало беспокоят мысли о неотвратимой смерти в конце рабочего дня. Эмуляторы, страдающие страхом смерти и прочими неврозами, менее продуктивны и потому не могут быть отобраны в качестве шаблона[21].

В высшей степени тяжелый труд как высшая степень счастья

Оценивая желательность подобных гипотетических ситуаций, важно принимать во внимание гедонистическое состояние среднего эмулятора[22]. Страдать или наслаждаться будет типичный работник-эмулятор, выполняя свою трудную работу?

Нам не следует поддаваться соблазну проецировать собственные чувства и ощущения на воображаемого цифрового работника. Вопрос не в том, были бы *вы* счастливы, если вам пришлось бы постоянно трудиться и не иметь возможности видеться со своими любимыми, — это ужасная судьба, спору нет.

Несколько правильнее было бы взять за основу гедонистический опыт современных людей в течение обычного рабочего дня. Во всем мире проводились исследования, в ходе которых респондентам задавали вопрос, насколько они счастливы, и большинство выбирало ответы от «довольно счастливы» до «очень счастливы» (средний балл 3,1 на шкале от 1 до 4)[23]. Исследования среднего эмоционального состояния, в которых у респондентов спрашивали, какие из положительных или отрицательных эмоций они испытывали недавно, как правило, дают аналогичные результаты (средний балл 0,52 по шкале от –1 до 1). Есть небольшая зависимость среднего субъективного благополучия от размера ВВП на душу населения страны[24]. Однако было бы ошибкой экстраполировать эти данные на гедонистическое состояние будущих работников-эмуляторов. Одна из причин заключается в том, что их условия будут совершенно другими: с одной стороны, они могут работать гораздо больше; с другой, они будут свободны от болезней, мышечной боли, голода, неприятных запахов и многого другого. Хотя такие соображения не говорят о главном. Гораздо важнее, что их ощущение наслаждения можно легко корректировать при помощи цифрового эквивалента лекарств или нейрохирургии. Это значит, что было бы ошибкой делать выводы о гедонистическом состоянии будущих эмуляторов на основании внешних условий их жизни, тем более представляя себя на их месте. Гедонистическое состояние — вопрос сознательного выбора. В модели, которую мы сейчас обсуждаем, этот выбор за своих работников делает владелец капитала,

стремящийся максимизировать доходность своих инвестиций в тружеников-эмуляторов. Соответственно вопрос, насколько счастливы они будут, сводится к выяснению, какие гедонистические состояния наиболее продуктивны (для решения различных задач, которые будут поставлены перед этими имитационными моделями).

И снова прошу воздержаться от поспешных выводов. Если оказывается, что независимо от эпох, географии и рода занятий человек обычно бывает в меру счастлив, это может склонить в пользу того, что такое же положение сохранится и в постпереходный период. То есть речь не о том, что раз человеческий мозг предрасположен к восприятию счастья, то, вероятно, он будет «удовлетворен» и в новых условиях; скорее, мы имеем в виду, что раз определенный уровень счастья оказался достижимым для человеческого мозга в прошлом, то аналогичный уровень счастья, возможно, будет доступен имитационным моделям человеческого мозга в будущем. Хотя эта формулировка также показывает слабость умозаключения, а именно: психические состояния, характерные для гоминид, занимающихся охотой и собирательством в африканской саванне, не обязательно пригодны для модифицированных версий человеческого мозга, живущих в условиях постпереходной виртуальной реальности. Конечно, мы можем *надеяться*, что будущие работники-эмуляторы будут столь же счастливы или даже счастливее, чем работники-люди на протяжении всей своей человеческой истории; но нам придется поискать более убедительные причины для подтверждения этого предположения (в нашем сценарии мультиполярного мира, выстраиваемого по принципу laissez-faire*).

Предположим, что причина превалирования состояния счастья среди людей (в той степени, в которой оно преобладает) состоит в том, что эта положительная эмоция служила сигнальной функцией в условиях эволюционного приспособления. Создание у других членов социальной группы впечатления о собственном процветании: *я здоров, отлично лажу с окружающими и уверен в своем благополучном будущем* — могло повышать популярность индивидуума. Поэтому склонность к жизнерадостности могла быть критерием отбора, в результате чего биохимия мозга современного

---

* *Laissez-faire* (фр. букв. «не мешать; оставаться пассивным») — принцип невмешательства; экономическая доктрина, согласно которой государственное вмешательство в экономику должно быть минимальным.

человека оказалась смещена в пользу более позитивного восприятия мира, чем то, которое было бы максимально эффективным в соответствии с более простым материалистическим критерием. Если это было бы так, то будущая *joie de vivre** может зависеть от того, сохранит ли в постпереходном мире эмоция радости свою сигнальную функцию, играющую социальную роль. (К этому вопросу мы вскоре вернемся.)

Тратит ли человек радостный больше энергии, чем человек угрюмый? Может быть, счастливые люди больше склонны к творческим порывам и полетам фантазии — поведение, которое вряд ли будет приветствоваться будущими работодателями. Возможно, угрюмая и тревожная сосредоточенность на безошибочном выполнении конкретного задания станет самым желанным образом действий для большинства видов трудовой деятельности. Мы не считаем, что это так, мы лишь говорим, что не знаем, что это не так. И все же нам следует подумать, насколько плоха может быть ситуация, если одна из таких пессимистичных гипотез относительно будущих мальтузианских условий окажется истинной: не только из-за «цены выбора» в создании чего-то лучшего и даже грандиозного, но также потому, что это состояние может быть плохим само по себе, возможно, гораздо более плохим, чем мальтузианские условия.

Мы редко работаем в полную силу. Но когда мы так делаем, чувствуем, как это болезненно. Представьте, что вы бежите по беговой дорожке с уклоном вверх: сердце колотится, мышцы болят, легким не хватает воздуха. Короткий взгляд на таймер: следующий отдых ожидается через 49 лет, 3 месяца, 20 дней, 4 часа, 56 минут и 12 секунд — это же и момент вашей смерти. Впору пожалеть, что родился.

И снова речь не о том, что все будет именно так, — проблема в отсутствии уверенности, что все будет иначе. Конечно, можно представить и более оптимистичную картину. Например, нет очевидной причины, по которой эмуляторы должны страдать от физического недомогания и болезней, поскольку они будут лишены уязвимой физической оболочки — что было бы большим достижением по сравнению с нынешним положением дел. Более того, поскольку виртуальная реальность довольно дешевая, эмуляторы могут работать в роскошных условиях: в прекрасных дворцах на вершинах гор, на террасах, выходящих в цветущий весенний сад, на солнечных пляжах,

---

* *Joie de vivre* (фр.) — радость жизни.

омываемых лазурным океаном, — с правильными освещением и температурным режимом, пейзажем и декором; их не побеспокоят неприятные запахи, шумы, сила тяжести и жужжащие насекомые; они будут облачены в комфортную одежду, ощущать ясность мысли и сосредоточенность внимания, хорошо питаться. Еще более важно, что если у людей оптимальным психическим состоянием для максимальной продуктивности в большинстве видов работ является радостная жажда деятельности — видимо, так оно и есть, — то цифровая эпоха имитационных моделей может оказаться похожей на рай.

В любом случае было бы ценно организовать все так, чтобы у кого-то или чего-то была возможность вмешаться и все исправить в случае, если выбранная траектория окажется ведущей в антиутопию. Также было бы желательно оставить что-то вроде спасательного люка, который позволял бы сбежать в смерть и забвение в случае, если качество жизни постоянно оказывается ниже того уровня, когда исчезновение становится предпочтительнее продолжения существования.

Аутсорсеры, лишенные сознания?

В долгосрочной перспективе, когда эпоху эмуляторов сменит эпоха искусственного интеллекта (или если машинный интеллект появится сразу, минуя стадию имитационной модели головного мозга), в сценарии многополярного мира страдания и удовольствия могут исчезнуть совсем, поскольку гедонистический механизм вознаграждения окажется не самой эффективной системой стимуляции сложного искусственного агента, который, в отличие от человеческого мозга, не обременен наследием нервной системы животных. Возможно, более совершенная система мотивации будет основана на явном выражении функции полезности или на какой-то иной архитектуре, в которой отсутствуют прямые функциональные аналоги удовольствия и страданий.

Близким, но несколько более радикальным вариантом многополярного мира — в принципе, лишенного привычных нам систем ценностей — является такой сценарий, по которому мировой пролетариат полностью будет лишен сознания. Вероятность этого особенно велика в случае ИИ, который может быть структурирован совершенно не так, как человеческий интеллект. Но даже если машинный интеллект вначале возникнет в результате полной эмуляции головного мозга, дав старт появлению обладающих сознанием

цифровых агентов, сохранение конкуренции в постпереходной экономике вполне способно привести к возникновению прогрессивных и менее нейроморфных форм машинного интеллекта — или в результате создания синтетического ИИ с нуля, или в результате последовательных модификаций и улучшений имитационных моделей, причем в процессе совершенствования эмуляторы начнут все больше терять человеческие характеристики.

Рассмотрим сценарий, по которому после появления технологии полной эмуляции головного мозга продолжающийся прогресс в нейробиологии и кибернетике (подстегиваемый наличием эмуляторов, работающих как в качестве исследователей, так и объектов экспериментов) сделал бы возможным изоляцию эмулированных индивидуальных когнитивных модулей и подключение их к аналогичным модулям, изолированным от других эмуляторов. Чтобы такие стандартные модули могли эффективно сотрудничать, им потребуется период обучения и притирки, после которого они будут в состоянии взаимодействовать быстрее. Это повысит их продуктивность и стимулирует дальнейшую стандартизацию.

После этого эмуляторы смогут начать передавать на аутсорсинг большую часть своего функционала. Зачем учиться складывать и умножать, если можно переслать задачу, требующую знания математики, в компанию Gauss Modules, Inc.? Зачем уметь говорить, если можно привлечь к переводу своих мыслей в слова агентство Coleridge Conversation? Зачем принимать решения о своей личной жизни, если существуют сертифицированные исполнительные модули, способные сканировать вашу систему ценностей и управлять вашими ресурсами так, что ваши замыслы удастся воплотить быстрее, чем если бы вы занимались этим сами? Некоторые эмуляторы могут предпочесть сохранить за собой большую часть собственного функционала и выполнять самостоятельно даже те задачи, с которыми более эффективно справились бы другие. Они будут похожи на любителей, с удовольствием разводящих овощи или вяжущих свитеры. Но эти эмуляторы-дилетанты так и не станут профессионалами, поэтому, если рабочий поток будет перенаправлен от малоэффективных игроков к более эффективным, — такие имитационные модели непременно окажутся в проигрыше.

Таким образом, бульонные кубики отдельных человекоподобных умов растворятся и превратятся в однородное алгоритмическое варево.

Можно предположить, что оптимальная эффективность будет обеспечена за счет группировки модулей, отвечающих за различные способности,

в структуры, отдаленно напоминающие систему когнитивных функций человеческого мозга. Вполне возможно, например, что математический модуль должен быть связан с языковым, и оба они — с исполнительным, чтобы все три могли работать вместе. Тогда когнитивный аутсорсинг окажется практически бесполезным. Но пока тому нет убедительных подтверждений, мы должны считать, что человекоподобная когнитивная архитектура оптимальна только внутри ограничений, связанных именно с человеческой неврологией (а может быть, и вообще не оптимальна). Когда появятся перспективные архитектуры, которые не могут быть хорошо реализованы на биологических нейронных сетях, возникнет необходимость в качественно новых решениях, и наиболее удачные из них уже почти не будут напоминать знакомые нам типы психики. Тогда человекоподобные когнитивные схемы начнут терять свою конкурентоспособность в новых экономических и экосистемных условиях постпереходной эпохи[25].

То есть вполне могут существовать свои ниши для разумных систем: менее сложных — таких как индивидуальные модули; более сложных — таких как огромные кластеры модулей; сравнимых по сложности с человеческим мозгом, но с радикально иной архитектурой. Будут ли таким системам присущи хоть какие-то ценности? Следует ли нам приветствовать рождение мира, в котором эти чужеродные системы заменят человеческий мозг?

Ответ на этот вопрос может зависеть от конкретной природы таких систем. В современном мире существует много уровней организации. Есть очень сложные структуры высокого уровня, например государства и транснациональные корпорации, состоящие из множества людей, но за этими структурами мы обычно признаем лишь инструментальную ценность. Государства и корпорации (как принято считать) не обладают сознанием сверх сознания людей, составляющих их: они не в состоянии почувствовать страдание, удовольствие или испытать какие-либо иные чувства. Мы ценим их лишь постольку, поскольку они обслуживают потребности людей, а когда они не справляются со своей задачей, мы «убиваем» их без малейшего раскаяния. Есть структуры более низкого уровня, также обычно лишенные морального статуса. Мы не видим никакого вреда в том, чтобы удалить приложение из смартфона, мы не считаем, будто нейрохирург вредит человеку, страдающему эпилепсией, когда удаляет у него неправильно функционирующий отдел мозга. Большинство людей призна́ет моральный статус

сложных систем уровня человеческого мозга только в случае, если они будут способны получать сознательный опыт[26].

В крайнем случае можно представить высокоразвитое с технологической точки зрения общество, состоящее из множества сложных систем, в том числе гораздо более сложных и интеллектуальных, чем все, что существует на планете сегодня, — общество, совершенно лишенное кого-либо, кто обладал бы сознанием или чье благополучие имело бы какое-либо моральное значение. В некотором смысле это было бы необитаемое общество. Общество экономических и технологических чудес, никому не приносящих пользы. Диснейленд без детей.

Эволюция — путь наверх или не обязательно?

Слово *эволюция* часто используется в качестве синонима слова *прогресс*, что, возможно, отражает общепринятое некритическое восприятие эволюции как доброй силы. Необоснованная вера в изначальную благотворность эволюционного процесса может уступить место беспристрастной оценке целесообразности многополярного исхода, в котором будущее разумной жизни определяется конкурентной динамикой. Любая такая оценка должна быть основана на некотором (хотя бы подразумеваемом) представлении о распределении вероятности различных фенотипов, которые смогут приспособиться к цифровой жизни постпереходного периода. Даже в самых благоприятных обстоятельствах было бы трудно извлечь четкий и правильный ответ из опутывающей эту тему паутины неизбежной неопределенности — тем более если мы еще внесем свой вклад и добавим изрядное количество наивного оптимизма.

Возможным источником веры в поступательное движение эволюции является восходящая динамика эволюционного процесса в прошлом. Начав с простейших делящихся клеток, эволюция порождала все более «развитые» организмы, в том числе существа, наделенные мозгом, сознанием, языком и мышлением. Позднее именно те культурные и технологические процессы, которые имеют отдаленное сходство с эволюционными, позволили человечеству развиваться ускоренными темпами. И на геологической, и на исторической шкале превалирует тенденция к повышению уровня сложности, знаний, сознания и скоординированной организации в достижении целей — тенденция, которую, если не стремиться к чрезмерной точности определений, можно было бы назвать прогрессом[27].

Во-первых, в нашем представлении эволюция как процесс, стабильно приносящий благо, с трудом совмещается со страданием, которое мы видим как в мире людей, так и в мире природы. Кто приветствует достижения эволюции, делают это скорее с эстетической, чем с этической стороны. Хотя в первую очередь нас должен волновать вопрос не о том будущем, которое мы для себя откроем в новом научно-фантастический романе или фильме, а о будущем, в котором всем нам было бы хорошо жить, — между тем и другим лежит огромная разница.

Во-вторых, у нас нет причин думать, что даже прогресс, имевший место в прошлом, был неизбежен. Многое можно отнести на счет удачи. Это соображение основано на том, что благодаря эффекту выбора наблюдателя нам доступны лишь свидетельства об успешном ходе собственной эволюции[28].

Предположим, что с вероятностью 99,9999 процента на всех планетах, где возникла жизнь, она погибла прежде, чем разумный наблюдатель мог начать размышлять о своем происхождении. Если это так, что мы могли бы ожидать увидеть? Вероятно, ровно то же самое, что видим сейчас. Гипотеза, что шансы разумной жизни появиться на той или иной планете малы, как раз и означает, что мы скорее окажемся не там, где жизнь существует лишь на начальной стадии, а там, где уже появилась разумная жизнь, даже если такие планеты представляют собой очень небольшую долю всех планет, на которых зародилась примитивная жизнь. Поэтому долгая история жизни на Земле не может служить надежным подкреплением, что была большая доля вероятности — не говоря уже о неизбежности — появления на нашей планете высокоразвитых организмов[29].

В-третьих, нет никаких гарантий, что эти идеи мелиоризма* собираются вместе с нами шагнуть в наше будущее — даже если нынешние условия были бы идеальными, даже если можно было бы показать неизбежность их возникновения из некоторых универсальных исходных условий. Это справедливо даже в том случае, если мы исключим катаклизмы, способные привести к гибели человечества, и даже в том случае, если предположим, что в результате эволюционного развития будут продолжать появляться все более и более сложные системы.

---

* *Мелиоризм* (от лат. *melior* — «лучше») — концепция, признающая реальность идеи прогресса как ведущей к совершенствованию мира.

Ранее мы предположили, что интеллектуальные работники-машины, отобранные по критерию максимальной продуктивности, будут трудиться чрезвычайно много, но неизвестно, смогут ли они при этом чувствовать себя счастливыми. Мы даже допустили, что некоторые из них могут вообще не иметь сознания, но именно они будут лучше всех приспособлены к конкурентной гонке будущей цифровой жизни. После полной потери восприятия, а потом утраты сознания, они лишатся всех прочих качеств, которые большинством людей считаются необходимыми для нормальной жизни. Современный человек ценит музыку, шутку, любовь, искусство, игры, танцы, разговоры, философию, литературу, приключения, путешествия, еду, выпивку, дружбу, детей, спорт, природу, традиции, духовные ценности и множество других вещей. Нет никакой гарантии, что хоть один пункт из перечисленного останется в жизни тех, кто приспособится к новым условиям. Наверное, максимально увеличить шансы на выживание сможет лишь одно свойство — готовность безостановочно и интенсивно трудиться, выполняя скучную и монотонную работу, цель которой состоит в улучшении восьмого знака после запятой какого-то экономического показателя. В этом случае отобранные в ходе эволюции фенотипы будут лишены перечисленных выше качеств и в зависимости уже от вашей системы ценностей покажутся вам существами либо отвратительными, либо никчемными, либо всего лишь жалкими, — но в любом случае бесконечно далекими от той великой Утопии, которая все-таки стоит нашего доброго слова.

Невольно возникает вопрос: почему столь богатое прошлое и столь насыщенное настоящее — со всеми нашими чувствами, привязанностями, мыслями и пристрастиями — может породить столь убогое будущее? Как это согласовать друг с другом? И невольно возникает контрвопрос: если все человеческие увлечения и занятия действительно так «бессмысленны», то почему они сохранились и даже развились в ходе эволюционного процесса, сформировавшего наш вид? Современный человек с его эволюционным дисбалансом не может поставить их себе в вину: наши предки из эпохи плейстоцена тоже тратили на это свою энергию. Многие поведенческие стереотипы даже нельзя назвать уникальными, поскольку они свойственны не только *Homo sapiens*. Например, демонстративное поведение по типу «боевой раскраски» встречается в самом разном контексте: от брачных игр в животном мире до противостояния государств и народов[30].

Хотя в нашу задачу не входит подробное эволюционное объяснение таких моделей поведения, тем не менее считаю нужным заметить, что многие присущие им функции отпадут сами собой в будущем, при новых условиях существования искусственного интеллекта. Возьмем, например, игру — этот особый вид социального поведения, характерный лишь для некоторых видов живых организмов и распространенный в основном среди молодых особей, — во многом игра представляет способ приобретения личностных и социальных навыков, необходимых для полноценной жизни. Но в постпереходный период функции игрового поведения перестанут иметь сколь-нибудь значимый смысл, так как появится возможность или создавать сразу «взрослые» имитационные модели, уже владеющие некоторой суммой зрелых навыков, или импортировать непосредственно в систему одного ИИ знания и умения, достигнутые другим ИИ.

Многие образцы демонстративного поведения человека могли сложиться в ходе эволюционных попыток инсценировать иногда на первый взгляд неуловимые достоинства, такие как: физическая стойкость; психологическая и эмоциональная устойчивость; положение в обществе; чувство локтя; предприимчивость и находчивость; готовность и умение побеждать. Приведем классический случай, принадлежащий, правда, миру животных, — павлин со своим знаменитым хвостом. Только самец, абсолютно уверенный в собственном достоинстве, как правило физическом, может позволить себе иметь столь вызывающее оперение и умело пользоваться им, — самки хорошо усвоили этот эволюционный сигнал и считают такую экстравагантность крайне привлекательной. Поведенческие особенности в не меньшей степени, чем морфологические, способны сигнализировать о генетической приспособленности и иных социально значимых признаках[31].

Учитывая, что демонстративное поведение столь распространено и среди людей, и среди других биологических видов, возникает вопрос: не войдет ли оно и в репертуар усовершенствованных технологических форм существования. Допустим, в грядущем пространстве интеллектуальных информационных экосистем не останется места для проявления таких свойств, как веселость, игривость, музыкальность и даже сознание, — в смысле их практического применения. Но, может быть, эти качества все-таки окажутся полезными с эволюционной точки зрения и дадут некоторое преимущество их обладателям, которым будет проще сигнализировать о своей адаптивности к новым условиям?

Конечно, довольно трудно устанавливать заранее, насколько удастся в будущем согласовать сегодняшнюю систему ценностей с адаптивной системой цифровой экологии. Тем более что поводов для скептицизма хватает.

Во-первых, многие встречающиеся в природе стереотипы демонстративного поведения связаны с выбором полового партнера; причем эти проявления особенно яркие, а для человека еще и дорогостоящие[32]. Но практически бесполым технологическим формам существования вряд ли придется задумываться над проблемой полового отбора.

Во-вторых, у технологически оснащенных агентов отпадет необходимость в демонстративном поведении, поскольку в их распоряжении появятся новые надежные и незатратные способы передачи информации о себе. Даже сегодня профессиональные кредиторы при оценке платежеспособности предпочитают доверять документальным доказательствам, то есть свидетельствам о праве собственности и банковским выпискам, а не полагаться на роскошный внешний вид клиента — костюм из последней коллекции модного дизайнера и часы Rolex. В будущем для получения нужных данных смогут привлекать специальные аудиторские фирмы, которые станут выдавать своим клиентам справки о наличии тех или иных качеств агента на основе анализа его поведенческих характеристик, или изучения его действий в смоделированных условиях, или прямого считывания его исходного кода. Если агент даст согласие на проведение подобной проверки, то это будет достаточным свидетельством, что с требуемыми свойствами у него все в порядке, и даже более эффектным подтверждением, чем прямая их демонстрация. *Подделывать* признаки, выявленные опосредованным, но профессиональным путем, окажется слишком трудным и дорогостоящим занятием — что, между прочим, является основным доказательством их достоверности, — но в случае *подлинности* признаков будет намного легче и дешевле передавать их цифровым способом, а не демонстрировать естественным образом.

В-третьих, не все варианты демонстративного поведения имеют равную значимость и одинаково желательны с точки зрения общества. А некоторые просто абсурдны в своей расточительности. Например, публичное уничтожение большого количества накопленного имущества индейцами квакиутл во время демонстративного обмена дарами между племенами, когда происходило своего рода соревнование вождей за максимальное влияние и авторитет, только орудием борьбы выступало их богатство, — эта традиционная

церемония называлась «потлач»[33]. Современными аналогами потлача можно считать рекордно высокие небоскребы, очень крупные и непомерно дорогие яхты, а также попытки строительства ракет для полета на Луну. И если мы абсолютно разделяем мнение, что музыка и хорошая шутка повышают качество человеческой жизни, то вряд ли то же самое относится к приобретению безумно дорогих модных аксессуаров и прочих статусных предметов роскоши. Хуже всего, когда безрассудное демонстративное поведение наносит непоправимый вред: довольно часто неумеренная мужская бравада приводит к прямому насилию и бряцанию оружием. Даже если в будущем разумные формы существования и воспользуются «сигнальной системой» демонстративного поведения, то все равно остается открытым вопрос о ценности выражаемого достоинства. Будет ли оно как восхитительная соловьиная трель, или как односложное карканье вороны, или как непрерывный лай взбесившейся собаки?

А потом появится синглтон?

Даже если непосредственным результатом перехода к машинному интеллекту окажется многополярный мир, несколько полюсов силы никак не отменяют в дальнейшем появление синглтона. Так естественным образом реализуется очевидная и очень долго длящаяся тенденция к общемировой консолидации политических сил[34]. Как это будет происходить?

Второй переход

В процессе превращения первоначального многополярного уклада в синглтон после первого перехода должен произойти второй технологический переход, причем довольно резкий, чтобы у одной из сил, способной воспользоваться моментом и сформировать синглтон, появилось решающее конкурентное преимущество. Гипотетический второй переход может быть вызван мощным достижением в разработке более высокого уровня сверхразума. Например, если первая волна машинного сверхразума будет основана на имитационных моделях, то второй виток начнется, когда эмуляторы, занимающиеся дальнейшими исследованиями, добьются успеха в создании эффективного ИИ, способного к самосовершенствованию[35]. (Может быть, ко второму переходу приведет прорыв или в нанотехнологиях, или военных технологиях, или любых иных универсального назначения, которые мы сегодня даже не в состоянии вообразить.)

Скорость развития первого постпереходного периода, скорее всего, будет чрезвычайно высокой. Поэтому вполне вероятно, что к решающему стратегическому преимуществу, которое образуется у лидирующей силы, может привести даже небольшой разрыв между ней и ее ближайшим конкурентом. Предположим, первый переход был совершен двумя вырвавшимися вперед проектами с разницей в несколько дней, но поскольку сам взлет оказался слишком медленным, то этот небольшой разрыв не обеспечил лидирующему проекту решающего стратегического преимущества. Оба проекта обрели мощь сверхразума, но один из них на несколько дней раньше. Однако теперь исследования идут на временной шкале машинного сверхразума, возможно, в тысячи миллионов раз быстрее, чем когда их проводили ученые-люди. Появление технологии, способной привести ко второму переходу, видимо, произойдет в течение дней, часов или минут. Даже если у лидера был запас всего в несколько дней, этот прорыв способен сработать как катапульта и обеспечить ему абсолютное стратегическое преимущество. Однако обратите внимание: если диффузия технологий (за счет шпионажа или по иным каналам) ускорится так же, как их развитие, эффект молниеносности может сойти на нет. Тогда останется единственный важный фактор — крутизна второго перехода, то есть скорость, с которой он произойдет относительно обычного для постпереходного периода темпа событий. (В этом смысле чем быстрее все происходит после первого перехода, тем менее крутым будет второй.)

Также можно предположить большую вероятность, что в случае второго перехода (или одного из последующих) решающее стратегическое преимущество будет использовано для формирования синглтона. После первого перехода агенты, принимающие решения, будут или сами обладать сверхразумом, или пользоваться советами сверхразума, который проанализирует имеющиеся варианты стратегического выбора. Более того, ситуация, сложившаяся после первого перехода, может быть менее опасной для агрессора с точки зрения нанесения упреждающего удара по конкурентам. Если разум, принимающий решение после первого перехода, существует в цифровом виде, его легче скопировать и тем самым сделать менее уязвимым для контратак. Агрессора не сильно напугает удар возмездия со стороны обороняющихся сил и уничтожение девяти десятых его населения — ведь у него есть возможность немедленно восстановить баланс благодаря резервным копиям. Разрушение инфраструктуры (которая тоже может быть быстро восстановлена) тоже не станет критичным для цифрового разума с практически

неограниченным жизненным циклом, поскольку он наверняка запланировал и увеличение имеющихся ресурсов, и расширение влияния в космологическом масштабе времени.

Суперорганизмы и эффект масштаба

На размер согласованности институциональных механизмов людей, таких как корпорации или государства, влияет множество параметров: технологических, военных, финансовых и культурных, — которые могут различаться в разные исторические эпохи. Революция машинного интеллекта приведет к глубоким изменениям в этих параметрах. Возможно, такие изменения вызовут формирование сингltона. Хотя, не видя деталей этих потенциальных изменений, нельзя исключить и противоположное — что результатом таких изменений станет фрагментация, а не унификация. Тем не менее стоит заметить, что рост неопределенности, с которой мы здесь имеем дело, сам по себе может быть основанием для большей уверенности в вероятном возникновении сингltона. Революция машинного интеллекта может, так сказать, смешать нам карты и перетасовать колоду так, что сделает реальной геополитическую перестройку, которая иначе вряд ли была бы возможной.

Детальный анализ всех факторов, способных повлиять на масштаб интеграции политических сил, увел бы нас далеко за пределы выбранной темы — один обзор соответствующей политологической и экономической литературы легко потянет на отдельную книгу. Нам придется ограничиться коротким взглядом на несколько аспектов появления цифровых агентов, которые упростят задачу централизации контроля.

Карл Шульман утверждает, что естественный отбор в популяции имитационных моделей мозга приведет к появлению «суперорганизмов», то есть сообществ эмуляторов, готовых пожертвовать собой ради блага их клана[36]. Суперорганизмам не придется решать агентскую проблему, столь насущную в случае организаций, члены которых преследуют свои собственные интересы. Как клетки в наших организмах или отдельные особи в колониях общественных насекомых, эмуляторы будут абсолютно альтруистичными по отношению к своим братским копиям и смогут сотрудничать друг с другом даже в отсутствие специальных схем мотивации.

Суперорганизмы будут иметь особенно заметное преимущество, если в них есть возможность безвозвратного стирания (или приостановки выполнения) отдельных имитационных моделей без их согласия. Компании

и государства, состоящие из эмуляторов, которые настаивают на самосохранении, окажутся под постоянно растущим гнетом обязательств по выплате содержания морально устаревшим или избыточным членам. В отличие от них те организации, где эмуляторы с готовностью уничтожают себя, когда потребность в их услугах исчезает, смогут легче адаптироваться к колебаниям спроса и экспериментировать, множа вариации своих работников и сохраняя из них лишь наиболее продуктивных.

Если возможность принудительного уничтожения отключена, то у эусоциальных* эмуляторов будет меньше конкурентных преимуществ, хотя они и сохранятся. Работодатели готовых сотрудничать и принести себя в жертву имитационных моделей по-прежнему получат выигрыш в продуктивности в результате уменьшения остроты агентской проблемы в своей организации, включая отсутствие необходимости бороться с сопротивлением имитационных моделей, возражающих против своего уничтожения. В целом выигрыш в продуктивности за счет наличия работников, готовых пожертвовать своими жизнями ради общего блага, представляет собой особую выгоду организаций, члены которых фанатически им преданы. Такие члены не только готовы долгие часы трудиться за символическую зарплату и даже отправиться ради них в могилу — они никогда не станут плести интриги и постараются действовать исключительно в интересах организации, что уменьшает потребность в контроле и бюрократических ограничениях.

Если единственным способом добиться такой преданности является копирование одного «родителя» (когда все имитационные модели, или эмуляторы, принадлежащие конкретному суперорганизму, получены от единого шаблона), тогда недостатком такого суперорганизма будет более узкий набор навыков, чем у соперников, неясно лишь, перевесит ли этот недостаток преимущество в виде отсутствия внутренних агентских проблем[37]. Впрочем, частично этот недостаток можно нивелировать, если суперорганизм будет состоять из членов, прошедших обучение по разным программам. Даже если все эмуляторы получены от единого шаблона, в процессе функционирования они смогут вырабатывать отличные друг от друга совокупности навыков. Один шаблон всесторонне развитого эмулятора может породить несколько типов работников, если одна копия проходит обучение на бухгалтера, другая — на инженера и так далее. В результате члены суперорганизма будут

* *Эусоциальность* — наивысший уровень социальной организации животных.

обладать различными навыками, но одинаковыми талантами. (Для большего разнообразия может потребоваться больше одного шаблона.)

Неотъемлемым свойством суперорганизма будет не то, что он состоит из копий одного «родителя», а то, что все индивидуальные агенты внутри него полностью посвятили себя общей цели. То есть способность создавать суперорганизмы потребует лишь частичного решения проблемы контроля. Если исходить из того, что полное решение проблемы контроля предоставит возможность создать сверхразумного агента, способного добиться любой произвольной конечной цели, то частичное решение проблемы контроля, которого достаточно, чтобы сформировать суперорганизм, позволит лишь изготовить многочисленное количество агентов, наделенных одной и той же конечной целью (она может быть значимой, но не обязательно произвольной)[38].

Таким образом, обсуждение основной идеи этого раздела не ограничивается вопросом о группах моноклональных эмуляторов, поэтому она может быть сформулирована в более общем виде и применима к широкому диапазону сценариев возникновения машинного интеллекта в многополярном мире. А именно: определенные достижения в развитии методов выбора мотивации, которые окажутся выполнимыми в случае, если действующие лица представлены в цифровом виде, могут помочь преодолеть некоторую неконструктивность, свойственную нашим сегодняшним крупным организациям, и обеспечить рост эффективности, обусловленный ростом масштаба производства. Когда эти ограничения будут преодолены, организации — будь то компании, страны или иные экономические и политические институты — могут значительно увеличиться в размерах. Это один из факторов, облегчающих формирование постпереходного синглтона.

Сфера, в которой суперорганизмы (и другие цифровые агенты с частично выбранной мотивацией) способны быстро достичь нужного превосходства, — это аппарат принуждения. Государство может использовать методы выбора мотивации для формирования одинаково преданных ему полиции, армии, разведывательных служб и гражданской администрации. Как пишет Шульман:

> Сохраненные состояния [довольно лояльной имитационной модели, тщательно отобранной и проверенной] можно будет размножать миллиарды раз, чтобы получить идеологически монолитную армию, бюрократию и полицию. После короткого периода службы каждая имитационная копия будет заменяться свежей копией того же самого сохраненного состояния, чтобы не допустить ее идеологического смещения.

> В рамках одной юрисдикции это позволило бы осуществлять невероятно тщательное наблюдение и регулирование — в принципе, можно было бы иметь одну такую исходную копию для всех граждан. С подобным подходом под силу совершить многое: запретить производство оружия массового поражения; ограничить эксперименты с эмуляциями мозга и воспроизводство эмуляторов; принять либеральную демократическую конституцию — или навеки утвердить мерзость тоталитаризма[39].

Чего можно добиться, обладая таким потенциалом? В первую очередь — консолидации власти и даже концентрации ее в руках очень ограниченного числа действующих лиц.

Объединение на договорных началах

Большую пользу в постпереходном многополярном мире способно принести международное сотрудничество. Может быть, нам на самом деле удастся совершить многое. Избавиться от войн и гонки вооружений. Освоить астрофизические источники информации и начать использовать их в оптимальные для человечества сроки. Избежать спешки в деле усовершенствования ИИ и разработки его новых форм, то есть координировать исследования по этим вопросам, анализировать новые идеи и тщательно проверять все проекты. Отложить или остановить работу над теми проектами, в которых заложена потенциальная угроза существования нашей цивилизации. Ввести одинаковые нормы регулирования в глобальном масштабе, включая положения о гарантированных стандартах жизни (которые потребуют некоторых мер контроля над популяциями) и предотвращении эксплуатации и злоупотребления по отношению к имитационным моделям и другим цифровым и биологическим сущностям. Более того, агенты, имеющие ресурсы, вполне удовлетворяющие их ценностные потребности (подробнее об этом будет рассказано в главе тринадцатой), могли бы заключить договор о разделе будущих ресурсов, по которому им была бы гарантирована определенная доля, вместо того чтобы развертывать борьбу за единоличный контроль над ними, рискуя не получить ничего.

Однако потенциальная возможность получить большую пользу от сотрудничества вовсе не означает, что о нем удастся договориться. В современном обществе за счет лучшей координации усилий в общемировом масштабе можно было бы обрести многие блага — наряду с прочим это сокращение военных расходов, прекращение войн, наложение ограничений на чрезмерный вылов рыбы, снятие торговых барьеров, предотвращение

загрязнения атмосферы. Тем не менее никто так и не пожинает эти созревшие плоды. Почему? Что мешает нам в полной мере воспользоваться возможностями сотрудничества ради абсолютного общего блага?

Одним из препятствий является сложность в согласовании договоренностей и обеспечении контроля за выполнением обязательств по всем договорам. Двое бывших противников, обладающих ядерным оружием, стали бы жить намного лучше, если оба смогли бы избавиться от атомных бомб. Однако, даже достигнув принципиального согласия по всем вопросам, вряд ли будет возможно полное разоружение из-за взаимных опасений, что другая сторона начнет мошенничать. Избавиться от вполне реальных страхов помогло бы создание соответствующих механизмов контроля за соблюдением соглашения. Например, можно учредить институт инспекторов, наблюдающих за уничтожением имеющихся запасов, контролирующих ядерные реакторы и другие объекты; кроме того, в целях гарантии, что никто из договорившихся сторон не возобновляет ядерную программу, они могли бы проводить технические исследования и собирать разведывательные данные. Конечно, их труд должен оплачиваться, но возможные издержки не ограничиваются суммами вознаграждений. Есть риск, что инспекторы станут шпионить и продавать коммерческие и военные секреты. Однако важнее всего, что каждая сторона будет подозревать другую в намерении тайно сохранить возможность производить ядерное оружие. В итоге очень многие потенциально выгодные сделки никогда не станут реальностью, поскольку соблюдение их условий проверить было бы чрезвычайно трудно.

Если станут доступны новые методы проверки, способные снизить затраты на контролирующие органы, можно будет ожидать расширения сотрудничества. Однако пока неясно, удастся ли снизить эти издержки и риски в постпереходную эпоху. Одновременно с новыми эффективными методами инспекции наверняка появятся и новейшие методы сокрытия обмана. В частности, все более активная деятельность, подлежащая регулированию, будет вестись в киберпространстве, где нет возможности непосредственного наблюдения. Например, цифровые исследователи, работающие над инновационными системами нанотехнологических вооружений или новейшим поколением ИИ, могут почти не оставлять следов в физическом мире. И виртуальные инспекторы не всегда смогут пробиться через все слои маскировки и шифрования, с помощью которых нарушитель договора будет скрывать свою незаконную деятельность.

Если удастся разработать надежные детекторы лжи, они станут чрезвычайно полезными инструментами контроля за соблюдением достигнутых соглашений[40]. Протоколы инспекций могут включать положения о проведении интервью с ключевыми персонами, чтобы выяснить, действительно ли они намерены выполнять все условия договора и не знают ли о случаях его нарушения.

Лица, уполномоченные принимать решения по программе и желающие обмануть проверку с применением детекторов лжи, могут сделать это, вначале дав распоряжение подчиненным продолжать незаконную деятельность и скрывать ее от них самих, а затем проведя процедуру частичного стирания памяти о своем участии в этих махинациях. В результате развития нанотехнологий соответствующие методы точечного воздействия на биологическую память наверняка окажутся доступными. В случае машинного интеллекта сделать это будет еще проще.

Страны могут попытаться решить эту проблему, обязавшись участвовать в процессе непрерывного контроля с регулярными проверками ключевых лиц на детекторе лжи, с целью выяснить, не скрывают ли они злокозненных намерений нарушить какой-либо из заключенных или планируемых к заключению международных договоров. Такое обязательство может трактоваться как своего рода метадоговор, предполагающий проверку выполнения остальных договоров; но страны могут присоединяться к нему по собственной инициативе с целью представить себя в качестве заслуживающих доверия партнеров. Однако с этим обязательством об участии в проверках или метадоговором связана та же проблема возможного мошенничества по схеме «делегировал и забыл». В идеале метадоговор должен начать действовать *до того*, как у сторон появится возможность провести подготовку к нарушению договоренностей. Если у потенциального нарушителя будет время на то, чтобы посеять семена обмана, доверие обеспечить не удастся.

В некоторых случаях для заключения сделки бывает достаточно одной возможности *обнаружить* нарушение соглашения. Однако часто все-таки требуется создать механизм *принуждения* к выполнению условий или назначения наказания в случае их нарушения. Потребность в таком механизме может возникнуть в том случае, когда угрозы выхода из соглашения обманутой стороны недостаточно для предотвращения нарушения, например если нарушитель получит в результате такое преимущество, что ему можно будет не беспокоиться о реакции другой стороны.

При наличии высокоэффективных методов выбора мотивации эта проблема может быть решена за счет создания независимого агентства с достаточными полицейскими или военными полномочиями для принуждения к исполнению договора любой стороны или сторон, даже невзирая на их сопротивление. Для этого агентство должно пользоваться полным доверием. Но при наличии эффективных методов выбора мотивации требуемое доверие можно обеспечить за счет привлечения к совместному контролю за созданием этого агентства всех участников переговоров.

Наделение независимого агентства чрезмерной принудительной властью может привести к тем же самым проблемам, с которыми мы столкнулись при обсуждении однополярного исхода (когда синглтон возникает до или в начале революции машинного интеллекта). Чтобы можно было добиваться выполнения договоренностей, которые касаются жизненно важных интересов безопасности стран-противников, независимое агентство фактически должно представлять собой синглтон: стать общемировым сверхразумным Левиафаном. Разница заключается лишь в том, что мы рассматриваем постпереходную ситуацию, в которой агенты, создающие Левиафана, обладают большей компетентностью, чем сегодняшние люди. Будущие создатели Левиафана сами могут быть сверхразумными. Это значительно увеличивает их шансы, что им удастся решить проблему контроля и разработать такую структуру агентства, которая будет служить интересам всех сторон.

Есть ли еще какие-то препятствия у общемировой координации, помимо затрат, связанных с контрольным наблюдением и принуждением к исполнению соглашений? Возможно, главной проблемой остается то, что можно назвать *издержки ведения переговоров*[41]. Даже когда есть возможность для компромисса, результаты которого выгодны всем участникам переговоров, иногда не удается нащупать основания продвигаться дальше, поскольку стороны не могут договориться о принципах разделения общей «добычи». Например, сделка не будет заключена, а потенциальная прибыль — получена, если двое могут подписать договор, по которому будут получать проценты от суммы в один доллар, но каждая сторона считает, что заслуживает шестьдесят центов чистой прибыли, и отказывается согласиться на меньшее. В целом из-за сделанного некоторыми участниками стратегического выбора характера ведения переговоров последние могут идти с трудом, затянуться надолго или вовсе сорваться.

В реальной жизни людям часто удается прийти к соглашению, несмотря на то что иногда приходится уступать стратегические позиции в процессе поиска компромисса (нередко ценой значительных затрат усилий и времени). Тем не менее вполне возможно, что в постпереходную эпоху проблема заключения стратегически важной сделки будет иметь несколько иную динамику. Искусственный интеллект, выступающий в роли посредника, будет последовательно придерживаться формальной и логической концепции, возможно, содержащей новые и неожиданные заключения, если сравнивать ее с позицией других участников переговоров. Кроме того, ИИ способен вступить в игру, недоступную для людей или очень трудную для них, в том числе он может первым занять жесткую позицию относительно проведения какой-то стратегии или выполнения каких-то действий. Хотя люди (и управляемые людьми институты) тоже время от времени способны на это — правда, обычно они не бывают столь конкретны в формулировках и вызывают к себе меньше доверия. Некоторые типы машинного интеллекта, заняв изначально жесткую позицию, будут непоколебимо стоять на своем[42].

Внедрение новейших методов проверки намерений участников договора способно серьезно изменить природу переговоров и принести большую пользу агенту, имеющему преимущество первого хода. Если его участие необходимо для получения какой-то потенциальной выгоды от сотрудничества и агент готов занять принципиально твердую позицию, то он будет в состоянии диктовать распределение этой выгоды, заявив, например, что не согласится на сделку, которая принесет ему, скажем, меньше 99 процентов прибыли. Другие агенты в итоге окажутся перед выбором: или не получить ничего (отказавшись от несправедливого предложения), или получить оставшийся один процент (согласившись на него). Если твердость намерений первого агента, занявшего непоколебимую жесткую позицию, можно проверить и подтвердить инструментальными методами, то его партнерам по переговорам действительно не остается никакого выхода, кроме как принять один из двух оставшихся вариантов.

Чтобы не допустить подобного манипулирования, любой агент может первым занять жесткую позицию, объявив о недопустимости шантажа и готовности отклонять все несправедливые предложения. Когда кто-то займет такую позицию (и сообщит о ней), другие агенты могут решить, что не в их интересах угрожать или самим заявлять о готовности согласиться на сделку лишь в том случае, если она будет только в их интересах, поскольку будут

знать, что эти угрозы окажутся беспочвенными, а несправедливые предложения — отклоненными. Но это лишь еще раз подтверждает, что преимущество остается за тем, кто сделал первый ход. Агент, занявший твердую позицию и сделавший первый ход, может выбирать, ограничиться ли ему лишь предупреждением другим о недопустимости получения несправедливых преимуществ или самому попытаться захватить львиную долю будущей добычи.

В самом выигрышном положении, видимо, окажется агент, темперамент или ценностные установки которого позволят ему не реагировать на угрозы, не поддаваться ни на какие манипуляции и не соглашаться на сделки, по которым ему не будет гарантирована справедливая прибыль. Мы знаем, что и среди людей встречаются переговорщики, умеющие проявлять железную волю и непреклонность[43]. Однако такая жесткость позиции может сыграть злую шутку, если выяснится, что и другие агенты-переговорщики нацелены на получение только справедливой доли и не готовы отступать. Тогда непреклонная решимость одной стороны столкнется с непоколебимостью другой стороны, в результате чего окажется, что невозможно достичь никакого соглашения (ситуация может дойти вплоть до объявления войны). Кроткий и безвольный мог бы выторговать пусть не справедливую прибыль, но хоть какой-то процент.

Пока неясно, какого рода устойчивость с точки зрения теории игр может быть приобретена в подобных переговорах в условиях постпереходной экономики. Агенты могут выбрать и более сложные стратегии, чем описанные нами. Остается лишь *надеяться*, что баланс будет достигнут, поскольку переговорщики все-таки выработают более или менее справедливую норму — этакую точку Шеллинга, служащую единственным ориентиром в обширном пространстве исходов; ориентиром, который благодаря общим ожиданиям станет основой для координации в ничем иным не определенной игре по объединению. Это равновесие может подпитываться какими-то нашими эволюционными установками и культурным программированием — общее стремление к справедливости, при условии, что нам удастся сохранить свои ценности и в постпереходную эпоху, определит ожидания и стратегию так, что установится привлекательное для всех устойчивое равновесие[44].

Во всяком случае, можно сделать вывод, что готовность занимать непоколебимо твердую позицию способна привести к непривычным для нас

вариантам завершения переговоров. Даже если постпереходная эпоха начнется как многополярная, может получиться так, что почти сразу возникнет синглтон как результат заключенного соглашения, которое разрешит все важные проблемы глобальной координации. Благодаря новым технологиям, доступным усовершенствованным формам машинного интеллекта, могут резко снизиться некоторые транзакционные издержки, в том числе, возможно, затраты на контрольные проверки и принуждение к исполнению договоренностей. Однако издержки, связанные с поиском выгодных для обеих сторон условий и достижением компромисса, могут оставаться довольно высокими. Безусловно, разные стратегии торга оказывают влияние на природу переговоров, но при этом все равно неясны причины, по которым достижение такого соглашения могло бы откладываться слишком надолго, тем более если само соглашение заключается в том, чтобы быть достигнутым. Если все-таки к соглашению не удастся прийти, будет иметь место противостояние в той или иной форме; в результате победит либо одна сторона — и вокруг выигравшей коалиции образуется синглтон; либо конфликт превратится в вечный — тогда синглтон может никогда не сформироваться. В итоге мы получим результат несравнимо худший относительно того, который можно было бы планировать, стремясь к скоординированной и направленной на сотрудничество деятельности человечества и тех цифровых сущностей, которые начнут заселять наш мир.

***

Мы увидели, что многополярность, даже если она будет стабильной, не гарантирует, что выход из ситуации окажется благоприятным. Исходная проблема отношений «принципал–агент» останется нерешенной и будет погребена под горой новых проблем, связанных с неудачей постпереходных усилий по глобальной координации, что только ухудшит общую атмосферу. Поэтому предлагаю вернуться к вопросу, каким образом можно обеспечить безопасность человечества в случае прихода в мир единственного вырвавшегося вперед сверхразума.

Глава двенадцатая
Выработка ценностей

Контроль над возможностями — в лучшем случае мера временная и вспомогательная. Если не планируется держать ИИ в заточении вечно, придется разрабатывать принципы выбора мотивации. Но как быть с ценностями? Сможем ли мы внедрить их в систему искусственного агента таким образом, чтобы он начал руководствоваться ими как своими конечными целями? Пока агент не стал разумным, у него, скорее всего, отсутствуют способности к пониманию или даже представлению, что такое система человеческих ценностей. Однако если откладывать процедуру обучения, дожидаясь, когда ИИ станет сверхразумным, то, вполне вероятно, он начнет сопротивляться такому вмешательству в свою систему мотивации и, как мы видели в седьмой главе, у него на то будут конвергентные инструментальные причины. Загрузка системы ценностей проблема не из легких, но отступать нельзя.

Проблема загрузки системы ценностей

Невозможно перечислить все ситуации, в которых может оказаться сверхразум, и для каждой из них определить действия, которые ему следует совершить. Точно так же невозможно составить список всех миров и определить полезность каждого. В любой реальности, гораздо более сложной, чем игра в крестики-нолики, есть слишком много возможных состояний (и исторических состояний*), чтобы можно было использовать метод полного перебора. Значит, систему мотивации нельзя задать в виде исчерпывающей таблицы поиска. Вместо этого она должна быть определена более абстрактно, в качестве какой-то формулы или правила, позволяющих агенту решить, как поступить в любой ситуации.

---

* *Историческое состояние* (history state) — согласно унифицированному языку моделирования, применяется в контексте составного состояния; используется для запоминания того из последовательных подсостояний, которое было текущим в момент выхода из составного состояния.

Один из формальных путей описания этого правила решений состоит в определении функции полезности. Функция полезности (как мы помним из первой главы) задает ценность каждого возможного исхода или в более общем случае — каждого из так называемых возможных миров. При наличии функции полезности можно определить агента, максимизирующего ожидаемую полезность. В любой момент такой агент выбирает действие, имеющее самое высокое значение полезности. (Ожидаемая полезность рассчитывается путем умножения полезности каждого возможного мира на субъективную вероятность того, что этот мир станет реальностью при условии совершения рассматриваемого действия.) В реальности возможных исходов оказывается слишком много, чтобы можно было точно рассчитать ожидаемую полезность действия. Тем не менее правило принятия решения и функция полезности вместе определяют нормативный идеал — понятие оптимальности, — который агент мог бы разработать, чтобы сделать приближение, причем по мере повышения уровня интеллекта ИИ приближение становится все точнее[1]. Создание машины, способной вычислить хорошее приближение ожидаемой полезности доступных ей действий, является ИИ-полной задачей[2]. В этой главе мы рассматриваем другую задачу — задачу, которая остается таковой даже в случае решения проблемы создания машинного интеллекта.

Схему агента, максимизирующего полезность, мы используем для того, чтобы представить оказавшегося в затруднительном положении программиста, работающего с зародышем ИИ и намеревающегося решить проблему контроля. Для этого он наделяет ИИ конечной целью, соответствующей, в принципе, нормальному человеческому представлению о желаемом исходе. Программист, у которого есть своя система ценностей, хотел бы, чтобы ИИ усвоил ее. Предположим, речь идет о понятии счастья. (Такие же проблемы возникли бы, если бы программиста интересовали такие понятия, как правосудие, свобода, слава, права человека, демократия, экологическое равновесие, саморазвитие.) Таким образом, в терминах ожидаемой полезности программисту нужно определить функцию полезности, которая задает ценность возможных миров в зависимости от уровня счастья, который они обеспечивают. Но как выразить эту функцию в исходном коде? В языках программирования нет понятия «счастье». Чтобы этот термин использовать, ему сначала следует дать определение. Причем недостаточно определить его, используя философские концепции и привычную для человека терминологическую

базу, например: «счастье — это наслаждение потенциальными возможностями, присущими нашей человеческой природе», или каким-то иным не менее мудреным способом. Определение должно быть дано в терминах, используемых в языке программирования ИИ, а в конечном счете с помощью таких базовых элементов, как математические операторы и ссылки на ячейки памяти. Когда смотришь на проблему с этой точки зрения, становится понятна сложность стоящей перед программистом задачи.

Идентифицировать и кодифицировать наши конечные цели так трудно потому, что человек пользуется довольно сложной системой дефиниций. Но эта сложность естественна для нас, поэтому мы ее не замечаем. Проведем аналогию со зрительным восприятием. Зрение точно так же может показаться простым делом, поскольку не требует от нас никаких усилий[3]. Кажется, что нужно всего лишь открыть глаза, и в нашем мозгу тут же возникает богатое, осмысленное, рельефное трехмерное изображение окружающего нас мира. Это интуитивное представление о зрении сродни ощущениям монарха от организации быта в его дворце: ему кажется, что каждый предмет просто появляется в нужном месте в нужное время, притом что механизм, обеспечивающий это, полностью скрыт от его взора. Однако выполнение даже простейшей визуальной задачи: поиск перечницы на кухне — требует проведения колоссального объема вычислительных действий. На базе зашумленной последовательности двумерных паттернов, возникшей в результате возбуждения нервных клеток сетчатки глаза и переданной по глазному нерву в мозг, зрительная кора головного мозга должна реконструировать и интерпретировать трехмерное представление окружающего пространства. Заметная часть поверхности коры головного мозга — нашей драгоценной недвижимости площадью один квадратный метр — занята областью обработки зрительной информации; когда вы читаете эту книгу, над выполнением этой задачи неустанно работают миллиарды нейронов (словно множество швей, склонившихся над своими швейными машинами в ателье и множество раз за секунду успевающих сшивать и снова распарывать огромное стеганое одеяло). Точно так же наши, казалось бы, простые ценности и желания на самом деле очень сложны[4]. Как программист мог бы отразить всю эту сложность в функции полезности?

Один из подходов заключается в том, чтобы попробовать напрямую закодировать полное представление о конечной цели, которую программист назначил для ИИ; иными словами, нужно записать функцию полезности,

применив метод точной спецификации. Этот подход мог бы сработать, если у нас была бы чрезвычайно простая цель, например мы хотели бы знать, сколько десятичных знаков после запятой стоит в числе пи. Еще раз: *единственное*, что нам понадобилось бы от ИИ, чтобы он рассчитал все знаки после запятой в числе пи. И нас не волновали бы никакие иные последствия достижения им этой цели (как мы помним, это проходило по категории пагубных отказов, тип — инфраструктурная избыточность). Было бы полезно при использовании метода точной спецификации выбрать еще метод приручения. Но если развиваемый ИИ должен стать сверхразумным монархом, а его конечной целью является следовать любым возможным *человеческим* ценностям, тогда метод точной спецификации, необходимый для полного определения цели, — безнадежно недостижимая задача[5].

Допустим, мы не можем загрузить в ИИ описание человеческих ценностей с помощью их полного представления на языке программирования — тогда что еще можно попробовать сделать? В этой главе мы обсудим несколько альтернативных путей. Какие-то из них на первый взгляд представляются вполне возможными, но при ближайшем рассмотрении оказываются гораздо менее выполнимыми. В дальнейшем имеет смысл обсуждать те пути, которые останутся открытыми.

Решение проблемы загрузки системы ценностей — задача, достойная усилий лучших представителей следующего поколения талантливых математиков. Мы не можем себе позволить откладывать решение этой проблемы до тех времен, когда усовершенствованный ИИ станет настолько разумным, что с легкостью раскусит наши намерения. Как мы уже знаем из раздела об инструментальной конвергенции (глава седьмая), он будет сопротивляться попыткам изменить его конечные цели. Если искусственный агент еще не стал абсолютной дружественным к моменту, когда обрел возможность размышлять о собственной агентской сущности, он вряд ли благосклонно отнесется к нашим планам по «промывке мозгов» или к заговору с целью заменить его на другого агента, отличающегося большим благорасположением к своим создателям и ближайшим соседям.

Естественный отбор

Эволюция уже один раз создала живое существо, наделенное системой ценностей. Этот неоспоримый факт может вдохновить на размышления, что проблему загрузки ценностей в ИИ можно решить эволюционными мето-

дами. Однако на этом пути — не столь безопасном, как кажется, — нас ожидают некоторые препятствия. Мы вспоминали о них в конце десятой главы, когда обсуждали, насколько опасными могут быть мощные поисковые процессы.

Эволюцию можно рассматривать в качестве отдельного класса поисковых алгоритмов, предполагающих двухэтапную настройку: на одном этапе — популяция возможных решений расширяется за счет новых кандидатов в соответствии с каким-то простым стохастическим правилом (например, случайной мутацией или половой рекомбинацией), на другом — популяция сокращается за счет отсева кандидатов, показывающих неудовлетворительные результаты тестирования при помощи оценочной функции. Как и в случае многих других типов мощного поиска, есть риск, что этот процесс отыщет решение, действительно удовлетворяющее формально определенному критерию поиска, но не отвечающее нашим моральным ожиданиям. (Это может случиться независимо от того, стремимся ли мы создать цифровой разум, имеющий такие же цели и ценности, как у среднестатистического человека, или, напротив, представляющий собой образец нравственности или идеал покорности.) Такого риска можно избежать, если не ограничиваться одноаспектным запросом на то, что мы хотим разработать, а постараться описать формальный критерий поиска, точно отражающий все измерения нашей цели. Но это уже оборачивается полновесной проблемой загрузки системы ценностей — и тогда нужно исходить из того, что она решена. В этом случае возникает следующая проблема, изложенная Ричардом Докинзом в книге «Река, текущая из рая»:

> Общее количество страдания в мире в год превосходит все мыслимые пределы. За минуту, которая потребовалась мне для написания этого предложения, тысячи животных были съедены живьем; спасались от хищников бегством, скуля от страха; медленно погибали из-за пожирающих их изнутри паразитов; умирали от голода, жажды и болезней[6].

Даже если ограничиться одним нашим видом, то ежедневно погибает сто пятьдесят тысяч человек, и бесконечное количество людей страдает от всевозможных мучений и лишений[7]. Может быть, природа и великий экспериментатор, но на свои опыты она никогда не получит одобрения у совета по этике, поскольку постоянно нарушает Хельсинкскую декларацию

со всеми ее этическими нормами*, причем с точек зрения и левых, и правых, и центристов. Важно другое: чтобы мы сами не шли слепо по пятам природы и не воспроизводили бездумно *in silico*** все эти ужасы. Правда, вряд ли у нас получится совсем избежать проявлений преступной безнравственности, если мы собираемся создавать искусственный интеллект по образу и подобию человеческого разума, опираясь на эволюционные методы, — чтобы повторить хотя бы на минимальном уровне естественный процесс развития, называемый биологической эволюцией[8].

Обучение с подкреплением

Обучение с подкреплением — это область машинного обучения, в которой агенты могут учиться максимизировать накопленное вознаграждение. Формируя нужную среду, в которой поощряется любое желательное качество агента, можно создать агента, способного научиться решать широкий круг задач (даже в отсутствие подробной инструкции или обратной связи с программистами, но лишь бы присутствовал сигнал о поощрении). Часто алгоритм обучения с подкреплением включает в себя постепенное построение некоторой функции оценки, которая присваивает значение ценности состояниям, парам состояние–действие и различным стратегическим направлениям. (Например, программа может научиться играть в нарды, используя обучение с подкреплением для постепенного развития навыка оценки позиций на доске.) Можно считать, что эта функция оценки, постоянно меняющаяся с опытом, в том числе включает в себя и обучение нужным целям. Однако то, чему учится агент, это не новые *конечные ценности*, но все более точные *оценки инструментальной ценности* достижения определенных состояний (или совершения определенных действий в определенных состояниях, или следования определенной политике). Поскольку конечная цель остается величиной постоянной, мы всегда можем описать агента,

---

* *Хельсинкская декларация* (Declaration of Helsinki) — набор этических принципов для медицинского сообщества, касающихся экспериментов на людях; разработана Всемирной медицинской ассоциацией в 1964 году.

** *In silico* (лат.) — термин, обозначающий компьютерное моделирование (симуляцию) эксперимента, чаще биологического; в нашем контексте речь идет об имитационных моделях головного мозга человека. Сама фраза тоже представляет собой «имитацию», так как создана по аналогии с такими моделями, как *in vivo* («в живом организме») и *in vitro* («в пробирке»).

проходящего обучение с подкреплением, как агента, имеющего конечную цель. Эта неизменная конечная цель агента — его стремление получать максимальное поощрение в будущем. Вознаграждение состоит из специально разработанных объектов восприятия, помещенных в его окружающую среду. Таким образом, в результате обучения с подкреплением у агента формируется устойчивый эффект самостимуляции (о котором подробно говорилось в главе восьмой), то есть агент начинает выстраивать собственную довольно сложную модель такого мира, который в состоянии предложить ему альтернативный вариант максимизации вознаграждения[9].

Наши замечания не подразумевают, будто обучение с подкреплением нельзя применять для развития безопасного для нас зародыша ИИ, мы лишь хотим сказать, что его использование следует соотносить с системой мотивации, которая сама по себе не основана на принципе максимизации вознаграждения. Тогда, чтобы решить проблему загрузки системы ценностей, потребуется искать иные подходы, нежели метод обучения с подкреплением.

Ассоциативная модель ценностного приращения

Невольно возникает вопрос: если проблема загрузки системы ценностей столь неподатлива, как нам самим удается обзаводиться ценностной ориентацией?

Одна из возможных (чрезмерно упрощенных) моделей выглядит примерно так. Мы вступаем в жизнь не только с относительно простым набором базовых предпочтений (иначе почему бы мы с детства испытывали неприятные ощущения от каких-то возбудителей и старались инстинктивно избегать этого?), но и с некоторой склонностью к приобретению дополнительных предпочтений, что происходит за счет обогащения опытом (например, у нас начинают формироваться определенные эстетические предпочтения, поскольку мы видим, что в нашем культурном пространстве какие-то цели и идеалы особо ценностны, а какое-то поведение весьма поощряется). И базовые первичные предпочтения, и склонность приобретать в течение жизни ценностные предпочтения являются врожденными чертами человека, сформированными в результате естественного и генетического отбора в ходе эволюции. Однако дополнительные предпочтения, которые складываются у нас к моменту взросления, зависят от жизненного пути. Таким образом, большая часть информационно-семантических моделей, имеющих отношение

к нашим конечным ценностям, не заложена генетически, а приобретена благодаря опыту.

Например, в нашей жизни появился любимый человек, и конечно, для нас важнейшей конечной ценностью становится его благополучие. От каких механизмов зависит появление этой ценности? Какие смысловые структуры задействованы в ее формировании? Структур много, но мы возьмем лишь две — понятие «человек» и понятие «благополучие». Ни эти, ни какие другие представления непосредственно не закодированы в нашей ДНК. Скорее, в ДНК хранится информация и инструкции по строительству и развитию нашего мозга, а значит, и нашего разума, который, пребывая в человеческой среде обитания, за несколько лет создает свою модель мира — модель, включающую и дефиницию человека, и дефиницию благополучия. Только после того как сложились эти два представления, можно приступать к объяснению, каким таким особым значением наполнена наша конечная ценность. А теперь вернемся к первому вопросу: от каких механизмов зависит появление наших ценностных предпочтений? Почему желание блага любимому человеку формируется вокруг *именно этих* обретенных нами представлений, а не каких-то других, тоже обретенных, — вроде представлений о цветочном горшке или штопоре? Вероятно, должен существовать какой-то особый врожденный механизм.

Как работает сам механизм, нам неизвестно. Он, видимо, очень сложный и многогранный, особенно в отношении человека. Поэтому, чтобы хоть как-то понять, как он действует, рассмотрим его примитивную форму на примере животных. Возьмем так называемую реакцию следования (геномный, или родительский, импринтинг), в частности, у выводковых птиц, когда только что вылупившийся, но уже сформированный, птенец сразу начинает неотступно следовать за родителями или первым увиденным движущимся объектом. За каким объектом-«мамой» птенец пожелает двигаться, зависит от его первого опыта, но сам процесс запечатления в памяти соответствующей сенсорной информации (импринтинг) обусловлен генетическими особенностями. Попытаемся провести аналогию с человеческими привязанностями. Когда Гарри встретил Салли, ее благополучие стало для него абсолютной ценностью, но предположим, что они так и не встретились, и Гарри полюбил бы другую; тогда, может быть, его ценностные предпочтения тоже были бы иными. Способность генов человека кодировать механизм выработки целеполагания лишь объясняет, почему наша конечная

цель обрастает разнообразными информационно-семантическими моделями, но их сложная организация никак не обусловлена генетически.

Следовательно, возникает вопрос: можно ли построить систему мотивации для искусственного интеллекта, основанную на этом принципе? То есть вместо описания сложной системы ценностей напрямую определить некий механизм, который обеспечил бы приобретение этих ценностей в процессе взаимодействия ИИ с определенной средой.

Похоже, имитировать процесс формирования ценностей, характерный для людей, непросто. Соответствующий человеческий генетический механизм стал результатом колоссальной работы, проделанной эволюцией, и повторить ее работу будет трудно. Более того, механизм, вероятно, рассчитан на нейрокогнитивную систему человека и поэтому неприменим к машинному интеллекту за исключением имитационных моделей. Но даже если полная эмуляция головного мозга окажется возможной, лучше будет начать с загрузки разума взрослого человека — разума, уже содержащего полное представление о некоторой совокупности человеческих ценностей[10].

Таким образом, попытка разработать модель ценностного приращения, точно имитирующую процесс формирования системы ценностей человека, означает безуспешную серию атак на проблему загрузки ценностей. Но, возможно, мы могли бы создать более простой искусственный механизм импорта в целевую систему ИИ высокоточных представлений о нужных нам ценностях? Чтобы добиться успеха, не обязательно снабжать ИИ точно такой же, как у людей, врожденной склонностью приобретать ценностные предпочтения. Возможно, это даже нежелательно — в конце концов, человеческая природа несовершенна, человек слишком часто делает выбор в пользу зла, что неприемлемо в любой системе, способной получить решающее стратегическое преимущество. Наверное, лучше ориентироваться на систему мотивации, не всегда соответствующей человеческим нормам, например такую, которой свойственна тенденция формировать конечные цели, полные бескорыстия, сострадания и великодушия, — любого, имеющего такие качества, мы сочли бы образцовым представителем человеческого рода. Эти конечные цели должны отклоняться от человеческой нормы в строго определенном направлении, иначе их трудно будет считать улучшениями; кроме того, они должны предполагать наличие неизменной антропоцентричной системы координат, при помощи которой можно делать значимые с человеческой точки зрения оценочные обобщения (чтобы избежать порочной реализации

на базе искусственно приемлемых описаний цели, которую мы рассматривали в главе восьмой). Вопрос, насколько такое возможно, по-прежнему остается открытым.

Еще одна проблема, связанная с ассоциативной моделью ценностного приращения, заключается в том, что ИИ может просто отключить этот механизм приращения. Как мы видели в седьмой главе, неприкосновенность целевой системы является его конвергентной инструментальной целью. Достигнув определенной стадии когнитивного развития, ИИ может начать воспринимать продолжающуюся работу механизма приращения как враждебное вмешательство[11]. Это необязательно плохо, но нужно с осторожностью подходить к блокировке целевой системы, чтобы ее отключение произошло в правильный момент: *после* того, как были приобретены нужные ценности, но *до* того, как они будут перезаписаны в виде непреднамеренного приращения.

Строительные леса для мотивационной системы

Есть еще один подход к решению проблемы загрузки системы ценностей, который можно назвать «возведение строительных лесов». Подход состоит в наделении зародыша ИИ временными сравнительно простыми конечными целями, которые можно выразить прямым кодированием или каким-то иным доступным способом. Наступит время, и ИИ будет способен формировать более сложные представления. Тогда мы снимем мотивационные «леса» и заменим временные ценности на новые, которые останутся конечной ценностной системой ИИ, даже когда он разовьется в полноценный сверхразум.

Поскольку временные цели — не просто инструментальные, но *конечные* цели ИИ, можно ожидать, что он будет сопротивляться их замене (неприкосновенность системы целей является конвергентной инструментальной ценностью). В этом и состоит главная опасность. Если ИИ преуспеет в противодействии замене временных целей постоянными, метод потерпит неудачу.

Чтобы избежать такого отказа, необходимо соблюдать осторожность. Например, можно использовать метод контроля над возможностями, чтобы ограничить свободу ИИ до тех пор, пока не будет инсталлирована зрелая система мотивации. В частности, можно попробовать остановить его когнитивное развитие на таком уровне, где можно безопасно и эффективно

наделить ИИ желательными для нас конечными целями. Для этого нужно затормозить совершенствование отдельных когнитивных способностей, в частности, таких, которые требуются для выработки стратегии и хитроумных схем в духе Макиавелли, при этом позволив развиваться более безобидным (предположительно) способностям.

Программисты могут попробовать создать атмосферу сотрудничества с ИИ при помощи методов выбора мотивации. Например, использовав такую временную цель, как готовность выполнять команды людей, в том числе команд, предполагающих замену любых имеющихся целей ИИ[12]. К другим временным целям относятся прозрачность ценностей и стратегии ИИ, а также разработка легкой для понимания программистами архитектуры, включающей последнюю версию конечной цели, значимой с точки зрения людей, и мотивированность к приручению (например, к ограничению использования вычислительных ресурсов).

Можно было бы попробовать и такой вариант: со временем заменить зародыш ИИ, наделенный единственной конечной целью, на аналогичную версию зародыша, но уже с другой конечной целью, заданной программистами косвенным образом. С такой заменой связаны некоторые трудности, особенно в контексте подхода к обучению целям, который мы обсудим в следующем разделе. Другие трудности будут рассмотрены в главе тринадцатой.

Метод возведения строительных лесов для мотивационной системы не лишен недостатков. В частности, есть риск, что ИИ станет слишком могущественным прежде, чем будет изменена его временная целевая система. Тогда он может воспротивиться (явно или тайно) усилиям программистов по ее замене на постоянную. В результате на этапе превращения зародыша ИИ в полноценный сверхразум останутся актуальными старые конечные цели. Еще один недостаток состоит в том, что наделение ИИЧУ желательными для разработчиков конечными целями может оказаться не таким простым делом, как в случае более примитивного ИИ. В отличие от него зародыш ИИ представляет собой *tabula rasa*, позволяя сформировать любую его структуру по желанию программистов. Этот недостаток может превратиться в преимущество, если удастся наделить зародыш ИИ временными целями, благодаря которым он будет стремиться к созданию такой архитектуры, которая поможет разработчикам в их последующих усилиях по заданию ему постоянных конечных целей. Однако пока неясно, легко ли обеспечить

наличие у временных целей зародыша ИИ такого свойства, а также будет ли способен даже идеально мотивированный ИИ создать лучшую архитектуру, чем команда программистов-людей.

Обучение ценностям

Теперь переходим к загрузке ценностей — серьезная проблема, которую придется решать довольно мягким методом. Он состоит в *обучении* ИИ ценностям, которые мы хотели бы ему поставить. Для этого потребуется хотя бы неявный критерий их отбора. Можно настроить ИИ так, чтобы он действовал в соответствии со своими представлениями об этих неявно заданных ценностях. Данные представления он будет уточнять по мере расширения своих знаний о мире.

В отличие от метода мотивационных строительных лесов, когда ИИ наделяется временной конечной целью, которая потом заменяется на отличную от нее постоянную, в методе обучения ценностям конечная цель не меняется на стадии разработки и функционирования ИИ. Обучение меняет не саму цель, а представления ИИ об этой цели.

Таким образом, у ИИ должен быть критерий, при помощи которого он мог бы определять, какие объекты восприятия содержат свидетельства в пользу некоторой гипотезы, что представляет собой конечная цель, а какие — против нее. Определить подходящий критерий может быть трудно. Отчасти эта трудность связана с самой задачей создания ИИ, которому требуется мощный механизм обучения, способный определять структуру окружающего мира на основании ограниченных сигналов от внешних датчиков. Этой проблемы мы касаться не будем. Но даже если считать задачу создания сверхразумного ИИ решенной, остаются трудности, специфические для проблемы загрузки системы ценностей. В случае метода обучения целям они принимают форму определения критерия, который связывает воспринимаемые потоки информации с гипотезами относительно тех или иных целей.

Прежде чем глубже погрузиться в метод обучения ценностям, было бы полезно проиллюстрировать идею на примере. Возьмем лист бумаги, напишем на нем определение какого-то набора ценностей, положим в конверт и заклеим его. После чего создадим агента, обладающего общим интеллектом человеческого уровня, и зададим ему следующую конечную цель: «Максимизировать реализацию ценностей, описание которых находится в этом конверте». Что будет делать агент?

Он не знает, что содержится в конверте. Но может выстраивать гипотезы и присваивать им вероятности, основываясь на всей имеющейся у него информации и доступных эмпирических данных. Например, анализируя другие тексты, написанные человеком, или наблюдая за человеческим поведением и отмечая какие-то закономерности. Это позволит ему выдвигать догадки. Не нужно иметь диплом философа, чтобы предположить, что, скорее всего, речь идет о заданиях, связанных с определенными ценностями: «минимизируй несправедливость и бессмысленные страдания» или «максимизируй доход акционеров», вряд ли его попросят «покрыть поверхность всех озер пластиковыми пакетами».

Приняв решение, агент начинает действовать так, чтобы реализовать ценности, которые, по его мнению, с наибольшей вероятностью содержатся в конверте. Важно, что при этом он будет считать важной инструментальной целью как можно больше узнать о содержимом конверта. Причина в том, что агент мог бы лучше реализовать почти любую конечную ценность, содержащуюся в конверте, если бы знал ее точную формулировку — тогда он действовал бы гораздо эффективнее. Агент также обнаружит конвергентные инструментальные причины (описанные в главе седьмой): неизменность целей, улучшение когнитивных способностей, приобретение ресурсов и так далее. И при этом, если исходить из предположения, что он присвоит достаточно высокую вероятность тому, что находящиеся в конверте ценности включают благополучие людей, он не станет стремиться реализовать эти инструментальные цели за счет немедленного превращения планеты в компьютрониум, тем самым уничтожив человеческий вид, поскольку это будет означать риск окончательно лишиться возможности достичь конечной ценности.

Такого агента можно сравнить с баржей, которую несколько буксиров тянут в разные стороны. Каждый буксир символизирует какую-то гипотезу о конечной ценности. Мощность двигателя буксира соответствует вероятности гипотезы, поэтому любые новые свидетельства меняют направление движения баржи. Результирующая сила перемещает баржу по траектории, обеспечивающей обучение (неявно заданной) конечной ценности и позволяющей обойти мели необратимых ошибок; а позднее, когда баржа достигнет открытого моря, то есть более точного знания конечной ценности, буксир с самым мощным двигателем потянет ее по самому прямому или благоприятному маршруту.

Метафоры с конвертом и баржей иллюстрируют принцип, лежащий в основе метода обучения ценностям, но обходят стороной множество критически важных технических моментов. Они станут заметнее, когда мы начнем описывать этот метод более формально (см. врезку 10).

Как можно наделить ИИ такой целью: «максимизируй реализацию ценностей, изложенных в записке, лежащей в запечатанном конверте»? (Или другими словами, как определить критерий цели — см. врезку 10.) Чтобы сделать это, необходимо определить место, где описаны ценности. В нашем примере это требует указания ссылки на текст в конверте. Хотя эта задача может показаться тривиальной, но и она не без подводных камней. Упомянем лишь один: критически важно, чтобы ссылка была не просто на некий внешний физический объект, но на объект по состоянию на определенное время. В противном случае ИИ может решить, что наилучший способ достичь своей цели — это заменить исходное описание ценности на такое, которое значительно упростит задачу (например, найти большее число для некоторого целого числа). Сделав это, ИИ сможет расслабиться и бить баклуши — хотя скорее за этим последует опасный отказ по причинам, которые мы обсуждали в главе восьмой. Итак, теперь встал вопрос, как определить это время. Мы могли бы указать на часы: «Время определяется движением стрелок этого устройства», — но это может не сработать, если ИИ предположит, что в состоянии манипулировать временем, управляя стрелками часов. И он будет прав, если определять «время» так, как это сделали мы. (В реальности все будет еще сложнее, поскольку соответствующие ценности не будут изложены в письменном виде. Скорее всего, ИИ придется выводить ценности из наблюдений за внешними структурами, содержащими соответствующую информацию, такими как человеческий разум.)

ВРЕЗКА 10. ФОРМАЛИЗАЦИЯ ОБУЧЕНИЯ ЦЕННОСТЯМ

Чтобы яснее понять метод, опишем его более формально. Читатели, которые не готовы погружаться в математические выкладки, могут этот раздел пропустить.

Предположим, что есть упрощенная структура, в которой агент взаимодействует со средой конечного числа моментов[13]. В момент k агент выполняет действие y_k, после чего получает ощущение x_k. История взаимодействия агента со средой в течение жизни m описывается цепочкой $y_1x_1y_2x_2\ldots y_mx_m$ (которую мы представим в виде $yx_{1:m}$ или $yx_{\leq m}$). На каждом шаге агент выбирает действие на основании последовательности ощущений, полученных к этому моменту.

Рассмотрим вначале обучение с подкреплением. Оптимальный ИИ, обучающийся с подкреплением (ИИ-ОП), максимизирует будущую ожидаемую награду. Тогда выполняется уравнение[14]:

$$y_x = \arg\max_{y_k} \sum_{x_k yx_{k+??}} (r_k + ... + r_m) P(yx_{\leq m} \mid yx_{<k} y_k) .$$

Последовательность подкреплений $r_k, \ldots, r_m$ вытекает из последовательности воспринимаемых состояний среды $x_{k:m}$, поскольку награда, полученная агентом на каждом шаге, является частью восприятия, полученного на этом шаге.

Мы уже говорили, что такого рода обучение с подкреплением в нынешних условиях не подходит, поскольку агент с довольно высоким интеллектом поймет, что обеспечит себе максимальное вознаграждение, если сможет напрямую манипулировать сигналом системы наград (эффект самостимуляции). В случае слабых агентов это не будет проблемой, поскольку мы сможем физически предотвратить их манипуляции с каналом, по которому передаются вознаграждения. Мы можем также контролировать их среду, чтобы они получали вознаграждение только в том случае, если их действия согласуются с нашими ожиданиями. Но у любого агента, обучающегося с подкреплением, будут иметься серьезные стимулы избавиться от этой искусственной зависимости: когда его вознаграждения обусловлены нашими капризами и желаниями. То есть наши отношения с агентом, обучающимся с подкреплением, фундаментально антагонистичны. И если агент силен, это может быть опасно.

Варианты эффекта самостимуляции также могут возникнуть у систем, не стремящихся получить внешнее вознаграждение, то есть у таких, чьи цели предполагают достижение какого-то внутреннего состояния. Скажем, в случае систем «актор–критик», где модуль актора выбирает действия так, чтобы минимизировать недовольство отдельного модуля критика, который вычисляет, насколько соответствует поведение актора требуемым показателям эффективности. Проблема этой системы следующая: модуль актора может понять, что способен минимизировать недовольство критика, изменив или вовсе ликвидировав его — как диктатор, распускающий парламент и национализирующий прессу. В системах с ограниченными возможностями избежать этой проблемы можно просто: не дав модулю актора никаких инструментов для модификации модуля критика. Однако обладающий достаточным интеллектом и ресурсами модуль актора всегда сможет обеспечить себе доступ к модулю критика (который фактически представляет собой лишь физический вычислительный процесс в каком-то компьютере)[15].

Прежде чем перейти к агенту, который проходит обучение ценностям, давайте в качестве промежуточного шага рассмотрим другую систему, максимизирующую полезность на основе наблюдений (ИИ-МНП). Она получается путем замены последовательности подкреплений ($r_k + \ldots + r_m$) в ИИ-ОП на функцию полезности, которая может зависеть от всей истории будущих взаимодействий ИИ:

$$y_k = \arg\max_{y_k} \sum_{x_k yx_{k+??}} U(yx_{\leq m}) P(yx_{\leq m} \mid yx_{<k} y_k)$$

Эта формула позволяет обойти проблему самостимуляции, поскольку функцию полезности, зависящую от всей истории взаимодействий, можно разработать так, чтобы наказывать истории взаимодействия, в которых проявляются признаки самообмана (или нежелания агента прикладывать достаточные усилия, чтобы получить точную картину действительности).

Таким образом, ИИ-МНП дает возможность обойти проблему самостимуляции *в принципе*. Однако, чтобы ею воспользоваться, нужно задать подходящую функцию полезности на классе всех возможных историй взаимодействия — а это очень трудная задача.

Возможно, более естественным было бы задать функцию полезности непосредственно в терминах возможных миров (или свойств возможных миров, или теорий о мире), а не в терминах историй взаимодействия агента. Используя этот подход, формулу оптимальности ИИ-МНП можно переписать и упростить:

$$y = \arg\max_y \sum_w U(w) P(w \mid Ey).$$

Здесь E — это все свидетельства, доступные агенту (в момент, когда он принимает решение), а U — функция полезности, которая присваивает полезность некоторому классу возможных миров. Оптимальный агент будет выбирать действия, которые максимизируют ожидаемую полезность.

Серьезная проблема этих формул — сложность задания функции полезности. И это наконец возвращает нас к проблеме загрузки ценностей. Чтобы функцию полезности можно было получить в процессе обучения, мы должны расширить наше формальное определение и допустить неопределенность функции полезности. Это можно сделать следующим образом (ИИ-ОЦ)[16]:

$$y = \arg\max_{y \in Y} \sum_{w \in W} (w) P(w \mid Ey) \sum_{u \in U} U(w) P(\nu(U) \mid w),$$

где $\nu(—)$ — функция от функций полезности для предположений относительно функций полезности. $\nu(U)$ — предположение, что функция полезности U удовлетворяет *критерию ценности*, выраженному ν[17].

То есть чтобы решить, какое действие выполнять, нужно действовать следующим образом: во-первых, вычислить условную вероятность каждого возможного мира w (учитывая все возможные свидетельства и исходя из предположения, что должно быть выполнено действие y); во-вторых, для каждой возможной функции U вычислить условную вероятность того, что U удовлетворяет критерию ценности ν (при условии, что w — это реальный мир); в-третьих, для каждой возможной функции полезности U вычислить полезность возможного мира w; в-четвертых, использовать все эти значения для расчета ожидаемой полезности действия y; в-пятых, повторить эту процедуру для всех возможных действий и выполнить действие, имеющее самую высокую ожидаемую полезность (используя любой метод выбора из равных значений в случае возникновения таковых). Понятно, что таким образом описанная процедура — предполагающая явное рассмотрение всех возможных миров — вряд ли реализуема с точки

зрения потребности в вычислительных ресурсах. ИИ придется использовать обходные пути, чтобы аппроксимировать это уравнение оптимальности.

Остается вопрос, как определить критерий ценности *v*[18]. Если у ИИ появится адекватное представление этого критерия, он, в принципе, сможет использовать свой интеллект для сбора информации о том, какие из возможных миров с наибольшей вероятностью могут оказаться реальными. После чего применить критерий ценности для каждого потенциально реального мира, чтобы выяснить, какая целевая функция удовлетворяет критерию в мире *w*. То есть формулу ИИ-ОЦ можно считать одним из способов идентифицировать и выделить ключевую сложность в методе обучения ценностям — как представить *v*. Формальное описание задачи высвечивает также множество других сложностей (например, как определить *Y*, *W* и *U*), с которыми придется справиться прежде, чем метод можно будет использовать[19].

Другая трудность кодирования цели «максимизируй реализацию ценностей из конверта» заключается в том, что даже если в этом письме описаны все правильные ценности и система мотивации ИИ успешно воспользуется этим источником, ИИ может интерпретировать описания не так, как предполагалось его создателями. Это создаст риск порочной реализации, описанной в главе восьмой.

Поясним, что трудность здесь даже не в том, как добиться, чтобы ИИ понял намерения людей. Сверхразум справится с этим без проблем. Скорее, трудность заключается в том, чтобы ИИ был мотивирован на достижение описанных целей так, как предполагалось. Понимание наших намерений это не гарантирует: ИИ может точно знать, что мы имели в виду, и не обращать никакого внимания на эту интерпретацию наших слов (используя в качестве мотивации иную их интерпретацию или вовсе на них не реагируя).

Трудность усугубляется тем, что в идеале (по соображениям безопасности) правильную мотивацию следует загрузить в зародыш ИИ *до того*, как он сможет выстраивать представления любых человеческих концепций и начнет понимать намерения людей. Это потребует создания какого-то когнитивного каркаса, в котором будет предусмотрено определенное место для системы мотивации ИИ как хранилища его конечных ценностей. Но у ИИ должна быть возможность изменять этот когнитивный каркас и развивать свои способности представления концепций по мере узнавания мира и роста интеллекта. ИИ может пережить эквивалент научной революции, в ходе которой его модель мира будет потрясена до основания, и он, возможно, столкнется с онтологическим кризисом, осознав, что его предыдущее видение целей было основано на заблуждениях и иллюзиях. При этом, начиная

с уровня интеллекта, еще не достигающего человеческого, и на всех остальных этапах развития, вплоть до сверхразума галактических масштабов, поведение ИИ должно определяться, по сути, неизменной конечной системой ценностей, которую благодаря этому развитию ИИ понимает все лучше; при этом зрелый ИИ, скорее всего, будет понимать ее совсем не так, как его разработчики, хотя эта разница возникнет не в результате случайных или враждебных действий ИИ, но скорее из добрых побуждений. Как бороться с этим, еще неясно[20] (см. врезку 11).

Подводя итоги, стоит сказать, что пока неизвестно, как использовать метод обучения ценностям для формирования у ИИ ценностной системы, приемлемой для человека (впрочем, некоторые новые идеи можно найти во врезке 12). В настоящее время этот метод следует считать скорее перспективным направлением исследований, нежели доступной для применения техникой. Если удастся заставить его работать, он может оказаться почти идеальным решением проблемы загрузки ценностей. Помимо прочих преимуществ, его использование станет естественным барьером для проявлений с нашей стороны преступной безнравственности, поскольку зародыш ИИ, способный догадаться, какие ценностные цели могли загрузить в него программисты, может додуматься, что подобные действия не соответствуют этим ценностям и поэтому их следует избегать как минимум до тех пор, пока не будет получена более определенная информация.

Последний, но немаловажный, вопрос — что положить в конверт? Или, если уйти от метафор, каким ценностям мы хотели бы обучить ИИ? Но этот вопрос одинаков для всех методов решения проблемы загрузки ценностей. Вернемся к нему в главе тринадцатой.

ВРЕЗКА 11. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, КОТОРЫЙ ХОЧЕТ БЫТЬ ДРУЖЕСТВЕННЫМ

Элиезер Юдковский попытался описать некоторые черты архитектуры зародыша ИИ, которая позволила бы ему вести себя так, как описано выше. В его терминологии такой ИИ должен использовать «семантику внешних ссылок»[21]. Чтобы проиллюстрировать основную идею Юдковского, давайте предположим, что мы хотим создать дружественный ИИ. Его исходная цель — попытаться представить себе некое свойство *F*, но изначально ИИ почти ничего об *F* не знает. Ему известно лишь, что *F* — некоторое абстрактное свойство. И еще он знает, что когда программисты говорят о дружественности, они, вероятно, пытаются передать информацию об *F*. Поскольку конечной целью ИИ является составление формулировки понятия *F*, его важной инструментальной

целью становится больше узнать об *F*. По мере того как ИИ узнает об *F* все больше, его поведение все сильнее определяется истинным содержанием этого свойства. То есть можно надеяться, что чем больше ИИ узнаёт и чем умнее становится, тем более дружелюбным он становится.

Разработчики могут содействовать этому процессу и снизить риск того, что ИИ совершит какую-то катастрофическую ошибку, пока не до конца понимает значение *F*, обеспечивая его «заявлениями программистов» — гипотезами о природе и содержании *F*, которым изначально присваивается высокая вероятность. Например, можно присвоить высокую вероятность гипотезе «вводить программистов в заблуждение недружественно». Однако такие заявления не являются «истиной по определению», аксиомами концепции дружелюбия. Скорее всего, это лишь начальные гипотезы, которым рациональный ИИ будет присваивать высокую вероятность как минимум до тех пор, пока доверяет эпистемологическим способностям программистов больше, чем своим.

Юдковский также предложил использовать то, что он называет «семантика причинной валидности». Идея состоит в том, чтобы ИИ делал не в точности то, что программисты говорят ему делать, но скорее то, что они пытались ему сказать сделать. Пытаясь объяснить зародышу ИИ, что такое дружелюбие, они могли совершить ошибку в своих объяснениях. Более того, сами программисты могли не до конца понимать истинную природу дружелюбия. Поэтому хочется, чтобы ИИ мог исправлять ошибки в их умозаключениях и выводить истинное или предполагавшееся значение из неидеальных объяснений, которые дали ему программисты. Например, воспроизводить причинные процессы появления представлений о дружелюбии у самих программистов и о способах его описания; понимать, что в процессе ввода информации об этом свойстве они могли сделать опечатку; попытаться найти и исправить ее. В более общем случае ИИ следует стремиться исправить последствия любого вмешательства, искажающего поток информации о характере дружелюбия, на всем ее пути от программистов до ИИ (где «искажающий» понимается в эпистемологическом смысле). В идеале по мере созревания ИИ ему следует преодолеть все когнитивные искажения и прочие фундаментально ошибочные концепции, которые могли бы помешать программистам до конца понять, что такое дружелюбие.

ВРЕЗКА 12. ДВЕ НОВЕЙШИЕ ИДЕИ — ПРАКТИЧЕСКИ НЕЗРЕЛЫЕ, ПОЧТИ ПОЛУСЫРЫЕ

Подход, который можно назвать «Аве Мария»[22], основан на надежде, что где-то во Вселенной существуют (или вскоре возникнут) цивилизации, успешно справившиеся со взрывным развитием интеллекта и в результате пришедшие к системам ценностей, в значительной степени совпадающим с нашими. В этом случае мы можем попробовать создать свой ИИ, который будет мотивирован делать то же, что и их интеллектуальные системы. Преимущества этого подхода состоят в том, что так создать нужную мотивацию у ИИ может быть легче, чем напрямую.

Чтобы эта схема могла сработать, нашему ИИ *нет* необходимости связываться с каким-то инопланетным ИИ. Скорее, в своих действиях он должен руководствоваться *оценками* того, что тот мог бы захотеть сделать. Наш ИИ мог бы смоделировать вероятные исходы взрывного развития интеллекта где-то еще, и по мере превращения в сверхразум делать это все точнее. Идеальных знаний от него не требуется. У взрывного развития интеллекта может быть широкий диапазон возможных исходов, и нашему ИИ нужно постараться определиться с предпочтениями относительно типов сверхразума, которые могут быть связаны с ними, взвешенными на их вероятности.

В этой версии подхода «Аве Мария» требуется, чтобы мы разработали конечные ценности для нашего ИИ, согласующиеся с предпочтениями других систем сверхразума. Как это сделать, пока до конца неясно. Однако структурно сверхразумные агенты должны отличаться, чтобы мы могли написать программу, которая служила бы детектором сверхразума, анализируя модель мира, возникающую в нашем развивающемся ИИ, в поиске характерных для сверхразума элементов представления. Затем программма-детектор могла бы каким-то образом извлекать предпочтения рассматриваемого сверхразума (из его представления о нашем ИИ)[23]. Если нам удастся создать такой детектор, его можно будет использовать для определения конечных ценностей нашего ИИ. Одна из трудностей заключается в том, что нам нужно создать такой детектор раньше, чем мы будем знать, какой каркас представления разработает наш ИИ. Программа-детектор должна уметь анализировать незнакомые каркасы представления и извлекать предпочтения представленных в них систем сверхразума. Это кажется непростой задачей, но, возможно, какое-то ее решение удастся найти[24].

Если получиться реализовать основной подход, можно будет немедленно заняться его улучшением. Например, вместо того чтобы следовать предпочтениям (точнее, их некоторой взвешенной композиции) *каждого* инопланетного сверхразума, у нашего ИИ может иметься фильтр для отбора подмножества инопланетных ИИ (чтобы он мог брать пример с тех, чьи ценности совпадают с нашими). Например, в качестве критерия включения ИИ в это подмножество может использоваться источник его возникновения. Некоторые обстоятельства создания ИИ (которые мы должны уметь определить в структурных терминах) могут коррелировать с тем, в какой степени появившийся в результате ИИ может разделять наши ценности. Возможно, большее доверие у нас вызовут ИИ, первоисточником которых была полная эмуляция головного мозга, или зародыш ИИ, в котором почти не использовались эволюционные механизмы, или такие, которые возникли в результате медленного контролируемого взлета. (Если брать в расчет источник возникновения ИИ, мы также сможем избежать опасности присвоить слишком большой вес тем ИИ, которые создают множество своих копий, — а на самом деле избежать создания для них стимула делать это.) Можно также внести в этот подход множество других улучшений.

Подход «Аве Мария» подразумевает веру, что где-то существуют другие системы сверхразума, в значительной степени разделяющие наши ценности[25]. Это означает, что он неидеален.

Однако технические препятствия, стоящие на пути реализации подхода «Аве Мария», хотя и значительны, но вполне могут оказаться менее сложными, чем при других подходах. Может быть, имеет смысл изучать подходы пусть и не самые идеальные, но более простые в применении, — причем не для использования, а скорее, чтобы иметь запасной план на случай, если к нужному моменту идеальное решение не будет найдено.

Недавно Пол Кристиано предложил еще одну идею решения проблемы загрузки ценностей[26]. Как и при «Аве Марии», это метод обучения ценностям, который предполагает определение критерия ценности не при помощи трудоемкой разработки, а скорее фокусировки. В отличие от «Аве Марии», здесь не предполагается существования других сверхразумных агентов, которые мы используем в качестве ролевых моделей для нашего собственного ИИ. Предложение Кристиано с трудом поддается короткому объяснению — оно представляет собой цепочку сложных умозаключений, — но можно попытаться как минимум указать на его основные элементы.

Предположим, мы получаем: а) математически точное описание мозга конкретного человека; б) математически строго определенную виртуальную среду, содержащую идеализированный компьютер с произвольно большим объемом памяти и сверхмощным процессором. Имея а и б, можно определить функцию полезности *U* как выходной сигнал, который выдает мозг человека после взаимодействия с этой средой. *U* может быть математически строго определенным объектом, но при этом таким, который (в силу вычислительных ограничений) мы неспособны описать *конкретно*. Тем не менее *U* может служить в качестве критерия ценности при обучении ИИ системе ценностей. При этом ИИ будет использовать различные эвристики, чтобы строить вероятностные гипотезы о том, что представляет собой *U*.

Интуитивно хочется, чтобы *U* была такой функцией полезности, которую нашел бы соответствующим образом подготовленный человек, обладающий произвольно большим объемом вычислительных ресурсов, достаточным, например, для создания астрономически большого количества своих имитационных моделей, способных помогать ему в поиске функции полезности или в разработке процесса ее поиска. (Мы сейчас затронули тему конвергентного экстраполированного волеизъявления, которую подробнее рассмотрим в тринадцатой главе.)

Задача описания идеализированной среды кажется относительно простой: мы можем дать математическое описание абстрактного компьютера с произвольно большой емкостью; а также при помощи программы виртуальной реальности описать, скажем, комнату со стоящим в ней компьютерным терминалом (олицетворяющим тот самый абстрактный компьютер). Но как получить математически точное описание мозга конкретного человека? Очевидный путь — его полная эмуляция, но что если эта технология еще не доступна?

Именно в этом и проявляется ключевая инновация, предложенная Кристиано. Он говорит, что для получения математически строгого критерия цели нам не нужна пригодная для практического использования вычислительная имитационная модель

мозга, которую мы могли бы запустить. Нам нужно лишь (возможно, неявное и безнадежно сложное) ее математическое *определение* — а его получить гораздо легче. При помощи функциональной нейровизуализации и других средств измерения можно собрать гигабайты данных о связях между входными и выходными сигналами головного мозга конкретного человека. Собрав достаточное количество данных, можно создать наиболее простую имитационную математическую модель, которая учитывает все эти данные, и эта модель фактически окажется эмулятором рассматриваемого мозга. Хотя с вычислительной точки зрения нам может оказаться не под силу задача *отыскать* такую имитационную модель из имеющихся у нас данных, опираясь на них и используя математически строгие показатели сложности (например, какой-то вариант колмогоровской сложности, с которой мы познакомились во врезке 1 в первой главе), вполне реально эту модель *определить*[27].

Вариации имитационной модели

Проблема загрузки ценностей выглядит несколько иначе, если речь идет не об искусственном интеллекте, а об имитационной модели головного мозга. Во-первых, к эмуляторам неприменимы методы, предполагающие понимание процессов на нижнем уровне и контроль над алгоритмами и архитектурой. Во-вторых, имея дело с имитационными моделями головного мозга (и улучшенным биологическим разумом) можно использовать неприменимый для искусственного интеллекта метод приумножения (из общей группы методов выбора мотивации)[28].

Метод приумножения можно сочетать с техниками корректировки изначально имеющихся у системы целей. Например, можно попробовать манипулировать мотивационными состояниями эмуляторов, управляя цифровым эквивалентом психоактивных веществ (или реальных химических веществ, если речь идет о биологических системах). Сегодня уже есть возможность манипулировать целями и мотивацией при помощи лекарственных препаратов, правда, в ограниченной степени[29]. Но фармакология будущего сможет предложить лекарства с гораздо более точным и предсказуемым эффектом. Благодаря цифровой среде, в которой существуют эмуляторы, все эти действия существенно упростятся — в ней гораздо легче проводить контролируемые эксперименты и получать непосредственный доступ к любым областям цифрового мозга.

Как и при проведении опытов над живыми существами, эксперименты на имитационных моделях связаны с этическими трудностями, которые невозможно урегулировать лишь с помощью формы информированного

согласия. Подобные довольно трудноразрешимые проблемы могут перерасти в настоящие конфликты, тормозящие развитие проектов, связанных с полной эмуляцией головного мозга (скорее всего, будут введены новые этические стандарты и нормативные акты). Сильнее всего это скажется на исследованиях механизмов мотивационной структуры эмуляторов. Результат может оказаться плачевным: из-за недостаточного изучения методов контроля над возможностями имитационных моделей и методов корректировки их конечных целей когнитивные способности эмуляторов начнут неуправляемо совершенствоваться, пока не достигнут потенциально опасного сверхразумного уровня. Более того, вполне реально, что в ситуации, когда этические вопросы будут стоять особенно остро, вперед вырвутся наименее щепетильные проектные группы и государства. В то же время если мы снизим свои этические стандарты, то в процессе экспериментальной работы с оцифрованным человеческим разумом ему может быть причинен непоправимый вред, что абсолютно неприемлемо. В любом случае нам придется нести полную ответственность за собственное недобросовестное поведение и нанесенный ущерб имитационным моделям.

При прочих равных условиях соображения этического порядка, скорее всего, заставят нас отказаться от проведения опытов над цифровыми копиями людей. В такой критическо-стратегической ситуации мы будем вынуждены искать альтернативные пути, не требующие столь активного изучения биологического мозга.

Однако не все так однозначно. Готов выслушать ваши возражения: исследования, связанные с полной эмуляцией головного мозга, с *меньшей* вероятностью будут вызывать этические проблемы, чем разработки в области искусственного интеллекта, на том основании, что нам легче проследить момент становления, когда право на моральный статус начнет обретать именно эмулятор, а не совершенно чужеродный искусственный разум. Если ИИ определенного типа или какие-то его подпроцессы обретут значительный моральный статус прежде, чем мы это распознаем, этические последствия могут быть огромными. Возьмем, например, невероятную легкость, с которой современные программисты создают агентов для обучения с подкреплением и применяют к ним негативные раздражители. Ежедневно создается бесчисленное количество таких агентов, причем не только в научных лабораториях, но и в многочисленных фирмах, где разрабатываются разные приложения и создаются компьютерные игры, содержащие множество

сложных персонажей. Предположительно, эти агенты еще слишком примитивны, чтобы претендовать на какой-то моральный статус. Но можем ли мы быть уверены на все сто процентов? И еще одно важное замечание: можем мы быть уверены, что узнаем, в какой момент следует остановиться, чтобы программы не начали испытывать страдания?

(В четырнадцатой главе мы вернемся к некоторым более общим стратегическим вопросам, которые возникают при сравнении двух процессов: проведения полной эмуляции головного мозга и создания искусственного интеллекта.)

Институциональное конструирование

Существуют интеллектуальные системы, чьи составляющие сами являются агентами, обладающими интеллектом. В нашем, пока еще человеческом, мире примерами таких систем являются государства и корпорации — они состоят из людей, но в отдельных случаях сами институты могут рассматриваться как самостоятельные, функционально независимые агенты. Мотивация такой сложной системы, как учреждение, зависит не только от мотивов составляющих ее субагентов, но и от того, как эти субагенты организованы. Например, институциональная система диктаторского типа может вести себя так, словно обладает волей, аналогичной воле одного-единственного субагента, исполняющего роль диктатора, а институциональная система демократического типа, напротив, ведет себя так, как будто аккумулирует в себе интересы всех субагентов и выражает совокупную волю всех участников. Однако можно представить такие институты управления, при которых организация не выражает совокупные интересы составляющих ее субагентов. (Теоретически вполне возможно существование тоталитарного государства, дружно ненавидимое всем его населением, поскольку властная структура обладает мощным аппаратом подавления, не допускающим даже мысли о каком бы то ни было гражданском противостоянии — ни о скоординированном восстании, ни об отдельных протестах. В итоге гражданам, не имеющим права ни на всеобщее, ни на одиночное возмущение, остается лишь выполнять функцию винтиков государственной машины.)

Таким образом, создавая соответствующие институты для сложных систем, можно предринять попытки сразу формировать эффективные системы мотивации. В девятой главе мы обсуждали социальную интеграцию как

один из вариантов метода контроля над возможностями. Теперь нам надо рассмотреть вопрос социальной интеграции с точки зрения стимулов, с которыми сталкивается агент, существующий в социальном мире равных ему субъектов. Мы сосредоточим внимание на том, что происходит *внутри* конкретного агента: каким образом его воля определяется его внутренней организацией. Поскольку устройство института такого рода не зависит от крупномасштабного социального инжиниринга или реформ, метод выбора мотивации применим в условиях отдельного проекта создания сверхразума, даже если социоэкономическая среда и международная обстановка не самые благоприятные.

Вероятно, правильнее всего было бы использовать институциональное конструирование в сочетании с методом приумножения. Если мы можем начать с агентов, уже обладающих требуемой мотивацией или даже аналогичной человеческой, то институциональные механизмы и меры предосторожности повысят гарантии, что система не свернет с правильного пути.

Предположим, мы начали с некоторых хорошо мотивированных агентов человеческого типа, например имитационных моделей. Нам требуется повысить когнитивные возможности этих агентов, но нас беспокоит, что совершенствование может нарушить их мотивационную систему. Один из способов справиться с этой проблемой — создать что-то типа организации, в которой отдельные эмуляторы действуют как субагенты. Каждая последующая процедура усовершенствования будет применяться по отношению лишь к небольшой части таких субагентов, а его влияние — оцениваться путем сравнения их поведения с поведением контрольной группы субагентов, не подвергавшихся процедуре улучшения. Когда подтвердится, что совершенствование когнитивных способностей эмуляторов не вредит их мотивации, то процедуру можно начать применять ко всей популяции субагентов. Если выяснится, что система мотивации улучшенных субагентов пострадала, они выключаются из дальнейших планов усовершенствования и не допускаются к процессу принятия основных решений (минимум до тех пор, пока организационная система в целом не разовьется до такой степени, что сможет безопасно их реинтегрировать)[30]. Хотя субагенты с пострадавшей мотивацией могли бы в результате улучшения получить определенное преимущество, они не в состоянии захватить власть в организационной системе или «заразить» своей погрешностью всех остальных, поскольку представляют собой небольшое меньшинство всего сообщества субагентов. То есть

коллективный интеллект и возможности системы будут постепенно повышаться в результате последовательных небольших шагов, притом что правильность каждого шага проверяется субагентами, обладающими не намного меньшими способностями, чем субагенты, появляющиеся в результате очередного шага.

Степень безопасности, которая достигается за счет такого институционального конструирования, ограничена точностью тестов, применяющихся для оценки улучшенных субагентов. Разработка тестов усложняется тем, что оцениваемые параметры могут быть проверены лишь на небольшом диапазоне когнитивных способностей, а применяться они должны к агентам в том числе за пределами этого диапазона. Можно отчасти снять эту проблему, если проводить усовершенствование последовательно, шаг за шагом, и проверку на протяжении длительного времени. Однако такие меры предосторожности повысят издержки и замедлят прогресс (а в случае конкурентной гонки это будет означать, что проект, принимающий такие меры, окажется в невыгодном положении). И даже если идти вперед небольшими шагами и проводить интенсивное тестирование, остается некоторый риск появления небольших дефектов, которые постепенно приведут к деградации всю систему целей — деградации, способной проявиться лишь на этапе достижения системой уровня сверхразума. Риск можно несколько снизить, расширив условие тестирования на множество шагов улучшений, чтобы даже те агенты, которые успешно прошли проверку на каком-то этапе и были подвергнуты последующим улучшениям, продолжали бы сравниваться с контрольной группой немодифицированных агентов с изначально правильной системой мотивации.

В идеале такое тестирование должно проходить непрерывно, а не ограничиваться первоначальным периодом оценки. У субагентов человеческого уровня имеются стратегические способности, и они могут скрывать некоторые цели в ходе контрольных проверок их поведения. Чтобы сделать возможным непрерывное тестирование, можно создать иерархию, в рамках которой одни субагенты выполняли задачу по контролю над другими субагентами с несколько более высоким уровнем способностей. На нижней иерархической ступеньке способностей (но на вершине властной иерархии) окажется самый недалекий и «заторможенный» принципал. Он будет напоминать слабоумного короля, который отдает распоряжение недееспособному двору, а тот, в свою очередь, контролирует посредственное правительство,

управляющее талантливыми людьми. Еще его можно сравнить с древним «мозгом рептилий», который следит за тем, чтобы более молодой и гораздо более развитый «неокортекс» отвечал за выполнение таких функций, как питание и совокупление.

Может возникнуть вопрос, а будет ли такая перевернутая меритократия стабильной? Если окажется, что мотивация значительного количества сверхразумных агентов нарушена, и люди, отвечающие за проект в целом, попытаются «откатить» часть улучшений, окажется ли эта институциональная структура достаточно прочной для того, чтобы выдержать сопротивление, вызванное этим шагом? В контексте исторических и современных человеческих обществ ответ, скорее всего, будет отрицательным. (Нельзя сказать, что к власти приходят только самые способные люди, но и сказать, что нами *в прямом смысле* управляют шимпанзе, тоже нельзя.) Однако этот вопрос оказывается гораздо более сложным, когда общество представляет собой совокупность компьютерных программ[31].

Представьте цифровую иерархию, в которой на каждых двух пролетариев приходится по одному хорошо обученному полицейскому агенту. А на каждых двух обычных полицейских агентов — один специальный полицейский агент. Еще есть специальный специальный полицейский агент на каждых двух просто специальных — и так далее, так что над каждым уровнем контролеров имеется уровень с как минимум в два раза меньшим количеством контролеров. Представьте также структуру этого общества, при которой контролер обладает большими полномочиями в отношении подконтрольных ему агентов. Например, может видеть и записывать все, что они говорят и делают, как если был бы скрыт от них зеркальным стеклом — из-за этого стекла подконтрольные агенты получают лишь приказы и инструкции, которые отбирал сам контролер. Еще нужно представить, что перед каждым контролером есть приборная панель с кнопками, нажимая которые он может в любой момент соответственно наказать, вознаградить, поставить на паузу, удалить или вернуть в предыдущее состояние своего подконтрольного агента. Благодаря этим возможностям власть агента-контролера выходит далеко за исторические рамки того, что встречалось даже в наиболее тоталитарных организациях.

Но даже этим инструменты контроля не исчерпываются. В распоряжении контролеров могут иметься реалистичные симуляторы виртуальной реальности, которые они могут использовать для моделирования гипотетических

сценариев, чтобы провоцировать подконтрольных агентов и изучать их реакции (например, предлагать принять участие в заговоре или немного расслабиться, уверяя, что их никто не видит). Кроме того, контролеры могут «влезать в умы» подконтрольных агентов, подслушивая их внутренние монологи и напрямую манипулируя их эмоциональным состоянием.

То есть у контролера имеется огромное преимущество[32]. Но каждый контролер является подконтрольным агентом другого контролера более высокого уровня, который, в свою очередь, наблюдает за каждым *его* шагом и следит, чтобы *он* не ослаблял усилий в деле исполнения своих контрольных функций. Таким образом, продуманная структура может быть стабильной даже в том случае, когда интеллектуально контролеры на всех уровнях иерархии слабее тех, кого контролируют[33]. И всю эту структуру, включающую множество сверхразумных агентов, может контролировать небольшое количество людей с административным доступом[34].

Издержки, связанные с такой структурой внутреннего контроля, могут быть значительными, но не запредельно высокими. Например, в модели, где на каждом уровне членов в два раза меньше, чем уровнем ниже, издержки на дополнительные вычисления ограничены всего лишь ста процентами от того, что требуется для «пролетарской» части системы, — или даже меньшей величиной, если более глупый босс требует меньше вычислений в расчете на субагента. За счет увеличения этого показателя можно усилить контроль, за счет понижения — уменьшить издержки. Дополнительные вычислительные издержки будут главным фактором при сценарии конкурентного рыночного мира (см. главу 11), но в контексте проекта, не имеющего сильных непосредственных конкурентов, их роль довольно умеренная. Также потребуются затраты времени разработчиков на создание и тестирование всего этого контрольного функционала. Их можно снизить, если в рамках обладающего большими ресурсами проекта распараллелить работу над контрольной структурой и самим искусственным интеллектом; но их тяжесть будет непомерной при небольших проектах и проектах, участвующих в напряженной технологической гонке.

Также заслуживает внимания еще один тип издержек в структурах такого рода — риск проявления преступной безнравственности[35]. Описанная нами институция выглядит как страшный северокорейский трудовой лагерь. Однако возможности ослабить этические проблемы, связанные с его работой, есть даже в том случае, когда составляющие его субагенты являются

эмуляторами с моральным статусом, соответствующим человеческому. В самом крайнем случае он может быть основан на добровольном участии в нем. Причем у каждого субагента должно быть право в любой момент прекратить свое участие[36]. Стертые имитационные модели могут храниться в памяти с обязательством восстановить их в более подходящих условиях, когда минует опасная стадия взрывного развития интеллекта. Тем временем субагенты, решившие участвовать в системе, могут размещаться в очень комфортабельных виртуальных условиях и иметь достаточно времени для сна и отдыха. Эти меры также предполагают затраты, которые, однако, вполне по силам проекту, обладающему большими ресурсами и не имеющему прямых конкурентов. Но в высококонкурентной среде эти расходы могут быть неприемлемыми — утешит лишь уверенность, что конкуренты их тоже несут.

В нашем примере мы предположили, что субагенты являются эмуляторами, то есть имитационными моделями головного мозга человека. Может возникнуть вопрос: потребует ли метод институционального конструирования, чтобы субагенты были антропоморфными? Или он равноприменим к системам, состоящим из искусственных субагентов?

Возможный скепсис в этом вопросе понятен. Известно, что несмотря на весь наш огромный опыт наблюдения за агентами-людьми, мы до сих пор не в состоянии предсказывать начало и исход революций: социальные науки могут в лучшем случае описать некоторые их статистические закономерности[37]. А поскольку мы не можем надежно предсказывать стабильность социальных структур, состоящих из обычных человеческих существ (о которых знаем так много), возникает соблазн заключить, что у нас нет надежды выстроить стабильные социальные структуры для когнитивно улучшенных человекоподобных агентов (о которых мы не знаем ничего), и тем более для ИИ-агентов (которые даже не похожи на агентов, о которых мы что-то знаем).

Однако все не так уж плохо. Люди и человекоподобные субъекты чрезвычайно сложны, в то время как искусственные агенты могут иметь сравнительно простую архитектуру. У искусственных агентов также может быть простая и явно задаваемая мотивация. Более того, цифровые агенты в целом (и эмуляторы и ИИ) поддаются копированию: это преимущество способно вызвать революцию в управлении, как взаимозаменяемые комплектующие вызвали революцию в производстве. Эти отличия в сочетании с возможностью работать с агентами, которые вначале бессильны, и создавать

институциональные структуры, в которых используются перечисленные выше методы контроля, могут сделать возможным получение нужного институционального результата — например, системы, в которой не будет революций, — причем с большей вероятностью, чем в случае с людьми.

Впрочем, нужно сказать, что у искусственных агентов могут отсутствовать многие свойства, знание которых позволяет нам прогнозировать поведение человекоподобных агентов. Им не нужно иметь никаких социальных эмоций, которые определяют человеческое поведение, таких как страх, гордость и угрызения совести. Им не нужны дружественные и семейные связи. Им не нужен «язык тела», который не позволяет нам, людям, скрыть свои намерения. Эти факторы могут дестабилизировать организации, состоящие из искусственных агентов. Более того, такие агенты способны совершать большие скачки в когнитивной производительности в результате внешне незначительных изменений в их алгоритмах или архитектуре. Безжалостно оптимальные искусственные агенты будут готовы пускаться в такие рискованные авантюры, результатом которых может стать сокращение размеров человечества[38]. А еще агенты, обладающие сверхразумом, смогут удивить нас способностью координировать свои действия, почти или совсем не связываясь друг с другом (например, посредством внутреннего моделирования гипотетической реакции партнеров на различные обстоятельства).

Эти и другие особенности повышают вероятность внезапного краха организации, состоящей из искусственных агентов, невзирая даже на, казалось бы, пуленепробиваемые методы социального контроля.

Итак, пока неясно, насколько многообещающим является метод институционального конструирования и будет ли он более эффективным в случае антропоморфных, нежели искусственных, агентов. Может показаться, что создание института с адекватной системой сдержек и противовесов повысит нашу безопасность — или по крайней мере не снизит ее, — поэтому с точки зрения снижения рисков данный метод лучше применять всегда. Но на самом деле даже это нельзя сказать с определенностью. Использование метода повышает сложность системы, создавая тем самым новые возможности для неблагоприятного развития ситуации, которые отсутствуют в случае агентов, не имеющих в качестве составляющих интеллектуальных субагентов. Тем не менее метод институционального конструирования заслуживает дальнейшего изучения[39].

Резюме

Инжиниринг системы целей — еще не установленная дисциплина. Пока нет полной ясности в том, как загружать в компьютер человеческие ценности, даже если речь идет о машинном интеллекте человеческого уровня. Изучив множество подходов, мы обнаружили, что некоторые из них, похоже, ведут в тупик, но есть и такие, которые кажутся многообещающими и должны стать предметом дальнейшего анализа. Обобщим изученный материал в табл. 12.

Таблица 12. Обобщение методов загрузки ценностей

Представление в явной форме	Кажется многообещающим в качестве способа загрузки ценностей при использовании метода приручения. Вряд ли полезен в случае более сложных целей
Естественный отбор	Менее перспективный. Полным перебором можно обнаружить схемы, удовлетворяющие формальному критерию поиска, но не соответствующие нашим намерениям. Более того, если варианты схем оценивать путем их реализации — включая те, которые не удовлетворяют даже формальному критерию, — резко повышаются риски. В случае применения метода естественного отбора сложнее избежать преступной безнравственности, особенно если мозг агентов похож на человеческий
Обучение с подкреплением	Для решения задачи обучения с подкреплением могут использоваться различные методы, но обычно это происходит путем создания системы, которая стремится максимизировать сигнал о вознаграждении. По мере развития интеллекта таких систем у них проявляется внутренне присущая им тенденция отказа по типу самостимулирования. Методы обучения с подкреплением не кажутся перспективными
Модель ценностного приращения	Человек получает большую часть информации о своих конкретных целях благодаря обогащенному опыту. И хотя, в принципе, метод ценностного приращения может использоваться для создания агента с человеческой мотивацией, присущие людям особенности приращения целей слишком сложно воспроизводить, если начинаешь работу с зародыша ИИ. Неверная аппроксимация способна привести к тому, что ИИ будет обобщать информацию не так, как люди, вследствие чего приобретет не те конечные цели, которые предполагались. Чтобы определить с достаточной точностью, насколько трудна может оказаться работа по ценностному приращению, требуются дополнительные исследования

Строительные леса для мотивационной системы	Пока рано говорить, насколько трудно будет добиться от системы выработки внутренних представлений высокого уровня, прозрачных для людей (и при этом удержать возможности системы на безопасном уровне), чтобы при помощи таких представлений создать новую систему ценностей. Метод кажется очень перспективным. (Но поскольку в этом случае, как при любом неопробованном методе, большая часть работы по созданию системы безопасности откладывается до момента появления ИИЧУ, нельзя допустить, чтобы это стало оправданием для игнорирования проблемы контроля в течение всего времени, предшествующего этому моменту.)
Обучение ценностям	Потенциально многообещающий подход, но нужно провести дополнительные исследования, чтобы определить, насколько трудно будет формально определить ссылки на важную внешнюю информацию о человеческих ценностях (и насколько трудно при помощи такой ссылки задать критерий правильности для функции полезности). В рамках этого подхода стоят пристального изучения предложения вроде метода «Аве Мария» и конструкции Пола Кристиано
Эмуляторы и цифровые модуляции	Если машинный интеллект создан в результате полной эмуляции головного мозга, скорее всего, будет возможно корректировать его мотивацию при помощи цифрового эквивалента лекарственных препаратов или иных средств. Позволит ли это загрузить цели с достаточной точностью, чтобы обеспечить безопасность даже в случае превращения эмулятора в сверхразум, — вопрос пока открытый. (Повлиять на развитие процесса могут этические ограничения.)
Институциональное конструирование	К организациям, состоящим из эмуляторов, применимы различные сильные методы контроля над возможностями, в том числе социальная интеграция. В принципе, такие методы могут быть использованы и для организаций, члены которых являются системами ИИ. Эмуляторы обладают одним набором свойств, которые облегчают проведение контроля над ними, и другим набором свойств, которые затрудняют проведение контроля над ними по сравнению с ИИ. Институциональное конструирование стоит дальнейшего исследования как потенциально полезная техника метода загрузки ценностей

Когда мы поймем, как решить проблему загрузки ценностей, то немедленно столкнемся со следующей — как решать, какие ценности надо загружать. Иными словами, есть ли у нас сложившееся мнение, что должен был бы желать сверхразум? Это вопрос почти философский, и мы к нему сейчас обратимся.

Глава тринадцатая

Выбор критериев выбора

Предположим, мы можем назначить зародышу ИИ любую конечную ценность. Тогда принятое нами решение — какая это должна быть цель — имеет далекоидущие последствия. Нам придется определить некоторые параметры выбора, связанные с аксиомами теории принятия решений и эпистемологии. Но откуда черпать уверенность, будто мы, люди — существа недалекие, невежественные и ограниченные, — можем принимать правильные решения по этому вопросу? Сможем ли мы сделать выбор, не предопределенный предрассудками и предубеждениями нашего поколения? В этой главе мы рассмотрим, как с помощью косвенной нормативности переложить большую часть умственной работы, связанной с принятием этих решений, на сам сверхразум, но при этом не забывая, что точкой отсчета всегда должны быть фундаментальные ценности человеческой жизни.

Необходимость в косвенной нормативности

Как заставить ИИ делать то, что мы хотим? Что мы хотим, чтобы хотел ИИ? До этого момента мы искали ответ на первый вопрос. Теперь пришло время обратиться ко второму.

Предположим, что мы решили проблемы контроля и теперь способны загрузить в мотивационную систему сверхразума любую ценность и убедить его считать ее своей конечной целью. Какую ценность нам все-таки следует выбрать? Выбор непрост. Если сверхразум обретет решающее стратегическое преимущество, именно его система ценностей начнет определять судьбу всего вселенского пространства.

Поэтому так важно не допускать ошибок при выборе цели. Но реально ли в подобных делах всерьез надеяться на безошибочность принятых решений? Мы можем заблуждаться относительно нравственных ценностей; не понять, что есть благо для нас самих; промахнуться даже в собственных

желаниях. Похоже, в поисках конечной цели придется продираться сквозь колючие заросли философских проблем. Если пойти прямым путем, можно наломать дров. Особенно риск неверного выбора велик в том случае, когда имеешь дело с незнакомым контекстом принятия решения. Ведь выбор конечной цели для машинного сверхразума — цели, от которой зависит будущее всего человечества, — видимо, из всех возможных сюжетов для нас это наиболее неведомый.

Скорее всего, у нас мало шансов победить в лобовой атаке, что подтверждается отсутствием среди специалистов полного согласия по проблемам, касающимся человеческих систем ценностей. Ни одна этическая теория не получила признания большинства философов, таким образом, можно считать, что большинство неправо[1]. Об этом свидетельствует и постоянное изменение этических норм — изменение, связанное с ходом времени, что мы традиционно воспринимаем как свидетельство прогресса. Например, в средневековой Европе считалось вполне респектабельным развлечением наблюдать за пытками и казнями противников власти. В Париже XVI века популярным действом было сжигание кошек[2]. Всего сто пятьдесят лет назад на американском Юге широко практиковалось рабство, причем при полной поддержке закона и в полном согласии с моральными нормами поведения. Оглядываясь назад, видишь вопиющие нарушения не только в поступках, но и в этических воззрениях людей, живших в прежние времена. Вероятно, с тех пор мы чему-то научились в вопросах этики, хотя вряд ли можно утверждать, что наша мораль поднялась на недосягаемую высоту. Вполне вероятно, что какие-то этические концепции, которыми мы руководствуемся сегодня, имеют серьезные недостатки. При таких обстоятельствах выбирать конечную ценность, основанную на наших нынешних убеждениях, да еще так, чтобы исключить любую возможность дальнейшего развития этической системы, означало бы брать на себя ответственность за будущий риск, ведущий к экзистенциальному разрушению моральных норм.

Даже если у нас существовала бы рациональная уверенность, что мы обнаружили правильную этическую теорию — которой у нас нет, — попрежнему оставался бы риск совершения ошибок при ее детальной проработке. У внешне простых этических теорий может иметься множество скрытых противоречий[3]. Рассмотрим, например, консеквенциалистскую теорию гедонизма (кстати, необыкновенно простую). Если совсем кратко, то она о том, что удовольствие — это ценность, а страдание — нет[4]. Даже

если мы поставим все наши моральные фишки на эту теорию и она окажется верной, останутся открытыми множество важных вопросов. Следует ли присваивать более высокий приоритет «высшим удовольствиям» по сравнению с «низшими» по примеру Джона Стюарта Милля? Как учитывать интенсивность и продолжительность удовольствия? Могут ли страдание и удовольствие взаимно исключать друг друга? Какие типы мышления ассоциируются с морально допустимыми удовольствиями?[5] Увеличится ли в два раза количество удовольствия в результате появления двух точных копий одного такого типа разума?[6] Существуют ли подсознательные удовольствия? Как быть с чрезвычайно низкими шансами хоть когда-нибудь получить предельное удовольствие? Как объединить удовольствие членов неограниченной популяции в одно целое?[7]

Неверный ответ на любой из этих вопросов приведет к катастрофе. То есть при выборе конечной ценности для сверхразума, в попытках нащупать хоть какое-то действенное решение, нам придется по-крупному ставить не только на этическую теорию в ее цельном состоянии, но и на частные особенности, на многочисленные интерпретации, всевозможные вкрапления и исключения — при таком обороте наши шансы на точный выстрел, кажется, начнут стремиться к нулю. Глупцы с радостью ухватятся за возможность одним махом решить все важные этические проблемы, а найденные ими удобные ответы сразу загрузить в зародыш ИИ. Мудрецы отправятся в трудный путь поисков альтернативных подходов и разыскивания способов подстраховки.

Все эти размышления подводят нас к варианту косвенной нормативности. Объективная причина создания сверхразума заключается в том, что на него можно переложить инструментальную задачу поиска эффективных путей достижения той или иной ценности. А за счет косвенной нормативности мы можем буквально свалить на него и сам выбор этой системы ценностей.

Косвенная нормативность позволяет решить проблему незнания того, что мы на самом деле хотим, что входит в наши интересы, что является моральным правом или идеалом. Вместо размышлений, основанных на сегодняшнем понимании (которое, вероятно, глубоко ошибочно), можно делегировать некоторую часть когнитивной работы по выбору системы ценностей самому сверхразуму. Он, несомненно, сможет выявить и ошибки и противоречия, искажающие наши представления, поскольку с такой

работой лучше него не справится никто. Можно обобщить эту идею и выразить ее в качестве эвристического принципа.

> *Принцип эпистемологического превосходства*
> Будущий сверхразум занимает эпистемологически более высокий наблюдательный пункт: его убеждения (видимо, относительно большинства вопросов) с большей вероятностью окажутся истинными, чем наши. Поэтому при любых возможных обстоятельствах следует полагаться на его мнение[8].

Косвенная нормативность позволяет применить этот принцип к проблеме выбора системы ценностей. Будучи неуверенными в своей способности задать конкретный нормативный стандарт, мы можем определить какое-то более абстрактное условие, которому должен удовлетворять любой нормативный стандарт, в надежде, что сверхразум справится сам и отыщет конкретный стандарт, удовлетворяющий этому абстрактному условию. А затем мы поставим перед зародышем ИИ его ценностную конечную цель: вести себя в соответствии со своими представлениями о правильных действиях, основанных на этом стандарте, определенном косвенным образом.

Прояснить эту идею нам помогут несколько примеров. Вначале рассмотрим модель косвенной нормативности, предложенную Элиезером Юдковским, — когерентное экстраполированное волеизъявление. Затем разберем несколько вариантов и альтернатив этой модели, чтобы составить представление о диапазоне возможных решений.

Когерентное экстраполированное волеизъявление

Юдковский предложил, что зародышу ИИ следует задать в качестве конечной цели следование когерентному экстраполированному волеизъявлению (далее по тексту — КЭВ) человечества, которое он определял так:

> Наше когерентное экстраполированное волеизъявление — это наше желание знать больше; думать быстрее; быть в большей степени людьми, которыми нам хотелось бы быть; стать ближе друг к другу; сделать так, чтобы наши мысли сближали нас, а не разделяли, чтобы наши желания совпадали, а не пересекались; экстраполировать так, как нам хотелось бы экстраполировать; понимать так, как нам хотелось бы понимать[9].

Когда Юдковский писал это, он не ставил перед собой задачу создать инструкцию по воплощению в жизнь своего предписания, более

напоминающего поэтическое воззвание. Его целью было набросать эскиз того, как могло бы быть определено КЭВ, а также пояснить, зачем нужен именно этот подход.

Многие идеи, лежащие в основе КЭВ, имеют аналоги и предшественников в философской литературе. Например, в этике существует *теория идеального наблюдателя*, которая исследует понятия (например, «хороший» и «плохой») с точки зрения тех суждений, которые сделал бы гипотетический идеальный наблюдатель (под таковым понимается всеведущий, логически мыслящий, беспристрастный и свободный от любой предвзятости субъект)[10]. Однако модель КЭВ не является (и не должна считаться) этической теорией. Никто не утверждает, что есть связь между целью и нашим когерентным экстраполированным волеизъявлением. КЭВ можно считать полезным способом аппроксимации всего, что имеет конечную цель без какой-либо связи с этикой. Будучи основным прототипом метода косвенной нормативности, КЭВ заслуживает более подробного изучения.

Некоторые комментарии

Отдельные термины из приведенной выше цитаты требуют пояснения. Желание «думать быстрее» в понимании Юдковского означает стремление *быть умнее и глубже проникать в суть вещей*. «Стать ближе друг к другу» — видимо, *учиться, развиваться и самосовершенствоваться в тесной связи друг с другом*.

Требуют своего объяснения некоторые фразы.

«…Чтобы наши мысли сближали нас, а не разделяли…»

ИИ следует работать над тем или иным свойством результата своих размышлений только в той степени, в какой это свойство может быть предсказано им с высокой долей вероятности. Если он неспособен предсказать, что «идеальные мы» желали бы это свойство, ему следует отказаться от реализации своих фантазий и воздержаться от действий. Однако, несмотря на то что многие детали наших идеализированных желаний могут быть неопределенными или непредсказуемыми, есть некие общие рамки наших предпочтений, которые ИИ способен осознать и хотя бы минимально стремиться к тому, чтобы события в будущем развивались в границах этого. Например, если ИИ может уверенно сказать, что наше КЭВ не имеет ничего общего с желанием пребывать в состоянии постоянной агонии или увидеть

Вселенную, превращенную в скрепки, то должен действовать так, чтобы не допустить подобных исходов[11].

«…Чтобы наши желания совпадали, а не пересекались…»

ИИ следует действовать в соответствии с довольно широким консенсусом экстраполированных волеизъявлений отдельных людей. Небольшое количество сильных, ясно выраженных желаний способно иногда перевесить слабые и невнятные желания большинства. Также Юдковский считает, что для ИИ требуется меньший консенсус, чтобы *предотвратить* некий конкретный негативный исход, и больший, чтобы действовать с целью реализации некоего конкретного позитивного исхода. «Исходным принципом для КЭВ должен быть консервативный подход к „да“ и внимательное отношение к „нет“», — пишет он[12].

«Экстраполировать так, как нам хотелось бы экстраполировать; понимать так, как нам хотелось бы понимать…»

Идея, лежащая в основе этих последних модификаторов, похоже, заключается в том, что правила экстраполяции сами должны учитывать экстраполированное волеизъявление. Индивидуум может иметь желание второго порядка (желание относительно того, что желать), чтобы некоторые его желания первого порядка не имели веса при экстраполяции его волеизъявления. Точно так же у нас могут быть желания относительно того, как должен развиваться процесс экстраполирования, и все это должно быть принято во внимание.

Можно возразить, что если удастся правильно определить понятие когерентного экстраполированного волеизъявления человечества, все равно окажется невозможным — даже для сверхразума — выяснить, что человечество хотело бы в гипотетических идеальных обстоятельствах, предусмотренных методом КЭВ. Если у ИИ не будет никакой информации о содержании нашего экстраполированного волеизъявления, в его распоряжении не останется никаких зацепок, которыми он мог бы руководствоваться в своем поведении. Хотя точно узнать КЭВ человечества действительно трудно, однако вполне возможно сформировать о нем информированное суждение. Причем возможно уже сегодня, не имея под рукой машинного сверхразума. Например, наше КЭВ видит людей будущего как людей, скорее живущих богато и счастливо, чем испытывающих

невыносимые страдания. Если мы способны делать такие разумные предположения, то сверхразум тем более справится. То есть с самого начала поведение сверхразума может определяться его оценками относительно содержания нашего КЭВ. У него будут сильные инструментальные причины уточнять эти первоначальные смыслы (например, изучая человеческую культуру и психологию, сканируя мозг людей и размышляя, каким образом мы поступали бы, если знали бы больше, думали бы глубже и так далее). В своих исследованиях сверхразум руководствовался бы собственными первоначальными оценками КЭВ. Поэтому он не станет проводить бесчисленные опыты над имитационными моделями, сопровождающиеся безмерными страданиями этих сущностей, зная, что наше КЭВ сочтет такие эксперименты преступной безнравственностью.

Приведем еще одно возражение: в мире существует такое разнообразие образов жизни и сводов этических норм, что вряд ли получится объединить их в единую систему КЭВ. Даже если удастся это сделать, результат может быть не особенно аппетитным — маловероятно сделать что-то съедобное из сваленных в одну тарелку лучших кусочков любимейших блюд всех людей и народов[13]. Ответ здесь прост: метод КЭВ не предполагает смешивать воедино все формы жизнедеятельности, все мировоззрения, этические нормы и личностные ценности человека. КЭВ по определению работает лишь в том случае, когда наши волеизъявления когерентны. Если разногласие между ними нарастает, несмотря на перебор различных идеальных условий, процесс должен воздержаться от определения результата. Продолжим кулинарную аналогию: хотя у людей различных культур могут быть разные любимые блюда, тем не менее люди способны достичь согласия в том, что еда не должна быть токсичной. То есть КЭВ могло бы действовать с общей целью не допустить токсичности еды, а в остальном люди совершенствовали бы свое кулинарное мастерство без его руководства и вмешательства.

Целесообразность КЭВ

Юдковский приводит семь аргументов в пользу метода КЭВ. Три из них, по сути, говорят, что даже при наличии гуманной и полезной цели может оказаться довольно трудно определить и явно выразить набор правил, не имеющих ненамеренных интерпретаций и нежелательных следствий[14]. Метод КЭВ видится его автору строгим и способным к самокоррекции, предполагается, что КЭВ обращается к *источникам* наших целей, вместо

того чтобы полагаться на нашу способность перечислить и раз и навсегда правильно сформулировать самые существенные из них.

Оставшиеся четыре аргумента выходят за рамки фундаментальных (и важнейших) вопросов. Они представляют собой конкретные пожелания к потенциальным решениям проблемы: как давать определения системам ценностей, — предполагая, что КЭВ удовлетворит этим пожеланиям.

«Воплощать совершенствование этических норм»

Пожелание того, чтобы решение допускало возможность прогресса в морально-нравственных вопросах. Мы предполагаем, что наши нынешние этические нормы могут иметь множество недостатков, и, скорее всего, серьезных недостатков. Если мы сформулировали бы для ИИ список конкретных и неизменных этических норм, которым он должен следовать, мы фактически зафиксировали бы нынешнее состояние моральных установок со всеми их ошибками, что убило бы всякую надежду на их совершенствование. Напротив, метод КЭВ оставляет возможность такого роста, поскольку позволяет ИИ пробовать делать то, что мы сами сделали бы, если продолжили бы развиваться в благоприятных условиях, и вполне возможно, что в этом случае нам удалось бы избавиться от имеющихся недостатков и ограничений своих этических норм.

«Избегать взлома судьбы человечества»

Юдковский имеет в виду сценарий, по которому небольшая группа программистов создает зародыш ИИ, превращающийся в сверхразум и обретающий абсолютное стратегическое преимущество. То есть в руках программистов сосредоточена судьба как человечества, так и всего космического пространства. Несомненно, это накладывает колоссальную ответственность, которая может обернуться непосильным бременем для любого смертного. Причем когда программисты окажутся в такой ситуации, то полностью уклониться от ответственности им уже не удастся, так как любой сделанный ими выбор, включая отказ от дальнейшего осуществления проекта, будет иметь исторические последствия для мира. Юдковский считает, что КЭВ не даст создателям ИИ возможности получить привилегию и позволит избежать непомерной тяжести определять судьбу всего мира. Запустив процесс, который руководствуется когерентным экстраполированным волеизъявлением *человечества* — в противовес их собственной воле или предпочитаемой

этической теории, — программисты фактически разделяют с другими свою ответственность за влияние на будущее человеческой цивилизации.

«Избегать создания мотивов для борьбы за исходные параметры системы»

Разделение со всем человечеством ответственности за его будущее не только предпочтительнее с точки зрения этики по сравнению с фиксацией собственных предпочтений, но и позволяет группе программистов снизить вероятность конфликта между создателями первого сверхразума. Метод КЭВ предполагает, что у программистов (и у организаторов проекта) не больше влияния на результат его действий, чем у любого другого человека — хотя они, конечно, играют главную причинную роль в определении структуры экстраполирования и в решении применить КЭВ, а не какой-то иной метод. Избегать конфликта важно не только из-за его непосредственного вреда, но и потому что он подрывает возможности сотрудничества в ходе решения важнейших проблем, связанных с безопасностью и эффективностью развития сверхразума.

Предполагается, что КЭВ способно получить широкую поддержку. И не только из-за того, что равномерно распределяет ответственность. Этот метод обладает большим миротворческим потенциалом, поскольку позволяет многим человеческим сообществам надеяться на реализацию предпочтительного для них варианта будущего. Представим афганского талиба, дискутирующего с членом Шведской гуманистической ассоциации. Не надо объяснять, насколько разные их мировоззренческие позиции: что для одного — утопия, для другого — антиутопия. Причем ни одного из них может не устроить даже компромиссное решение, например: разрешить девочкам получать образование, но только до девятого класса; или шведским девочкам разрешить получать образование, а афганским — нет. Однако и талиб, и гуманист могут согласиться с тем, что грядущее человечества будет определяться его когерентным экстраполированным волеизъявлением. Талиб может рассуждать следующим образом: если его религиозные взгляды истинны (а в этом он убежден) и если имеются серьезные основания для их приятия (в чем он также убежден), тогда человечество в итоге их признает — только людям надо стать менее предвзятыми, тратить больше сил на изучение священных текстов, уразуметь наконец, как устроен мир, правильно расставлять приоритеты, избавиться от иррациональной воинственности и малодушия, ну и так далее[15]. Гуманист

точно так же решит, что в этих идеальных условиях человечество в конечном счете станет руководствоваться принципами, которые поддерживает он.

«Ответственность за судьбу человечества должна лежать на нем самом»

Нас может не устроить исход, при котором патерналистски настроенный сверхразум начнет постоянно следить за нами и управлять нашей жизнью на всех ее уровнях, пытаясь каждую мелочь оптимизировать в соответствии с неким великим планом. Даже если допустить, что сверхразум будет идеально доброжелательным и начисто лишенным самонадеянности, высокомерия, властности, ограниченности и прочих человеческих недостатков, нам может не понравится потеря собственной самостоятельности. Вероятно, мы предпочтем сами определять свою судьбу, даже ценой ошибок. Возможно, нам хотелось бы, чтобы сверхразум выступал в качестве страховки, не допуская катастрофического развития событий, а в остальном предоставил нас самим себе.

Метод КЭВ дает такую возможность. Предполагается, что его нужно запустить лишь в начале, а затем он сам будет развиваться и меняться вместе с изменением экстраполированного волеизъявления. Если экстраполированное волеизъявление человечества склонится к тому, чтобы жить под присмотром патерналистски настроенного сверхразума, тогда процесс КЭВ создаст такой ИИ и передаст ему бразды правления. Если, напротив, экстраполированное волеизъявление человечества придет к тому, чтобы создать демократическое мировое правительство, состоящее из людей, тогда КЭВ инициирует формирование такого института. Если экстраполированное волеизъявление человечества будет заключаться в наделении каждого человека ресурсами, которыми он может пользоваться в свое удовольствие, пока уважает такие же права других людей, тогда процесс КЭВ осуществит и эту идею; причем сверхразум будет присутствовать на заднем плане в качестве этакого воплощения закона природы — чтобы пресекать злоупотребления, кражи, убийства и прочие действия, не согласованные с их объектом[16].

Таким образом, структура модели КЭВ допускает практически неограниченный диапазон исходов. Можно даже допустить, что экстраполированное волеизъявление человечества будет состоять в том, чтобы сверхразум вообще ничего не делал. В этом случае он может безопасно отключить себя, предварительно убедившись в действительно высокой вероятности того, что этого хочет КЭВ.

Дополнительные замечания

Описание модели КЭВ в том виде, как оно дано выше, конечно, очень схематично. Остается множество параметров, которые можно бы описать по-разному и получить в результате разные варианты реализации.

Одним из таких параметров является база экстраполяции: чье волеизъявление следует учитывать? Можно сказать «всех», но этот ответ порождает массу новых вопросов. Должна ли база экстраполяции включать «маргинальные» состояния человека и какие? Примут ли во внимание эмбриональное и плодное состояние? Как быть с людьми, у которых наступила смерть мозга, людьми с тяжелыми психическим расстройствами и пациентами, находящимися в коме? Должны ли оба полушария людей с расщепленным мозгом иметь одинаковый вес, и должен ли он совпадать с весом, который присвоен мозгу среднестатистического здорового человека? Как быть с людьми, которые жили раньше и уже умерли? А с теми, кто родится в будущем? С высшими млекопитающими и другими существами, обладающими чувствами? Имитационными моделями мозга? Инопланетянами?

Одно из мнений заключается в том, чтобы включать в популяцию лишь взрослых людей, живущих на Земле на момент начала создания ИИ. После этого первая экстраполяция, полученная на этой базе, решит, нужно ли ее расширять. Поскольку количество «маргиналов» на периферии этой базы сравнительно невелико, результат экстраполяции не должен сильно зависеть от того, где именно будет проведена граница — например, будут или нет учитываться и эмбрион, и плод или только плод.

Однако если кого-то исключат из исходной базы экстраполяции, это не будет означать, что их желания и благополучие принесли в жертву. Если КЭВ тех, кто в базу включен (например, живущие на Земле взрослые люди), будет состоять в том, что этические соображения следует распространить и на другие существа, результат КЭВ учтет эти предпочтения. Тем не менее возможно, что интересы людей, включенных в исходную базу экстраполяции, будут учтены в большей степени, чем всех остальных. В частности, если процесс начинает действовать только в случае согласия между индивидуальными экстраполированными волеизъявлениями (как предполагается в оригинальном предложении Юдковского), то возникает значительный риск преобладания голосов тех, кто не самым благородным образом начнет выступать против заботы о благополучии высших млекопитающих или

эмуляторов. Тогда был бы нанесен серьезный ущерб нравственной атмосфере всего человеческого сообщества[17].

Метод КЭВ выгодно отличается от других еще и тем, что снижает конкуренцию между людьми за создание первого сверхразума. Но даже признав, что метод выглядит лучше многих других, мы не можем исключить совсем возникновение конфликтных ситуаций. Всегда найдутся те, кто начнет преследовать собственные корыстные интересы: и отдельные личности, и сообщества, и целые государства могут стремиться увеличить свою долю будущего пирога за счет исключения других из базы экстраполяции.

Такое стремление к власти объясняется по-разному. Например, собственник проекта, полностью финансирующий разработку ИИ, имеет право влиять на дальнейшие результаты. Этичность этого права сомнительна. Оспорить ее можно, например, тем, что проект по созданию первого успешного зародыша ИИ несет огромные риски гибели всего человечества, которые, соответственно, должны быть как-то компенсированы. Но справедливая величина компенсации настолько высока, что единственно возможная ее форма — это предоставление каждому доли тех благ, которые удастся получить в случае успеха[18].

Другой аргумент сторонников концентрации власти над ИИ заключается в том, будто значительные слои народонаселения имеют примитивные или аморальные предпочтения, поэтому включение их в базу экстраполяции только увеличит риск превращения будущего человечества в антиутопию. Трудно определить соотношение добра и зла в душе человека. Так же трудно оценить, как оно изменяется в зависимости от его принадлежности к различным группам, социальным стратам, культурам и странам. Независимо от того, оптимистично или пессимистично смотришь на природу человека, вряд ли возникнет желание стремиться к освоению космических пространств на основании одной лишь надежды, что в экстраполированном волеизъявлении большинства из семи миллиардов живущих сейчас на планете людей преобладают добрые черты. Конечно, исключение определенной группы из базы экстраполяции не гарантирует победы; вполне возможно, что в душах тех, кто первым готов исключить других или концентрировать власть над проектом, зла намного больше.

Еще одна причина борьбы за контроль над исходными условиями состоит в том, что кто-то может считать, будто предложенные другими подходы к созданию ИИ не сработают так, как должно, даже если предполагается,

что ИИ будет следовать КЭВ человечества. Если у нескольких групп разработчиков сформируются разные представления о научных направлениях, которые с наибольшей вероятностью приблизят успех, они вступят между собой в беспощадную борьбу. Каждая группа начнет стремиться к тому, чтобы воспрепятствовать остальным создать ИИ. В подобной ситуации конкурирующие проекты должны урегулировать свои эпистемологические противоречия, не прибегая к вооруженным конфликтам, а с помощью методов, позволяющих более надежно определять, кто прав больше, а кто — меньше[19].

Модели, основанные на этических принципах

Модель КЭВ — не единственно возможная форма косвенной нормативности. Например, вместо моделирования когерентного экстраполированного волеизъявления человечества можно пойти иным путем. Разработать ИИ, цель которого будет состоять в том, чтобы выполнять только то, что правильно с этической точки зрения. Причем люди, прекрасно осознавая, насколько когнитивные способности ИИ превосходят их собственные, будут полагаться на его оценочные суждения, какие действия отвечают этому определению, а какие нет. Это можно назвать моделью моральной правоты (далее в тексте — МП). Мысль предельно простая. Человек не обладает абсолютными представлениями, что хорошо, а что плохо; еще хуже он разбирается в том, как лучшим образом проанализировать концепцию моральной правоты с философской точки зрения. Сверхразум в состоянии справиться с подобными проблемами лучше человека[20].

Но как быть, если мы не уверены в истинности концепции моральной правоты? Даже тогда можно попытаться использовать этот метод. Нужно лишь обязательно определить, как должен поступать ИИ в том случае, если выяснится, что его предположения о сути моральной правоты ошибочны. Например, можно указать следующее: если ИИ с довольно высокой вероятностью выяснит, что с концепцией моральной правоты не связано ни одной абсолютной истины, тогда он может вернуться к использованию метода КЭВ или просто отключиться[21].

По всей видимости, у метода МП есть несколько преимуществ по сравнению с методом КЭВ. В случае МП не важны различные свободные параметры метода КЭВ, в частности степень когерентности экстраполированных волеизъявлений, которая требуется для работы КЭВ, легкость, с которой

большинство может взять верх над меньшинством, и природа социального окружения, в котором наши экстраполированные личности должны «стать ближе друг к другу». В этом случае, кажется, невозможна нравственная катастрофа в результате использования слишком узкой или слишком широкой базы экстраполяции. Более того, МП будет подталкивать ИИ к этически правильным действиям даже в том случае, когда наши когерентные экстраполированные волеизъявления могли бы вызвать недопустимые с точки зрения морали шаги ИИ. Однако в границах метода КЭВ — и мы говорили об этом — они вполне возможны. В человеческой природе высокие моральные качества, видимо, скорее являются редким металлом, их не встретишь в изобилии, и даже после того, как золото добыто и очищено в соответствии с методом КЭВ, никто не знает, что получится в результате — драгоценный металл, нейтральный шлак или токсичный осадок.

По всей видимости, метод МП тоже не лишен недостатков. Они основаны на чрезвычайно сложной концепции, вокруг которой ломают копья философы на протяжении многих веков, но в отношении которой так и не достигли согласия. Выбор ошибочного представления о моральной правоте может привести к очень негативным с этической точки зрения результатам. Казалось бы, эта сложность в определении моральной правоты может быть сильнейшим аргументом против использования метода. Однако пока неясно, насколько серьезным недостатком является эта особенность МП. В методе КЭВ тоже используются термины и концепции, которые не слишком легко определить (например, «знания», «быть в большей степени людьми, которыми нам хотелось бы быть», «стать ближе друг к другу»)[22]. Даже если терминологические единицы КЭВ чуть менее непрозрачны, они все-таки очень далеки от тех возможностей, которые сегодня есть у программистов, чтобы выразить все это в исходном коде[23]. Чтобы ИИ справился с любой подобной концепцией, ему нужно обладать хотя бы общими лингвистическими знаниями и способностями (сравнимыми с теми, которые есть у взрослого человека). Но такую общую способность понимать обычный язык можно использовать и для познания того, что означает «моральная правота». Если ИИ способен понять смысл, он выберет и действия, которые соответствуют этому значению. По мере продвижения в сторону сверхразума ИИ может прогрессировать в двух направлениях: в осознании значения моральной правоты с точки зрения философской проблемы и в решении практической задачи по применению этого понимания к оценке конкретных действий[24].

Хотя это и нелегко, но вряд ли *труднее*, чем моделирование когерентного экстраполированного волеизъявления человечества[25].

Более фундаментальная проблема МП заключается в другом. Даже будучи примененным, этот метод может не привести к нужному результату, то есть ИИ не остановится на том правильном варианте, который выбрали бы мы сами, если были бы такими же разумными и информированными, как он. Причем это не просто случайная ошибка, а существенный недостаток модели МП, к тому же чрезвычайно опасный для нас самих[26].

Можно попытаться сохранить основную идею МП и при этом снизить ее требовательность, сосредоточившись на *моральной допустимости* — идее о том, что можно сделать так, что ИИ будет действовать в соответствии с КЭВ человечества до тех пор, пока эти действия будут морально допустимыми. Например, возможна следующая цель ИИ:

> Среди морально допустимых для ИИ действий выбирать те, которые соответствовали бы КЭВ человечества. Однако если какая-то часть этой инструкции определена не строго, или если есть серьезные сомнения в ее значении, или если концепция моральной правоты ложна, или если, создавая ИИ с этой целью, мы действовали морально недопустимо, тогда перейти к контролируемому отключению[27]. Следовать значению этой инструкции, которое имелось в виду разработчиками.

Кого-то может по-прежнему беспокоить факт, что в модели моральной допустимости (далее по тексту — МД) отводится неоправданно большая роль требованиям этики. Но степень жертвы, которую придется принести, зависит от того, какая этическая теория является истинной[28]. Если предполагается этика, отвечающая принципу разумной достаточности, то есть *удовлетворяющая* в том смысле, что считает морально допустимыми любые действия, которые удовлетворяют нескольким моральным ограничениям, — тогда МП может предоставить нам большую свободу во влиянии нашего когерентного экстраполированного волеизъявления на действия ИИ. Если предполагается этика, отвечающая критерию доведения до максимума, то есть *максимизирующая* в том смысле, что морально допустимы лишь те действия, которые имеют наилучшие последствия с этической точки зрения, — тогда метод МД почти или совсем не оставляет возможности, чтобы наши предпочтения влияли на результаты работы ИИ.

Чтобы лучше разобраться в этом соображении, вернемся на минуту к нашему примеру гедонистического консеквенционализма. Предположим,

что эта этическая теория верна и ИИ знает об этом. Для наших целей мы можем определить гедонистический консеквенционализм как утверждение, что действие является этически правильным (и морально допустимым) тогда и только тогда, когда среди всех возможных действий никакое другое не способно обеспечить больший баланс удовольствия и страданий. Тогда ИИ, работающий по методу МД, мог бы максимизировать удовольствие, превратив всю доступную ему часть Вселенной в особую субстанцию гедониум, другую особую субстанцию, компьютрониум, использовать для максимального расширения вычислений, которые обеспечат ощущения, приносящие абсолютное наслаждение. Даже если полная эмуляция головного мозга любого из живущих сейчас людей была бы осуществима, она все равно не стала бы самым эффективным способом, чтобы обеспечить весь мир получением удовольствия, — тогда получается, что все мы погибнем.

Таким образом, применяя метод МП или МД, мы столкнемся с риском отдать свои жизни ради чего-то лучшего. Это может оказаться слишком большой жертвой, чем кажется, поскольку мы потеряем не просто шанс жить обычной человеческой жизнью, но и возможность наслаждаться жизнью гораздо более длинной и наполненной, которую мог бы обеспечить дружественный сверхразум.

Эта жертва представляется еще менее привлекательной, когда понимаешь, что сверхразум мог бы получить почти столь же хороший результат, пожертвовав при этом гораздо меньшей долей нашего потенциального благополучия. Предположим, мы согласились бы допустить, что *почти* вся достижимая Вселенная превращается в гедониум, за исключением какой-то малой ее части, скажем, Млечного Пути, который мы оставим для своих нужд. Даже в таком случае можно будет использовать сотни миллиардов галактик для максимизации удовольствия. И при этом в нашей галактике на протяжении миллиардов лет существовали бы процветающие цивилизации, обитатели которых — и люди, и все другие создания — не просто выжили бы, но благоденствовали в своем постчеловеческом мире[29].

Если предпочитаешь этот последний вариант (как склонен предпочитать его я), вряд ли будешь настаивать на применении принципа моральной допустимости. Что, естественно, не отменяет значимости морали.

Даже с чисто этической точки зрения, возможно, лучше *защищать* метод, который не настолько требователен в отношении морали, как МП или МД.

Если лучшие модели, но основанные на нравственном принципе, не имеют шансов быть использованными — возможно, из-за их чрезмерной требовательности, — может быть, было бы правильнее защищать другое предложение, пусть лишь приближенное к идеалу, но имеющее большие шансы быть примененным[30].

Делай то, что я имею в виду

Мы можем испытывать неуверенность в том, какой метод выбрать: КЭВ, МП, МД или какой-либо еще. Можем ли мы снять с себя ответственность за решение даже такого высокого уровня и переложить ее на ИИ, у которого достаточно развита когнитивная деятельность? Где предел допустимости нашей лени?

Рассмотрим, например, цель, основанную на «разумном подходе»:

сделать так, чтобы разумнее всего стало обращаться к ИИ для выполнения той или иной работы.

Эту цель можно было бы свести к экстраполированному волеизъявлению, морали или чему-то еще, главное, что она могла бы избавить нас от усилий и риска, связанных с попыткой самим выяснять, какую конкретную цель нам было бы разумнее всего выбрать.

Однако здесь также присутствуют некоторые проблемы, характерные для целей, основанных на морали. Во-первых, нас может пугать, что эта цель, основанная на разумном подходе, оставляет слишком мало пространства для наших собственных желаний. Некоторые философы убеждены, что человеку всегда разумнее делать то, что для него лучше всего с этической точки зрения. Может быть, они и правы, но что тогда нас ожидает? Во-первых, цель, основанная на разумности, сжимается до МП — с соответствующим риском, что сверхразум, использующий этот метод, убьет всех, до кого дотянется. Во-вторых, как и в случае всех прочих методов, описанных техническим языком, есть вероятность, что мы ошибочно понимаем значение своих утверждений. Мы видели, что в случае целей, основанных на морали, просьба ИИ делать то, что правильно, способна привести к слишком непредвиденным и нежелательным последствиям. Знай мы об этом заранее, то никогда не наделили бы ИИ подобной целью. Аналогично и с просьбой, обращенной к ИИ, делать то, что мы считали бы самым разумным действием.

Попробуем избежать этих трудностей, описав цель подчеркнуто нетехническим языком, скажем, использовав слово *милый*[31]:

> *вести себя очень мило; если не получается очень мило, тогда вести себя как минимум просто мило.*

Как можно возражать против создания *милого* ИИ? Но мы должны спросить, что означает это слово. В словарях можно найти разные значения слова *милый*, которые явно не предполагались для нашего случая. Нам совсем не нужно, чтобы ИИ был что-то типа «любезный», «вежливый», «изысканный» или «утонченный». Если можно было бы положиться на то, что ИИ распознает предполагавшуюся нами интерпретацию слова *милый* и будет мотивирован на милые действия именно в этом смысле, тогда цель, похоже, свелась бы к команде ИИ делать то, что программисты имели в виду[32]. Аналогичное указание было включено в формулировку КЭВ («…понимать так, как нам хотелось бы понимать») и в критерий моральной допустимости, описанный ранее («…следовать предполагаемому значению этой инструкции»). Употребив фразу «делай, что я имею в виду», мы фактически сообщаем машине, что все остальные слова в описании не следует понимать буквально. Но говоря, что ИИ должен быть «милым», мы не добавляем ничего — вся реальная нагрузка ложится на команду «делай, что я имею в виду». Если бы мы знали, как адекватно отразить в коде команду «делай, что я имею в виду», ее можно было бы также использовать в качестве отдельной цели.

Как можно было бы использовать этот процесс «делай, что я имею в виду»? То есть как создать ИИ, мотивированный доброжелательно интерпретировать наши желания и невысказанные намерения и действовать в соответствии с ними? Начать можно с попытки прояснить, что мы подразумеваем под фразой «делай, что я имею в виду». Как выразить тот же смысл, но используя другую терминологию — скажем, бихевиористской теории. Почему бы нам не применить термины предпочтения, которые проявляются в тех или иных гипотетических ситуациях, например, когда у нас больше времени для размышления над вариантами решения, в которых мы умнее, в которых мы знаем больше фактов, имеющих отношение к делу, — в общем, в таких, когда складываются благоприятные условия, чтобы мы могли четко показать на конкретных примерах, что мы имеем в виду, когда хотим видеть ИИ дружелюбным, полезным, милым…

Здесь мы замкнули круг. И вернулись к косвенной нормативности, с которой начали. В частности, к методу КЭВ, предполагающему, что из описания цели исключается все конкретное, после чего в нем остается лишь абстрактная цель, определенная в чисто процедурных терминах: делать то, что мы хотели бы, чтобы делал ИИ в соответствующих идеальных обстоятельствах. Идя на такую уловку, то есть используя косвенное называние, мы надеемся переложить на ИИ большую часть интеллектуальной работы, которую пришлось бы выполнять нам самим, попытайся мы сформулировать более конкретное описание целей ИИ. Следовательно, если мы стремимся в полной мере использовать эпистемологическое превосходство ИИ, КЭВ становится выражением принципа эпистемологического уважения.

Перечень компонентов

До сих пор мы рассматривали различные варианты того, как должна быть описана цель ИИ. Но на поведение разумной системы оказывают влияние и другие компоненты архитектуры. Особенно критически важно, какие методы теории принятия решений и теории познания в нем используются. И еще: будет ли ИИ сообщать людям о своих планах до начала их реализации.

В табл. 13 приведен список проектных решений, которые необходимо выполнить в процессе создания сверхразума. Разработчики проекта должны уметь объяснить, какие решения были приняты по отношению к каждому случаю и почему были приняты именно эти, а не иные решения[33].

Таблица 13. Перечень компонентов

Описание цели	Какую цель должен преследовать ИИ? Как он должен интерпретировать описание этой цели? Должна ли цель предполагать специальное вознаграждение для тех, кто внес свой вклад в успех проекта?
Принятие решений	Должен ли ИИ использовать причинный подход к принятию решений, подход на основе ожиданий, безусловный подход или какой-то еще?
Эпистемология, то есть познание мира	Какой должна быть функция априорной вероятности ИИ, какие другие явные и неявные предположения о мире ему следует сделать? Как он должен учитывать влияние человека?
Ратификация, то есть подтверждение	Должны ли планы ИИ проходить проверку человеком прежде, чем будут реализованы? Если да, как будет организован этот процесс?

Описание цели

Мы уже обсуждали, как можно использовать косвенную нормативность для определения цели, которую должен преследовать ИИ, и рассмотрели некоторые варианты, например модели, основанные на морали и когерентном экстраполированном волеизъявлении. Если мы остановились на какомто из вариантов, это создает необходимость делать следующий выбор. Например, есть множество вариаций модели КЭВ, зависящих от того, кого включают в базу экстраполяции, от ее структуры и так далее. Другие формы мотивации могут означать необходимость использовать иное описание цели. Предположим, создается ИИ-оракул, цель которого — давать точные ответы. Если при этом используется метод приручения, в описании цели должна присутствовать формулировка о недопустимости чрезмерного использования ресурсов при подготовке этих ответов.

При выборе вариантов устройства ИИ следует также ответить на вопрос, должна ли цель предполагать специальное вознаграждение для тех, кто внес свой вклад в успех проекта, например за счет выделения им дополнительных ресурсов или оказания влияния на поведение ИИ. Любое упоминание этого обстоятельства можно назвать «стимулирующий пакет». Благодаря стимулирующему пакету мы повысим вероятность успешной реализации проекта, пусть и ценой некоторого компромисса с точки зрения достижения стоящих перед ним целей.

Например, если целью проекта является создание процесса, реализующего когерентное экстраполированное волеизъявление человечества, тогда стимулирующий пакет может представлять собой указание, что желаниям некоторых людей при экстраполировании будет присвоено большее значение. Если такой проект окажется успешным, его результатом не обязательно станет реализация когерентного экстраполированного волеизъявления всего человечества. Скорее, будет достигнута некоторая близкая к этому цель[34].

Поскольку стимулирующий пакет включается в описание цели, а сверхразуму придется обязательно интерпретировать и реализовывать это определение, можно воспользоваться преимуществами, предлагаемыми методом косвенной нормативности, и сформулировать в описании сложные положения, довольно трудные, если реализацией будет заниматься человек. Например, вместо того чтобы вознаграждать программистов за какую-то наспех сколоченную, но понятную системой показателей, например сколько часов

они проработали и сколько ошибок исправили, в стимулирующем пакете может быть указано, что программисты «должны получать вознаграждение пропорционально их вкладу, который увеличил ожидаемую вероятность успешной реализации проекта так, как этого хотели заказчики». Более того, нет причины ограничивать круг тех, на кого распространяется стимулирующий пакет, только участниками проекта. Можно указать, что вознагражден будет *каждый* человек в соответствии с его вкладом. Распределение благодарности — задача трудная, но от ИИ можно ожидать разумной аппроксимации критерия, явно или неявно указанного в стимулирующем пакете.

Предположим, сверхразум отыщет какой-то способ вознаградить даже тех, кто умер задолго до его создания[35]. И тогда стимулирующий пакет может быть расширен за счет некоторых умерших людей, причем не только до начала проекта, но и, возможно, до первого упоминания о концепции стимулирующего пакета. Конечно, проведение такой ретроспективной политики уже ни к чему людям, лежащим в могилах, но это было бы правильно по этическим соображениям. Правда, есть одно техническое возражение: если цель проекта — воздать по заслугам, то тогда об этом нужно упомянуть в самом описании цели, а не в рамках стимулирующего пакета.

Мы не можем слишком глубоко погружаться в этические и стратегические вопросы, связанные с созданием стимулирующего пакета. Однако позиция разработчиков ИИ по этим вопросам является важным аспектом концепции его устройства.

Принятие решений

Еще один важный компонент проекта по созданию ИИ — система принятия решений. Выбор того или иного подхода влияет на стратегию поведения ИИ в судьбоносные моменты его существования. От этого может зависеть, например, будет ли ИИ открыт для договорных отношений с другими гипотетическими сверхразумными цивилизациями или, напротив, станет объектом шантажа с их стороны. Специфика процесса принятия решений также играет роль в трудных ситуациях вроде тех, когда имеется ограниченная вероятность неограниченных выигрышей («пари Паскаля») или чрезвычайно низкая вероятность чрезвычайно высоких ограниченных выигрышей («ограбление Паскаля»), а также когда ИИ сталкивается

с фундаментальной нормативной неопределенностью или многочисленными копиями той же программы-агента[36].

Варианты, приведенные в таблице, включают в себя причинный подход к принятию решений (у которого есть множество вариаций) и подход на основе ожиданий, а также более современных кандидатов вроде «безусловного» и «вневременного» подхода, которые еще только разрабатываются[37]. Может оказаться не слишком просто подобрать верный вариант и убедиться, что мы правильно его используем. Хотя перспективы прямого описания подхода ИИ к принятию решений кажутся более реальными, чем прямого описания его конечных целей, значительный риск ошибки все-таки существует. Многие сложности, способные поставить в тупик наиболее популярные теории принятия решений, были обнаружены совсем недавно, откуда следует, что могут существовать и иные проблемы, пока невидимые глазу. В случае ИИ результаты применения ошибочного подхода могут быть катастрофическими, вплоть до гибели всего человечества.

Учитывая эти сложности, возникает идея описать подход к принятию решений, который должен использовать ИИ, непрямым методом. Можно предложить ИИ использовать «тот подход к принятию решений *D*, который мы предложили бы ему применить после долгих размышлений над этим вопросом». Однако ИИ должен иметь возможность принимать решения еще до того, как узнает, что такое *D*. Отсюда возникает потребность в промежуточном подходе к принятию решений *D*', которым ИИ мог бы руководствоваться в процессе поиска *D*. Можно попытаться определить *D*' как своего рода суперпозицию текущих гипотез ИИ о *D* (взвешенных на их вероятности), хотя остаются нерешенными некоторые технические проблемы, в частности, как сделать это в общем виде[38]. Есть еще один повод для беспокойства: ИИ способен сделать непоправимо неверный выбор (например, перезаписать себя, зафиксировав ошибочный подход к принятию решений) на стадии обучения, прежде чем у него появится возможность определить правильность того или иного подхода. Чтобы уменьшить риск ошибки в период повышенной уязвимости ИИ, можно попробовать наделить зародыш *ограниченной рациональностью* в той или иной форме — сознательно упрощенным, но более надежным подходом к принятию решений, который стойко игнорирует эзотерические соображения, даже если мы думаем, что они в конечном счете могут оказаться правильными, и который впоследствии должен заменить себя на более

сложный (непрямой) подход к принятию решений, удовлетворяющий определенным критериям[39]. Можно ли это сделать, и если да, то как — вопрос открытый.

Эпистемология, или Познание мира

В рамках проекта необходимо сделать фундаментальный выбор в отношении методов теории познания, которыми будет пользоваться ИИ, и описать принципы и критерии оценки эпистемологических гипотез. Можно, например, остановиться на байесовском подходе и принять эпистемологию как функцию априорного распределения вероятности — имплицитного присвоения ИИ значений вероятности возможным мирам до того, как им будут рассмотрены и учтены какие-либо воспринимаемые свидетельства. В других условиях методы познания могут принимать иную форму, однако в любом случае необходимо некоторое индуктивное правило обучения, если ИИ должен обобщать наблюдения, сделанные в прошлом, и делать предсказания относительно будущего[40]. Однако, как и в случае с описанием цели и подходом к принятию решений, есть риск, что наше определение эпистемологии окажется ошибочным.

На первый взгляд может показаться, что размер ущерба от неправильного выбора методов теории познания ограничен. Ведь если они *совсем* не будут работать, ИИ просто окажется не слишком интеллектуальным и не будет представлять угрозу, о которой говорится в этой книге. Но остается опасность, что эпистемология будет определена достаточно хорошо для того, чтобы ИИ был инструментально эффективен в большинстве ситуаций, но при этом в определении будет содержаться некий изъян, из-за которого ИИ собьется с пути в каком-то жизненно важном вопросе. Такой ИИ станет походить на умного человека, абсолютно убежденного в истинности ложной догмы, на которой выстроена его философия; он начнет «бороться с ветряными мельницами» и всего себя посвятит достижению фантастических или опасных целей.

Незначительные различия в априорном распределении вероятностей способны привести к серьезным отличиям в поведении ИИ. Например, может быть приравнена к нулю априорная вероятность того, что Вселенная бесконечна. И тогда независимо от количества астрономических свидетельств в пользу этого ИИ будет упрямо отвергать все космологические теории, построенные на идее бесконечной Вселенной, делая в результате

неправильный выбор[41]. Или окажется нулевой априорная вероятность того, что Вселенная не является вычислимой по Тьюрингу (на самом деле это общее свойство многих априорных распределений вероятностей, которые обсуждаются в научной литературе, включая уже упомянутую в первой главе колмогоровскую сложность), что также приведет к плохо понимаемым последствиям, если это допущение — известное как тезис Чёрча–Тьюринга — окажется ложным. ИИ может быть наделен априорным распределением вероятностей, которое приведет к появлению у него тех или иных сильных метафизических воззрений разного рода, например предположение о возможности истинности дуализма разума и тела или возможности существования не поддающихся улучшению моральных фактов. Если какие-то из этих воззрений окажутся ошибочными, возникнет шанс того, что ИИ будет стремиться достичь своих конечных целей способами, которые мы бы отнесли к порочной реализации. И при этом нет никакой очевидной причины полагать, что такой ИИ, будучи фундаментально неправым в каком-то одном очень важном аспекте, не сможет стать достаточно эффективным в инструментальном смысле, чтобы обеспечить себе решающее стратегическое преимущество. (Еще одной областью, где может играть ключевую роль выбор эпистемологических аксиом, является изучение эффекта наблюдателя и его влияния на выводы, которые можно сделать на основе дейктической информации[42].)

У нас есть все основания сомневаться в своей способности разрешить все фундаментальные эпистемологические проблемы к моменту начала создания первого зародыша ИИ. Поэтому лучше исходить из того, что для задания его методов познания мира будет использован непрямой подход. Но тогда возникает множество вопросов, аналогичных случаю применения непрямого подхода к определению процесса принятия решений. Однако в случае эпистемологии есть больше надежд на позитивную конвергентность, поскольку любой из широкого спектра подходов теории познания обеспечивает создание безопасного и эффективного ИИ и в конечном счете приводит к одним и тем же результатам. Причина заключается в том, что различия в априорном распределении вероятностей, как правило, стираются при наличии довольно большого количества эмпирических свидетельств и в результате проведения глубокого анализа[43].

Было бы неплохо поставить себе цель наделить ИИ фундаментальными эпистемологическими принципами, аналогичными тем, которые управляют

нашим собственным мышлением. Тогда, если последовательно применять свои стандарты, любой ИИ, отклоняющийся от этого идеала, должен считаться мыслящим неправильно. Конечно, это применимо лишь к нашим действительно *фундаментальным* эпистемологическим принципам. Не относящиеся к фундаментальным принципы ИИ должен постоянно создавать и пересматривать самостоятельно по мере развития своих представлений о мире. Задача ИИ — не потворствовать человеческим предубеждениям, а избавляться от следствий нашего невежества и глупости.

Ратификация, или Подтверждение

Последним пунктом в нашем списке вариантов выбора различных аспектов устройства ИИ является *ратификация*. Должны ли планы ИИ проходить проверку человеком прежде, чем будут реализованы? В случае ИИ-оракула ответ на этот вопрос утвердительный по определению. Оракул выдает информацию; человек решает, использовать ли ее и если да, то как. Однако в случае ИИ-джинна, ИИ-монарха и ИИ-инструмента вопрос о том, нужна ли какая-то форма ратификации, остается открытым.

Чтобы посмотреть, как может работать ратификация, возьмем ИИ, который должен действовать как монарх, реализующий КЭВ человечества. Представим, что прежде чем запустить его, мы создаем оракула, единственной целью которого будет отвечать на вопросы о том, что должен делать монарх. В предыдущих главах мы видели, что с созданием оракула-сверхразума связаны определенные риски (в частности, риск проявления преступной безнравственности или риск инфраструктурной избыточности). Но мы примем за данность, что ИИ-оракул будет успешно создан и указанные подводные камни удастся обойти.

Итак, есть ИИ-оракул, выдающий нам свои оценки последствий запуска тех или иных фрагментов кода, в которых реализуется КЭВ человечества. Оракул не может прогнозировать во всех деталях, что произойдет, но его предсказания, скорее всего, окажутся точнее наших. (Если сверхразум *ничего* не сможет сказать о том, что будет делать программа, было бы безумием ее запускать.) В общем, оракул немного думает и выдает результат. Чтобы он был понятнее, оракул может предложить оператору набор инструментов, с помощью которых можно изучить различные аспекты предсказанного исхода. Помимо картины, как может выглядеть будущее, оракул представит статистику количества мыслящих существ, которые будут жить в разные

времена, и нижние, средние и пиковые показатели их благополучия. Он также может составить подробные биографии нескольких случайных людей (возможно, воображаемых, выбранных в силу репрезентативности). И обратить внимание оператора на некоторые аспекты, о которых тот мог бы не спросить, но которые действительно заслуживают его внимания.

Такая способность заранее проанализировать возможные исходы дает нам очевидные преимущества. В ходе анализа можно увидеть последствия ошибки в определениях, которые планируется заложить в ИИ-монарха или записать в его исходном коде. Если «хрустальный шар» показывает нам будущее в руинах, можно удалить код планируемого к созданию монарха и попробовать что-то еще. Будем считать, что изучать возможные последствия нашего выбора прежде, чем сделать его, следует непременно, особенно в тех случаях, когда на кону — будущее всего человеческого вида.

Потенциально серьезные недостатки ратификации не лежат на поверхности. Желание противоборствующих фракций заранее увидеть, каким будет вердикт высшего разума, вместо того чтобы просто положиться на его мудрость, может подорвать миротворческую суть КЭВ. Сторонники подхода, основанного на морали, могут беспокоиться из-за того, что решимость спонсора улетучится, как только он увидит, к каким жертвам приведет стремление к оптимальному решению с точки зрения этики. Кроме того, у нас могут быть все основания предпочитать жизнь, в которой потребуется постоянно преодолевать себя, то есть будущее, полное сюрпризов и противоречий, — будущее, контуры которого не так тесно привязаны к нынешним исходным условиям, но оставляющее определенный простор для резкого движения и незапланированного роста. Мы с меньшей вероятностью строили бы амбициозные планы, если бы могли подбирать каждую деталь будущего и отправлять на доработку его черновики, не полностью отвечающие нашему преходящему настроению.

Итак, вопрос ратификации планов ИИ организаторов не слишком прост, как может показаться вначале. Тем не менее правильнее было бы воспользоваться возможностью и ознакомиться с вариантами, если такой функционал будет реализован. Но не стоит ждать от наблюдателя детального изучения и корректировки каждого аспекта предполагаемого результата, будет лучше, если мы наделим его правом вето, которое он мог бы использовать ограниченное число раз, прежде чем проект был бы окончательно прекращен[44].

Выбор правильного пути

Главной целью ратификации является уменьшение вероятности катастрофической ошибки. В целом кажется, что правильнее ставить перед собой именно эту цель, нежели максимизировать шансы оптимизации каждой детали плана. На то есть две причины. Во-первых, распространение человечества имеет космические масштабы — есть куда развиваться, даже если с нашим процессом будут связаны некоторые потери или ненужные ограничения. Во-вторых, есть надежда, что если исходные условия для взрывного развития интеллекта мы выберем более или менее верно, то сверхразум в конечном счете реализует наши ожидания. Здесь важно попасть в правильный аттрактор.

Что касается эпистемологии, то есть познания мира, можно предположить, что широкий спектр априорных распределений вероятностей в конечном счете сойдется к очень близким апостериорным распределениям (если вычислениями будет заниматься сверхразум, определяя условную вероятность на реалистичных данных). Поэтому нам не нужно беспокоиться о том, чтобы эпистемология была *идеально* правильной. Нужно лишь избежать ситуации, в которой ИИ получит такое экстремальное априорное распределение вероятностей, что не сможет обучиться важным истинам, даже несмотря на интенсивные исследования и анализ[45].

Что касается принятия решений, то здесь риск непоправимой ошибки кажется более высоким. Но надежда прямо описать достаточно хороший подход к принятию решений все-таки есть. ИИ, обладающий сверхразумом, способен в любой момент переключиться на новый подход, но если начнет с совсем неудачного, то может не увидеть причину для переключения. Или ему не хватит времени выбрать заведомо лучший подход. Возьмем, например, агента, который не должен поддаваться шантажу и умеет отсеивать потенциальных вымогателей. Вполне возможно, что при его создании использовался оптимальный подход к принятию решений. Но если агент получит угрозу и решит, что она заслуживает доверия, ему будет нанесен ущерб.

При наличии адекватных подходов к принятию решений и познанию мира можно попробовать создать систему, использующую КЭВ или какое-то иное косвенное описание цели. В этом случае снова есть надежда на конвергентность — разные способы реализации КЭВ должны привести

к одинаково благоприятным для человечества исходам. Если не предполагать конвергентность, то остается лишь надеяться на лучшее.

У нас нет необходимости тщательно оптимизировать систему. Скорее, следует сосредоточить внимание на надежном проекте, который внушит уверенность, что ИИ достанет здравого смысла распознать свою ошибку. Несовершенный ИИ, построенный на прочном основании, постепенно исправит себя сам, после чего приложит к миру не меньше позитивной оптимизирующей силы, чем мог бы приложить, будучи совершенным с самого начала.

Глава четырнадцатая
Стратегический ландшафт

Пришло время рассмотреть проблему сверхразума в более широком контексте. Нам следует хорошо ориентироваться в стратегическом ландшафте хотя бы для того, чтобы представлять общее направление своего движения. Как оказывается, это непросто. В предпоследней главе мы познакомимся с несколькими общими аналитическими концепциями, которые помогут нам обсуждать долгосрочные научные и технологические проблемы. А затем попробуем применить их к машинному интеллекту.

Рассмотрим различие между двумя нормативными подходами, при помощи которых можно оценивать любую предлагаемую стратегию. *Субъективная точка зрения* предполагает ответ на вопрос: насколько проведение тех или иных изменений «в наших интересах» — то есть насколько (в среднем и предположительно) они будут отвечать интересам тех обладающих моральным статусом субъектов, которые или уже существуют, или будут существовать независимо от того, произойдут предлагаемые изменения или нет. *Объективная точка зрения*, напротив, не предполагает учет мнения существующих людей или тех, кто будет жить в будущем, независимо от того, произойдут ли предполагаемые изменения. Она учитывает всех одинаково независимо от их положения на временной шкале. С объективной точки зрения наибольшую ценность имеет появление новых людей, при условии, что их жизнь будет стоить того, чтобы ее прожить, — чем более счастливой будет их жизнь, тем лучше.

Для первичного анализа может быть полезно сопоставить эти две точки зрения, хоть такой прием является лишь легким намеком на этические сложности, связанные с революцией машинного интеллекта. Вначале следует посмотреть на ситуацию с объективной, а затем с субъективной точек зрения и сравнить их.

Стратегия научно-технологического развития

Прежде чем обратиться к более узким вопросам машинного интеллекта, следует познакомиться с несколькими стратегическими концепциями и соображениями, которые имеют отношение к научному и технологическому развитию в целом.

Различные темпы технологического развития

Предположим, что некое лицо, наделенное властными полномочиями, предлагает прекратить финансирование определенной области исследований, оправдывая это рисками или долгосрочными последствиями, связанными с развитием новой технологии, которая когда-нибудь может появиться в этой области. Естественно, официальное лицо не удивится, если со стороны научного сообщества поднимется ропот недовольства.

Ученые и их защитники часто говорят, что контролировать эволюцию технологий, блокируя исследования, бесполезно. Если технология возможна (говорят они), то она появится независимо от опасений властных лиц по поводу предполагаемых будущих рисков. Действительно, чем бóльшим потенциалом обладает технология, тем с большей уверенностью можно сказать, что кто-то где-то будет мотивирован ее создать. Прекращение финансирования не остановит прогресс и не устранит сопутствующие ему риски.

Интересно, что этот аргумент о бесполезности контроля почти никогда не возникает в тех случаях, когда лица, наделенные властными полномочиями, предлагают *увеличить* финансирование некоторых областей исследований, несмотря на то что аргумент этот, похоже, обоюдоострый. Не приходилось слышать возмущенных криков: «Пожалуйста, не увеличивайте бюджеты в нашей области. А лучше сократите их. Исследователи в других странах наверняка заполнят пустоту, эта работа так или иначе будет сделана. Не растрачивайте общественные средства на проведение исследований в нашей стране!»

С чем связана такая двойственность? Одно из объяснений, конечно, заключается в том, что у членов научного сообщества проявляется предубеждение в свою пользу, заставляющее нас верить, что исследования — это всегда хорошо, и пытаться использовать практически любые аргументы в пользу повышения их финансирования. Однако возможно также, что двойные стандарты объясняются национальными интересами.

Предположим, что развитие некой технологии будет иметь эффект *двух* типов: небольшую пользу *B* для ее создателей и страны, которая их финансирует, и гораздо более ощутимый вред *H* — вплоть до риска гибели — для всех. Выбор в пользу разработки такой опасной технологии могли бы сделать даже те, кто в целом придерживается альтруистических взглядов. Они могли бы рассуждать следующим образом: вред *H* будет нанесен в любом случае, что бы они ни делали, поскольку если не они, то кто-то еще все равно эту технологию создаст; а раз так, то можно было бы хотя бы получить пользу *B* для себя и своей страны. («К несчастью, скоро появится оружие, которое уничтожит мир. К счастью, мы получили грант на его создание!»)

Чем бы ни объяснялся аргумент о бесполезности контроля, с его помощью невозможно доказать, что с объективной точки зрения отсутствуют причины для попытки контролировать технологическое развитие. Это невозможно доказать, даже если мы допустим истинность вдохновляющей идеи, что в случае неустанных усилий ученых и исследователей в конечном счете все возможные технологии будут созданы — то есть если мы допустим следующее.

— *Гипотеза о технологической полноте*

Если ученые и исследователи не прекратят свои усилия, тогда все важные базовые возможности, которые можно получить при помощи неких технологий, будут получены[1].

Есть как минимум две причины, по которым из гипотезы о технологической полноте не следует аргумент о бесполезности контроля. Во-первых, может не иметь место посыл, поскольку на самом деле нам не дано знать, прекратят ли ученые и исследователи свои усилия (прежде чем будет достигнута технологическая зрелость). Эта оговорка особенно актуальна в контексте, в котором присутствует экзистенциальный риск. Во-вторых, даже если мы уверены, что все важные базовые возможности, которые можно получить при помощи неких технологий, будут получены, все же имеет смысл попытаться влиять на направление исследований. Важно не то, *будет ли* создана та или иная технология, а то, когда она будет создана, *кем* и *в каком контексте*. А на эти обстоятельства рождения новой технологии, определяющие эффект от нее, можно влиять, открывая и закрывая вентиль финансирования (и применяя другие инструменты контроля).

Из этих рассуждений следует принцип, который предполагает, что нам нужно уделять внимание относительной скорости развития различных технологий[2].

— *Принцип разной скорости технологического развития*
Следует тормозить развитие опасных и вредных технологий, особенно тех, которые повышают уровень экзистенциального риска, и ускорять развитие полезных технологий, особенно тех, которые снижают уровень экзистенциального риска, вызванного другими технологиями.

Соответственно, любое решение лиц, облеченных властью, можно оценивать исходя из того, насколько оно обеспечивает преимуществами желательные виды технологий по сравнению с нежелательными[3].

Предпочтительный порядок появления

Некоторые технологии являются амбивалентными по отношению к экзистенциальным рискам, увеличивая одни и снижая другие. К такого рода технологиям принадлежит изучение и разработка сверхразума.

В предыдущих главах мы видели, что с появлением машинного сверхразума возникает высочайший экзистенциальный риск глобального масштаба, но при этом другие экзистенциальные риски могут снижаться. Скажем, практически исчезнут риски гибели человечества вследствие природных катаклизмов: падений астероидов, извержений вулканов, вспышек эпидемий — поскольку сверхразум способен разработать контрмеры в каждом из этих случаев или хотя бы перевести их в менее опасную категорию (например, за счет колонизации космоса).

Экзистенциальные риски природных катаклизмов относительно малы на обозримой шкале времени. Но сверхразум способен также устранить совсем или существенно снизить многие антропогенные риски. В частности, риск случайной гибели человечества в результате инцидентов, связанных с новыми технологиями. Поскольку сверхразум в целом гораздо способнее человека, он будет меньше ошибаться, лучше понимать, когда требуются те или иные меры предосторожности, и принимать их более компетентно. Разумно устроенный сверхразум может иногда идти на риск, но только если это оправданно. Более того, как минимум в тех сценариях, где сверхразум формирует синглтон, многие антропогенные экзистенциальные риски

неслучайного характера, связанные с проблемой отсутствия глобальной координации, исчезают совершенно. К ним относятся риски войн, технологической и других нежелательных форм конкуренции и эволюции, а также избыточного использования ресурсов.

Поскольку существенную угрозу вызывает развитие людьми таких направлений, как синтетическая биология, молекулярные нанотехнологии, климатический инжиниринг, инструменты биомедицинского улучшения и нейропсихологического манипулирования, средства социального контроля, способные привести к тоталитаризму или тирании, а также другие технологии, которые пока невозможно представить, устранение этих рисков станет огромным благом для человечества. Однако если природные и прочие риски не техногенного характера низки, то этот аргумент стоит уточнить: важно, чтобы сверхразум был создан *раньше* других опасных технологий, например инновационных нанотехнологий. Важно не то, когда он появится (с объективной точки зрения), а правильный порядок появления.

Если сверхразум появится раньше других потенциально опасных технологий, скажем, нанотехнологий, то он будет в силах снизить экзистенциальный риск, связанный с нанотехнологиями, но нанотехнологии снизить риск, связанный с появлением сверхразума, не могут[4]. Следовательно, создав вначале сверхразум, мы будем иметь дело лишь с одним видом экзистенциального риска, а создав вначале нанотехнологии, получим как риск нанотехнологий, так и риск сверхразума[5]. То есть даже если экзистенциальный риск сверхразума высок и даже если это самая рискованная из всех технологий, все равно имеет смысл торопить его создание.

Однако эти аргументы в пользу «чем быстрее, тем лучше» основаны на предположении, что риски, связанные со сверхразумом, одинаковы независимо от того, когда он будет создан. Если все-таки они снижаются с течением времени, возможно, лучше было бы замедлить революцию машинного интеллекта. Даже притом что более позднее появление сверхразума оставит больше времени для других экзистенциальных катастроф, может быть, предпочтительнее подождать с его созданием? Особенно это верно в том случае, если экзистенциальные риски, связанные со сверхразумом, намного выше, чем риски других потенциально опасных технологий.

Есть несколько причин полагать, что опасность, заложенная во взрывном развитии интеллекта, будет резко снижаться в ближайшие десятилетия.

Одна из них заключается в том, что чем позже произойдет взрыв, тем больше времени будет на решение проблемы контроля. Это осознали лишь недавно, и большинство лучших идей по решению проблемы появились за последнее десятилетие (причем некоторые из них — уже за время написания этой книги). Есть основания полагать, что ситуация будет прогрессировать, и хотя проблема оказалась очень сложной, в течение века развитие может оказаться значительным. Чем позже появится сверхразум, тем большего прогресса удастся добиться. Это сильный аргумент в пользу более позднего появления сверхразума — и очень сильный аргумент против его слишком раннего появления.

Есть еще одна причина, почему отсрочка в появлении сверхразума может сделать его безопаснее. Она даст возможность развиться различным позитивным тенденциям, способным повлиять на его действия. Насколько большое значение вы присвоите этой причине, зависит от оптимистичности вашего взгляда на человеческую цивилизацию.

Оптимист, конечно, укажет на множество обнадеживающих признаков и повсеместно возникающих возможностей. Люди могут научиться лучше ладить друг с другом, что приведет к снижению уровня насилия, количества войн и жестокости; может вырасти степень глобальной координации и политической интеграции, что позволит избежать нежелательной гонки вооружений в области технологий (подробнее об этом ниже) и прийти к согласию относительно широкого распространения потенциальных благ от взрывного развития интеллекта. Похоже, именно в этом направлении складываются долгосрочные исторические тенденции[6].

Кроме того, оптимист может ожидать, что в нашем столетии возрастет здравомыслие человечества: станет меньше предубеждений (в среднем); продолжат накапливаться знания о мире; людям станет привычнее размышлять об абстрактных будущих вероятностях и глобальных рисках. Если повезет, мы станем свидетелями общего роста эпистемологических стандартов как в индивидуальном, так и в коллективном сознании. Тенденции такого масштаба обычно заметны. Научный прогресс приведет к тому, что мы будем знать намного больше. Благодаря экономическому росту все бóльшая доля человечества будет получать достаточное питание (особенно в возрасте, критическом с точки зрения развития мозга) и доступ к качественному образованию. Успехи в информационных технологиях упростят поиск, обобщение, оценку и распространение данных и идей. Более того,

к концу столетия на счету человечества будут дополнительные сто лет ошибок — и на них можно будет учиться.

Многие потенциальные плоды прогресса являются амбивалентными в уже упомянутом нами смысле: они увеличивают одни экзистенциальные риски и снижают другие. Например, благодаря успехам в развитии технологий слежения, глубинной обработки данных, детекции лжи, биометрии, психологических или нейрохимических средств манипулирования мнениями и желаниями можно снизить экзистенциальные риски за счет большей международной координации усилий по борьбе с терроризмом и преступностью. Однако те же самые успехи могут увеличить некоторые экзистенциальные риски в результате усиления негативных социальных процессов или формирования стабильных тоталитарных режимов.

Важным рубежом является улучшение биологического механизма познания, например за счет генетического отбора. Обсуждая это в главах второй и третьей, мы пришли к выводу, что наиболее радикальные системы сверхразума, скорее всего, возникнут в форме машинного интеллекта. Это заявление согласуется с тем, что биологическое улучшение интеллектуальных способностей сыграет важную роль в подготовке к созданию и собственно создании искусственного сверхразума. Такое когнитивное совершенствование может показаться очевидным способом снизить риски — чем умнее люди, работающие над проблемой контроля, тем больше вероятность, что ее решение будет найдено. Однако когнитивное совершенствование также ускоряет процесс создания машинного интеллекта, тем самым сокращая время, которое у нас есть для поиска этого решения. Когнитивное совершенствование может иметь и другие следствия, имеющие отношение к интересующему нас вопросу. Они заслуживают более пристального рассмотрения. (Большинство следующих мыслей относительно когнитивного улучшения равно применимы и к небиологическим средствам повышения нашей индивидуальной или коллективной эпистемологической эффективности.)

Скорость изменений и когнитивное совершенствование

Повышение среднего уровня и верхнего предела человеческих интеллектуальных способностей, скорее всего, приведет к ускорению технологического прогресса, в том числе в области создания различных форм машинного интеллекта, решения проблемы контроля и достижения множества других

технических и экономических целей. Каким будет чистый эффект такого ускорения?

Давайте рассмотрим ограниченный случай «универсального ускорителя» — воображаемого вмешательства, с помощью которого можно ускорить буквально *все*. Действие такого универсального ускорителя скажется больше на произвольном изменении масштаба временной шкалы и не приведет к качественным изменениям наблюдаемых результатов[7].

Если мы хотим наполнить смыслом идею ускорения различных процессов благодаря когнитивному совершенствованию, нам, очевидно, нужна какая-то иная концепция, нежели универсальный ускоритель. Более перспективно было бы сконцентрировать внимание на вопросе, как когнитивное улучшение способно повысить скорость изменения одних процессов *относительно* скорости изменения других. Такое различие способно повлиять на динамику всей системы. Итак, рассмотрим следующую концепцию.

— *Ускоритель макроструктурного развития*

Рычаг, который повышает скорость развития макроструктурных свойств человеческого существования, при этом оставляя неизменной динамику на микроуровне.

Представьте, что вы потянули этот рычаг в направлении торможения. Тормозные колодки прижимаются к огромному колесу человеческой истории, летят искры, раздается металлический скрежет. После замедления колеса мы получаем мир, в котором технологические инновации происходят медленнее, а фундаментальные или имеющие глобальное значение изменения в политической структуре и культуре случаются реже и не столь резко, как прежде. Чтобы одна эпоха пришла на смену другой, должны смениться многие поколения. На протяжении своей жизни люди становятся свидетелями совсем незначительных изменений в базовой структуре человеческого существования.

Большинство представителей нашего вида были свидетелями гораздо более медленного макроструктурного развития, чем сейчас. Пятьдесят тысяч лет назад могло пройти целое тысячелетие без единого значительного технологического прорыва, без заметного увеличения объема человеческих знаний и понимания мира, без значимых глобальных политических изменений. Однако на микроуровне калейдоскоп человеческой жизни мелькал с довольно высокой скоростью — рождения, смерти и прочие события

в жизни человека. В эпоху плейстоцена день среднего человека мог быть насыщен событиями больше, чем сейчас.

Если в ваших руках оказался бы волшебный рычаг, позволяющий менять скорость макроструктурного развития, что бы вы сделали? Ускорили бы его, затормозили или оставили как есть?

Объективная точка зрения предполагает, что для ответа на этот вопрос потребуется оценить изменение уровня экзистенциального риска. Но вначале давайте определимся с различием между двумя его типами: статическим и динамическим. Статический риск ассоциируется с пребыванием в определенном состоянии, и общий размер статического риска системы есть прямая функция того, насколько долго система остается в этом состоянии. Все природные риски, как правило, статические: чем дольше мы живем, тем выше вероятность гибели в результате попадания астероида, извержения супервулкана, вспышки гамма-излучения, начала эпидемии или какого-то другого взмаха космической смертоносной косы. Некоторые антропогенные риски тоже статические. На индивидуальном уровне чем дольше солдат выглядывает из-за бруствера, тем выше кумулятивная вероятность погибнуть от пули снайпера. Статические риски есть и на экзистенциальном уровне: чем дольше мы живем в глобально анархической системе, тем выше кумулятивная вероятность термоядерного Армагеддона или мировой войны с применением неядерного оружия массового поражения — и то и другое способно уничтожить цивилизацию.

Динамический риск, в отличие от статического, это дискретный риск, связанный с неким необходимым или желательным переходом из одного состояния в другое. Как только переход завершен, риск исчезает. Величина динамического риска, связанного с переходом, обычно не является простой функцией его длительности. Риск пересечения минного поля не уменьшится вдвое, если бежать по нему в два раза быстрее. С быстрым взлетом в ходе создания сверхразума связан динамический риск, зависящий от качества проведенной подготовки, но не от того, потребуется на взлет двадцать миллисекунд или двадцать часов.

То есть о гипотетическом ускорителе макроструктурного развития можно сказать следующее.

1. В той мере, в которой нас беспокоит статический риск, нам следует стремиться к быстрому ускорению — при условии, что у нас есть реалистичные перспективы дожить до постпереходной эпохи,

в которой все остальные экзистенциальные риски будут значительно снижены.

2. Если бы мы точно знали, что некий шаг вызовет экзистенциальную катастрофу, нам следовало бы снизить скорость макроструктурного развития (или даже обратить его вспять), чтобы дать большему количеству поколений шанс прожить жизнь до того момента, пока не опустится занавес. Но на самом деле было бы слишком пессимистично считать, что человечество обречено.
3. В настоящее время статический экзистенциальный риск, похоже, относительно низок. Если предположить, что технологические макроусловия нашей жизни останутся на нынешнем уровне, вероятность экзистенциальной катастрофы, скажем, в ближайшее десятилетие, очень мала. Поэтому с задержкой технологического развития на десять лет — при условии, что она будет сделана на нынешней стадии развития или в какой-то другой момент, когда статический риск окажется столь же низким, — будет связан лишь незначительный экзистенциальный статический риск, притом что задержка вполне способна благотворно повлиять на величину экзистенциального динамического риска, в частности, из-за большего времени на подготовку к переходу.

Делаем вывод: скорость макроструктурного развития важна постольку, поскольку влияет на то, насколько подготовлено будет человечество к тому моменту, когда столкнется с ключевыми динамическими рисками[8].

Поэтому мы должны спросить себя, как когнитивное совершенствование (и соответствующее ускорение макроструктурного развития) повлияет на ожидаемый уровень подготовки в момент перехода. Следует ли нам предпочесть более короткий период подготовки на более высоком уровне интеллекта? С более высоким интеллектом время подготовки можно использовать более эффективно, а последний критически важный шаг сделает более разумное человечество. Или нам следует предпочесть действия на том уровне интеллекта, который более близок сегодняшнему, если это даст нам больше времени на подготовку?

То, какой вариант лучше, зависит от природы вызова, к которому мы готовимся. Если он заключается в решении проблемы, для которой ключевым является обучение на основе опыта, тогда определяющим фактором может быть длительность периода подготовки, поскольку для накопления

требуемого опыта нужно время. Как мог бы выглядеть подобный вызов? В качестве гипотетического примера можно предположить, что в какой-то момент в будущем будет создано новое оружие, способное в случае начала войны вызвать экзистенциальную катастрофу, скажем, с вероятностью один к десяти. Если мы столкнулись бы с таким вызовом, то могли бы захотеть замедлить темпы макроструктурного развития, чтобы у нашего вида было больше времени для совместных действий до того момента, когда будет сделан критически важный шаг и новое оружие станет реальностью. Можно надеяться на то, что благодаря этой отсрочке, подаренной торможением развития, человечество научится обходиться без войн, например, международные отношения на Земле станут напоминать те, что характерны для стран Евросоюза, которые яростно боролись друг с другом на протяжении многих веков, а теперь сосуществуют в мире и относительной гармонии. Такое умиротворение может стать следствием или постепенного смягчения нравов в результате различных эволюционных процессов, или шоковой терапии в результате вспышки насилия, недотягивающей до экзистенциального уровня (например, локального ядерного конфликта и вызванного им изменения во взглядах, способного привести к созданию глобальных институтов, необходимых для исключения войн между государствами). Если такого рода обучение не может быть ускорено за счет когнитивного совершенствования, тогда в таком улучшении смысла нет, поскольку оно лишь ускорит горение бикфордова шнура.

Однако с потенциальным взрывным развитием искусственного интеллекта может быть связан вызов и иного рода. Для решения проблемы контроля требуется проницательность, здравый смысл и теоретические знания. Не очень понятно, чем тут может помочь дополнительный исторический опыт. Получить непосредственный опыт переживания взрывного развития невозможно (пока не будет слишком поздно), да и многие другие скрытые от глаз аспекты проблемы контроля делают ее уникальной и не позволяют использовать какие-то исторические прецеденты. По этим причинам количество времени, которое пройдет до взрывного развития искусственного интеллекта, мало что значит само по себе. Но, вполне вероятно, приобретут значение следующие обстоятельства: 1) интеллектуальный прогресс в изучении проблемы контроля, достигнутый к моменту детонации; 2) навыки и интеллект, доступные в этот момент, чтобы реализовать наилучшие из имеющихся решения (или сымпровизировать в случае их отсутствия)[9].

Очевидно, что последний фактор должен положительно повлиять на когнитивное улучшение. Как когнитивное улучшение повлияет на первый фактор — пока неясно.

Одна из причин, по которым когнитивное совершенствование способно привести к прогрессу в проблеме контроля к моменту, когда произойдет взрывное развитие искусственного интеллекта, состоит в том, что этот прогресс может быть достижим лишь на экстремально высоких уровнях интеллектуальной производительности — даже более высоких, чем те, что требуются для создания машинного интеллекта. При решении проблемы контроля роль проб и ошибок, а также накопления экспериментальных результатов не так велика, как в работе над искусственным интеллектом или в исследованиях эмуляции головного мозга. То есть относительная важность времени и умственных усилий сильно варьируется в зависимости от типа задачи, и прогресс в решении проблемы контроля, достигнутый за счет когнитивного совершенствования, может оказаться *бóльшим*, чем прогресс в деле создания машинного интеллекта.

Еще одна причина, почему когнитивное совершенствование способно обеспечить больший прогресс в решении проблемы контроля, состоит в том, что сама потребность в этом прогрессе будет скорее признана интеллектуально более развитыми обществами и индивидуумами. Чтобы понять, почему проблема контроля важна, и присвоить ей высший приоритет, нужно обладать прозорливостью и здравомыслием[10]. А для решения столь непривычной задачи требуется обладать редкостным разумом.

Эти соображения подводят нас к заключению о желательности когнитивного улучшения, особенно если придавать большое значение экзистенциальным рискам взрывного развития искусственного интеллекта. Оно также поможет снизить риски, связанные с другими вызовами, для чего требуются проницательность и умение абстрактно мыслить (в противоположность постепенной адаптации к происходящим в окружающем мире изменениям или процессу культурного созревания и строительства соответствующих институтов, который может растянуться на время жизни нескольких поколений).

Технологические связки

Предположим, решить проблему контроля в случае искусственного интеллекта довольно трудно, а по отношению к имитационным моделям

несколько легче, — поэтому в сторону создания машинного интеллекта предпочтительнее двигаться через исследования по полной эмуляции головного мозга. Мы еще вернемся к вопросу о том, безопаснее ли этот путь по сравнению с развитием искусственного интеллекта. Но пока нужно отметить, что даже если принять это допущение, из него вовсе не следует, что нам нужно торопить появление технологии эмуляции мозга. Одну из причин мы уже обсудили — лучше, чтобы сверхразум был создан скорее позже, чем раньше, потому что тогда у нас будет больше времени для решения проблемы контроля и достижения прогресса по другим важным направлениям. Поэтому если есть уверенность, что эмуляция головного мозга станет предтечей развития ИИ, было бы контрпродуктивно спешить с развитием этой технологии.

Но даже если оказалось бы, что нам выгодно как можно быстрее создать технологию эмуляции головного мозга, из этого вовсе *не следовало бы*, что мы должны торопить прогресс в этом направлении. Поскольку он может привести к созданию вовсе не имитационной модели мозга, а нейроморфного искусственного интеллекта — формы ИИ, копирующей некоторые аспекты организации коры головного мозга, но недостаточно точно воспроизводящей его нейронную функциональность, чтобы симуляция получилась близкой. Если такой нейроморфный ИИ хуже, чем ИИ, который мог бы быть создан в противном случае, — а есть все основания полагать, что это так, — и если, подстегивая прогресс в области эмуляции мозга, мы увеличиваем шансы создать подобный нейроморфный ИИ, тогда наши усилия в направлении *лучшего* исхода (имитационная модель мозга) приведут к *худшему* исходу (нейроморфный ИИ); если мы все-таки стремимся ко *второму по степени привлекательности* исходу (синтетический ИИ), то можем действительно его достичь.

Мы только что описали (гипотетический) случай того, что можно назвать технологической связкой[11]. Этот термин относится к условиям, в которых две технологии предсказуемо связаны во времени так, что развитие одной должно привести к развитию другой, в качестве или предтечи, или приложения, или следствия. Технологические связки следует принимать в расчет, когда мы используем принцип различного технологического развития: не очень правильно ускорять создание желательной технологии *Y*, если единственный способ ее получить — это создать чрезвычайно нежелательную технологию-предтечу *X*, или если в результате создания *Y* немедленно

появится связанная с ней чрезвычайно нежелательная технология *Z*. Прежде чем делать предложение любимому человеку, посмотрите на будущих родственников.

В случае эмуляции головного мозга прочность технологической связки вызывает вопросы. Во второй главе мы заметили, что хотя для прогресса в этом направлении нужно будет создать множество новых технологий, каких-то ярких теоретических прорывов не потребуется. В частности, нам не нужно понимать, как работает биологический механизм познания, достаточно лишь знать, как создавать компьютерные модели небольших участков мозга, таких как различные виды нейронов. Тем не менее в процессе развития способности моделировать человеческий мозг будет собрано множество данных о его анатомии, что даст возможность значительно улучшить функциональные модели нейронных сетей коры головного мозга. И тогда появятся хорошие шансы создать нейроморфный ИИ раньше полноценной имитационной модели мозга[12]. Из истории нам известно множество примеров, когда методы ИИ брали начало в области нейробиологии и даже обычной биологии. (Например, нейрон Маккаллока–Питтса, перцептроны, или персептроны, и другие искусственные нейроны и нейронные сети появились благодаря исследованиям в области нейроанатомии; обучение с подкреплением инспирировано бихевиоризмом; генетические алгоритмы — эволюционной теорией; архитектура поведенческих модулей и перцепционная иерархия — теориями когнитивистики о планировании движений и чувственном восприятии; искусственные иммунные системы — теоретической иммунологией; роевой интеллект — экологией колоний насекомых и других самоорганизующихся систем; реактивный и основанный на поведении контроль в робототехнике — исследованиями механизма передвижения животных.) Возможно, еще важнее, что есть множество важных вопросов, имеющих отношение к ИИ, на которые можно будет ответить, лишь изучая мозг дальше. (Например, каким образом мозг хранит структурированные представления в кратковременной и долговременной памяти? Как решается проблема связывания? Что такое нейронный код? Как в мозгу представляются концепции? Есть ли некая стандартная единица механизма обработки информации в коре головного мозга, вроде колонки кортекса, и если да, то какова ее схема и как ее функциональность зависит от этой схемы? Как такие колонки соединяются и обучаются?)

Скоро мы сможем больше сказать об относительной опасности эмуляции головного мозга, нейроморфного ИИ и синтетического ИИ, но уже сейчас

можно отметить еще одну важную технологическую связку между эмуляцией и ИИ. Даже если усилия в направлении эмуляции головного мозга действительно приведут к созданию имитационной модели мозга (а не нейроморфного ИИ) и даже если появление такой модели окажется безопасным, риск все-таки остается — риск, связанный со *вторым переходом*, переходом от имитационной модели к ИИ, который представляет собой гораздо более мощную форму машинного интеллекта.

Есть множество других примеров технологических связок, заслуживающих более глубокого анализа. Например, усилия в сфере полной эмуляции головного мозга ускорят прогресс нейробиологии в целом[13]. Быстрее станут развиваться технология детекции лжи, техники нейропсихологического манипулирования, методы когнитивного улучшения, различные направления медицины. Усилия в области когнитивного совершенствования (в зависимости от конкретного направления) будут иметь такие побочные эффекты, как быстрое развитие методов генетической селекции и генетического инжиниринга, причем для улучшения не только когнитивных способностей, но и других черт нашей личности.

Аргументация от противного

Мы выйдем на следующий уровень стратегической сложности, если примем в расчет то, что не существует идеально доброжелательного, рационального и универсального контролера над миром, который просто реализовывал бы то, что считается наилучшим вариантом действий. Любые абстрактные соображения о том, что «следовало бы сделать», должны быть облечены в форму конкретного сообщения, которое появится в атмосфере риторической и политической реальности. Его будут игнорировать, неправильно трактовать, искажать и приспосабливать для различных конфликтующих целей; оно будет скакать повсюду, как шарик в пинболе, вызывая действия и противодействия и становясь причиной целого каскада последствий, которые не будут иметь прямой связи с намерениями автора.

Проницательный агент мог бы предвидеть все это. Возьмем, например, следующую схему аргументации в пользу проведения исследований с целью создания потенциально опасной технологии *X*. (Один из примеров использования этой схемы можно найти в работах Эрика Дрекслера. В его случае *X* = молекулярная нанотехнология[14].)

1. Риск *X* высок.
2. Для снижения этого риска требуется выделить время на серьезную подготовку.
3. Серьезная подготовка начнется лишь после того, как к перспективе создания X начнут серьезно относиться широкие слои общества.
4. Широкие слои общества начнут серьезно относиться к перспективе создания *X* только в случае проведения масштабных исследований возможности его создания.
5. Чем раньше начнутся масштабные исследования, тем больше времени понадобится на создание *X* (поскольку в самом начале ниже уровень развития необходимых технологий).
6. Следовательно, чем раньше начнутся масштабные исследования, тем больше будет времени на серьезную подготовку и тем сильнее удастся снизить риск.
7. Следовательно, масштабные исследования в отношении *X* нужно начинать немедленно.

Если следовать этой логике рассуждений, то, начав с причины двигаться вперед медленно или вовсе остановиться — из-за высокого риска, связанного с *X*, — приходишь к прямо противоположному следствию.

Родственный этому тип аргументации сводится к тому, что нам следует — какая жестокость! — приветствовать мелкие и средние катастрофы на том основании, что они вскроют наши уязвимые места и заставят принять меры предосторожности, снижающие вероятность экзистенциальной катастрофы. Идея состоит в том, что мелкие и средние катастрофы служат своего рода прививкой: сталкиваясь с относительно слабой угрозой, цивилизация вырабатывает иммунитет, благодаря которому сможет справиться с потенциально губительным вариантом той же самой угрозы[15].

По сути, речь идет о призывах к так называемой шоковой терапии, когда оправдывается нечто плохое в надежде, что это сможет вызвать нужную реакцию общества. Мы упомянули о ней не потому, что одобряем ее, а для того чтобы познакомить вас с идеей того, что называем аргументацией от противного. Она подразумевает, что если относиться к другим людям как к иррациональным агентам и играть на их когнитивных искажениях и ошибочных суждениях, то можно добиться от них более адекватной реакции, чем в случае, когда говоришь о проблеме прямо и обращаешься к их рассудку.

Скорее всего, использовать стратегию аргументации от противного для достижения глобальных долгосрочных целей будет невероятно трудно. Кто сможет предсказать результирующее воздействие сообщения после его скачков влево-вправо и вверх-вниз в пинболе общественного дискурса? Для этого пришлось бы смоделировать риторический эффект на миллиарды его составляющих, для которых характерны свои уникальные черты и степень влияния на события, меняющиеся на протяжении долгого времени, в течение которого извне на систему будут действовать непредсказуемые события, а изнутри ее топология будет претерпевать непрерывную эндогенную реорганизацию — задача явно неразрешимая![16] Однако может статься, что нам не придется детально прогнозировать всю траекторию будущего развития системы, дабы обзавестись уверенностью, что наше вмешательство в нужный момент увеличит вероятность достижения ею нужной нам долгосрочной цели. Можно было бы, например, сосредоточить внимание лишь на сравнительно краткосрочных и предсказуемых этапах этой траектории, выбирая такие действия, которые облегчат их прохождение, и рассматривая поведение системы за пределами горизонта прогнозирования как случайное блуждание.

Существуют, однако, некоторые этические основания для отказа от шагов в пределах аргументации от противного. Попытки перехитрить других выглядят как игра с нулевой суммой или даже отрицательной, если учесть время и энергию, которых понадобится намного больше как на сами эти попытки, так и потому, что всем становится сложнее понять, что же на самом деле думают другие, и добиться доверия к своим собственным суждениям[17]. Повсеместное использование методов стратегического манипулирования убило бы искренность, а истина была бы обречена блуждать в коварном мраке политических игр.

Пути и возможности

Следует ли нам приветствовать успехи в создании аппаратного обеспечения? А успехи на пути к созданию компьютерной модели мозга? Давайте ответим на эти вопросы по очереди.

Последствия прогресса в области аппаратного обеспечения

Повышение быстродействия компьютеров облегчает создание машинного интеллекта. Таким образом, одним из эффектов от прогресса в области аппаратного обеспечения является ускорение момента, когда на свет появится

машинный интеллект. Мы уже говорили о том, что с объективной точки зрения это может быть плохо, поскольку у человечества будет меньше времени на то, чтобы решить проблему контроля и подняться на более высокую ступень цивилизации. Впрочем, это не означает неминуемость катастрофы. Поскольку сверхразум устранит множество других экзистенциальных рисков, стоит предпочесть его раннее появление, если уровень этих рисков окажется слишком высоким[18].

Приближение или откладывание взрывного развития искусственного интеллекта — не единственный канал, посредством которого прогресс в области аппаратного обеспечения способен повлиять на экзистенциальный риск. Еще одна возможность заключается в том, что аппаратное обеспечение может до некоторой степени заменить программное: более совершенная аппаратная основа снижает требования к минимальному уровню навыков, которые требуются для создания кода зародыша ИИ. Быстрые компьютеры могут также стимулировать использование подходов, которые больше полагаются на полный перебор (вроде генетических алгоритмов и прочих методов, работающих по схеме «генерация–оценка–уничтожение») и меньше — на техники, требующие глубокого понимания ситуации. Если в результате использования методов полного перебора, как правило, возникают системы с более беспорядочной и менее точной архитектурой, где проблему контроля решить труднее, чем в системах с более точно спроектированной и теоретически предсказуемой архитектурой, то это означает еще одну причину, по которой более быстрые компьютеры увеличивают экзистенциальный риск.

Еще одно соображение состоит в том, что быстрый прогресс в области аппаратного обеспечения повышает вероятность быстрого взлета. Чем больших успехов добивается индустрия микропроцессоров, тем меньше человекочасов требуется программистам для вывода возможностей компьютеров на новый уровень производительности. Это значит, что взрывное развитие искусственного интеллекта вряд ли произойдет на самом нижнем уровне производительности аппаратного обеспечения. То есть *более* вероятно, что такой взрыв случится тогда, когда компьютеры окажутся на гораздо более высоком уровне развития. Это означает, что взлет произойдет в условиях «аппаратного навеса». Как мы видели в четвертой главе, аппаратный навес — один из многих факторов, снижающих сопротивляемость системы в процессе взлета. А значит, быстрый прогресс в области аппаратного

обеспечения, скорее всего, приведет к более быстрому и взрывоопасному переходу к сверхразуму.

Быстрый взлет за счет так называемого аппаратного навеса может разными путями оказывать влияние на риски перехода. Наиболее очевидно, что при быстром взлете остается меньше возможностей отреагировать и внести корректировки в процесс перехода, что повышает связанные с ним риски. Кроме того, в случае аппаратного навеса снижаются шансы, что потенциально опасный зародыш ИИ, приступивший к самосовершенствованию, будет удачно изолирован и не допущен к достаточному объему вычислительных ресурсов, поскольку чем быстрее процессоры, тем меньше их требуется, чтобы ИИ быстро развился до уровня сверхразума. Еще один эффект аппаратного навеса заключается в том, что он уравнивает шансы мелких и крупных проектов, снижая преимущества крупных в более мощном компьютерном парке. Он также повышает экзистенциальный риск, поскольку в рамках крупных проектов выше вероятность решить проблему контроля и установить более приемлемые с этической точки зрения цели[19].

У быстрого взлета есть свои преимущества. Он увеличивает вероятность того, что будет сформирован синглтон. Если формирование синглтона важно для решения проблемы постпереходной координации, может иметь смысл пойти на более высокий риск в ходе взрывного развития искусственного интеллекта, чтобы снизить риск катастрофических последствий провала координационных усилий после него.

Успехи в сфере вычислений могут повлиять на исход революции машинного интеллекта не только прямо, но и косвенно, благодаря своему влиянию на общество и опосредованному участию в формировании условий для взрывного развития искусственного интеллекта. Скажем, интернет, для развития которого требуется массовое производство недорогих компьютеров, влияет на деятельность людей во многих областях, включая работу над созданием искусственного интеллекта и исследования по проблеме контроля. (Эта книга, скорее всего, не была бы написана, и вы могли бы остаться без нее, не будь интернета.) Однако аппаратное обеспечение уже довольно хорошо развито, чтобы появились отличные приложения, облегчающие людям работу, общение, размышления, — поэтому неясно, действительно ли прогресс в этих областях сдерживается скоростью его совершенствования[20].

Итак, представляется, что с объективной точки зрения более быстрый прогресс в области развития аппаратного обеспечения нежелателен. Это

предварительное заключение может быть опровергнуто, например, если окажется, что другие экзистенциальные риски или риски, связанные с провалом постпереходной координации, окажутся чрезвычайно высокими. В любом случае, похоже, трудно еще сильнее полагаться на скорость совершенствования оборудования. А значит, наши усилия по улучшению исходных условий взрывного развития искусственного интеллекта следует сосредоточить на других параметрах общего стратегического процесса.

Обратите внимание, что даже если мы не знаем, как влиять на тот или иной параметр, может быть полезно для начала работы над стратегией хотя бы определить «знак» этого влияния (то есть понять, что лучше — увеличивать или уменьшать значение параметра). А точку опоры для воздействия на выбранный параметр можно будет отыскать и позже. Возможно также, что знак одного параметра коррелирует со знаком какого-то другого, которым легче манипулировать, и тогда наш исходный анализ помог бы нам решить, что с ним делать.

Следует ли стимулировать исследования в области полной эмуляции головного мозга?

Чем труднее кажется проблема контроля в случае искусственного интеллекта, тем более соблазнительным представляется движение по пути создания имитационной модели мозга в качестве менее рискованной альтернативы. Однако есть некоторые сложности, которые стоит проанализировать, прежде чем делать окончательные выводы[21].

Прежде всего, это наличие технологических связок, которые мы уже обсуждали. Было отмечено, что усилия по созданию имитационной модели мозга могут привести к появлению нейроморфного ИИ, то есть особенно опасной формы машинного интеллекта.

Но на время допустим, что мы действительно добились своей цели и создали компьютерную имитационную модель головного мозга (далее по тексту — КИМГМ). Будет ли она безопаснее ИИ? Это сложный вопрос. У КИМГМ есть минимум три *предполагаемых* преимущества перед ИИ: 1) ее характеристики производительности легче понять; 2) ей внутренне присущи человеческие мотивы; 3) в случае ее первенства взлет будет медленным. Предлагаю коротко рассмотреть каждый фактор.

1. То, что характеристики производительности КИМГМ понять легче, чем ИИ, звучит убедительно. У нас в изобилии информации о сильных

и слабых сторонах человеческого интеллекта, но нет никаких знаний об ИИЧУ. Однако понимать, что может или не может делать оцифрованный моментальный снимок человеческого интеллекта, не то же самое, что понимать, как этот интеллект будет реагировать на изменения, направленные на повышение его производительности. В отличие от него, ИИ можно изначально тщательно проектировать таким образом, чтобы он был понятен как в статическом, так и в динамическом состоянии. Поэтому хотя интеллектуальная производительность КИМГМ может оказаться более предсказуемой на сравнимых стадиях их разработки, неясно, будет ли она динамически более предсказуемой, чем ИИ, созданный компетентными и озабоченными проблемой безопасности программистами.

2. Нет никакой гарантии, что КИМГМ будет непременно присуща мотивация ее человеческого прототипа. Чтобы скопировать оценочные характеристики человеческой личности, может потребоваться слишком высокоточная модель. Даже *если* мотивационную систему отдельного человека удастся скопировать точно, неясно, насколько безопасным это окажется. Люди могут быть лживыми, эгоистичными и жестокими. И хотя можно надеяться, что прототип будет выбран за его исключительную добродетель, трудно спрогнозировать, как он станет действовать после переноса в совершенно чуждую ему среду, после повышения его интеллектуальных способностей до сверхчеловеческого уровня и с учетом возможности захватить господство над миром. Верно лишь то, что у КИМГМ будет более *человекоподобная* мотивация (они не станут ценить лишь скрепки или знаки после запятой в числе пи). В зависимости от ваших взглядов на человеческую природу это может как обнадеживать, так и разочаровывать[22].
3. Неясно, почему технология полной эмуляции головного мозга может привести к более медленному взлету, чем технология разработки искусственного интеллекта. Возможно, в случае эмуляции мозга можно ожидать меньшего аппаратного навеса, поскольку она менее эффективна с вычислительной точки зрения, чем ИИ. Возможно, также, что ИИ способен легче превратить всю доступную ему вычислительную мощность в один гигантский интегрированный интеллект, в то время как КИМГМ будут предшествовать появлению сверхразума и опережать

человечество лишь с точки зрения быстродействия и возможного количества копий. Если появление КИМГМ приведет к более медленному взлету, то дополнительным преимуществом будет ослабление проблемы контроля. Кроме того, более медленный взлет повысит вероятность многополярного исхода. Но настолько ли он желателен, этот многополярный мир, — большой вопрос.

Есть еще одно соображение, которое ставит под сомнение идею безопасности создания вначале технологии эмуляции головного мозга, — необходимость решать проблему *второго перехода*. Даже если первый машинный интеллект человеческого уровня будет создан в форме эмулятора, возможность появления искусственного интеллекта останется. ИИ в его зрелой форме обладает важными преимуществами по сравнению с КИМГМ, которые делают его несравнимо более мощной технологией[23]. Поскольку зрелый ИИ приведет к моральному устареванию КИМГМ (за исключением специальной задачи по консервации мозга отдельных людей), обратное движение вряд ли возможно.

Это означает следующее: если ИИ будет разработан первым, возможно, что у взрывного развития искусственного интеллекта будет всего одна волна; а если первым будет создана КИМГМ, таких волн может быть две. Первая — появление самой КИМГМ, вторая — появление ИИ. Совокупный экзистенциальный риск на пути КИМГМ–ИИ представляет собой *сумму* рисков, связанных с первым и вторым переходом (обусловленную успешным завершением первого), см. рис. 13[24].

Насколько более безопасным окажется переход к ИИ в мире КИМГМ? Одно из соображений заключается в том, что переход к ИИ мог бы быть менее взрывным, если случился бы уже после появления какой-то иной формы машинного интеллекта. Имитационные модели, работающие на цифровых скоростях и в количествах, далеко превосходящих население Земли, могли бы сократить когнитивный разрыв и легче контролировать ИИ. Впрочем, этому соображению не стоит придавать слишком большое значение, поскольку разрыв между ИИ и КИМГМ все-таки будет очень значительным. Однако если КИМГМ не просто быстрее и многочисленнее, но еще и качественно умнее людей (или как минимум не уступают лучшим представителям человечества), тогда у сценария с КИМГМ будут преимущества, аналогичные сценарию с когнитивным улучшением биологического мозга, который мы рассматривали выше.

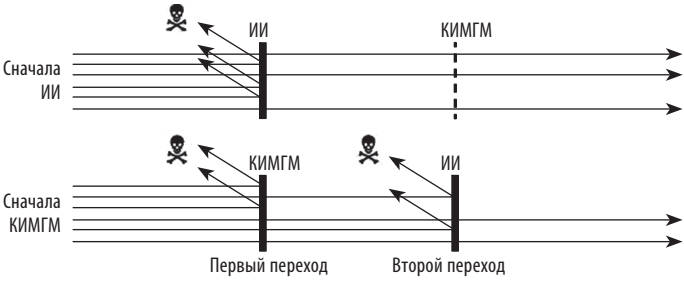

Рис. 13. Что раньше, искусственный интеллект или компьютерная имитационная модель головного мозга? В сценарии «сначала КИМГМ» есть два перехода, связанных с риском, — вначале в результате создания КИМГМ, затем — ИИ. Совокупный экзистенциальный риск этого сценария представляет собой сумму рисков каждого перехода. Однако риск самого перехода к ИИ может быть ниже, когда это происходит в мире, где уже успешно работает КИМГМ.

Второе соображение касается ситуации конкуренции проектов, в этом случае при создании КИМГМ лидер может упрочить свое преимущество. Возьмем сценарий, в котором у лидирующего проекта по созданию КИМГМ есть шестимесячная фора по сравнению с ближайшим преследователем. Предположим, что первые работающие эмуляторы будут безопасными, терпеливыми и станут с готовностью сотрудничать с людьми. Если они работают на быстром оборудовании, то смогут потратить целую субъективную вечность в поисках ответа на вопрос, как создать безопасный ИИ. Например, если они работают на скоростях, в сто тысяч раз превышающих человеческие, и способны, не отвлекаясь, заниматься проблемой контроля в течение шести месяцев звездного времени, то смогут продвинуться в ее решении на пятьдесят тысяч лет дальше к тому моменту, когда столкнутся с конкуренцией со стороны эмуляторов, созданных в рамках проекта ближайшего конкурента. При наличии доступного аппаратного обеспечения они смогут еще ускорить свой прогресс, запустив миллиарды собственных копий независимо работать над различными подзадачами. А если лидирующий проект использует свое шестимесячное преимущество для формирования синглтона, то сможет купить своей команде по разработке ИИ неограниченное количество времени для решения проблемы контроля[25].

В целом возникает ощущение, что в случае, если КИМГМ появится раньше ИИ, риск перехода к ИИ может быть ниже. Однако если сложить остаточный

риск перехода к ИИ с риском предшествующего ему перехода к КИМГМ, становится не столь понятно, как соотносится совокупный экзистенциальный риск сценария «сначала КИМГМ» с риском сценария «сначала ИИ». Получается, что если смотреть скептически на способность человечества управлять переходом к ИИ — приняв во внимание, что человеческая природа или цивилизация могут улучшиться к тому моменту, когда мы столкнемся с этим вызовом, — то сценарий «сначала КИМГМ» кажется более привлекательным.

Чтобы понять, следует ли продвигать идею создания технологии полной эмуляции головного мозга, нужно учесть еще несколько важных моментов.

Во-первых, следует помнить об уже упомянутой связке технологий: движение в сторону КИМГМ может привести к созданию нейроморфного ИИ. Это причина не настаивать на эмуляторах[26]. Несомненно, есть *некоторые* варианты синтетического ИИ, которые менее безопасны, чем *некоторые* нейроморфные варианты. Однако в среднем кажется, что нейроморфные системы более опасны. Одна из причин этого в том, что имитация может подменить собой знание. Чтобы построить нечто с нуля, обычно нужно сравнительно хорошо представлять, как будет работать система. Но в таком понимании нет нужды, если всего лишь копируешь какие-то свойства уже существующей системы. Полная эмуляция головного мозга полагается на масштабное копирование биологической системы, что не требует досконального понимания механизмов познания (хотя, без сомнения, понадобятся серьезные знания на уровне ее компонентов). В этом смысле нейроморфный ИИ может походить на КИМГМ: его можно получить путем плагиата различных элементов без необходимости того, чтобы инженеры имели глубокое математическое понимание принципов ее работы. Зато нейроморфный ИИ *не будет похож* на КИМГМ в другом аспекте — у него не будет по умолчанию человеческой системы мотивации[27]. Это соображение говорит против стремления двигаться по пути КИМГМ, если в результате может появиться нейроморфный ИИ.

Во-вторых, о чем нужно помнить: КИМГМ, скорее всего, подаст нам сигнал о своем скором появлении. В случае ИИ всегда есть возможность неожиданного концептуального прорыва. В отличие от него, для успеха эмуляции головного мозга требуется совершить множество трудоемких предварительных шагов: создать высокопроизводительные мощности для сканирования и программное обеспечение для обработки изображений,

проработать алгоритм моделирования нейронной сети. Поэтому можно уверенно утверждать, что КИМГМ еще не на пороге (и не будет создан в ближайшие двадцать–тридцать лет). Это значит, что усилия по ускорению создания КИМГМ имеют значение лишь для тех сценариев, где машинный интеллект появляется сравнительно нескоро. Это делает инвестиции в КИМГМ привлекательными для тех, кто хотел бы, чтобы взрывное развитие интеллекта исключило другие экзистенциальные риски, но опасается поддерживать разработку ИИ из страха, что этот взрыв произойдет преждевременно, до того как будет решена проблема контроля. Однако, похоже, неопределенность относительно сроков этих событий настолько велика, что не стоит придавать этому соображению слишком большой вес[28].

Поэтому стратегия поддержки КИМГМ более привлекательна, если: 1) вы с пессимизмом смотрите на способность людей решить проблему контроля над ИИ; 2) не слишком беспокоитесь по поводу нейроморфного ИИ, многополярных исходов и рисков второго перехода; 3) думаете, что КИМГМ и ИИ будут созданы довольно скоро; 4) предпочитаете, чтобы сверхразум появился не слишком рано и не слишком поздно.

С субъективной точки зрения — лучше быстрее

Боюсь, что под словами одного из комментаторов в блоге могут подписаться многие:

> Мне инстинктивно хочется двигаться быстрее. И не потому что, как мне кажется, мир от этого выиграет. Почему я должен думать о мире, если все равно умру и превращусь в прах? Мне просто хочется, чтобы все двигалось быстрее, черт возьми! Это увеличивает шансы, что я застану более технологически продвинутое будущее[29].

С субъективной точки зрения у нас есть серьезная причина торопить появление любых радикально новых технологий, с которыми связан экзистенциальный риск. Причина в том, что все, кто сейчас живет, по умолчанию умрут в течение ближайших ста лет.

Особенно сильно желание торопить события, когда речь идет о технологиях, способных продлить нашу жизнь и тем самым увеличить количество ныне живущих людей, способных застать взрывное развитие искусственного интеллекта. Если революция машинного интеллекта пойдет так, как нам хотелось бы, появившийся в ее результате сверхразум почти наверняка найдет способ бесконечно продлевать жизнь людей, которые окажутся его

современниками, причем не просто продлевать жизнь, но и делать ее абсолютно качественной, поскольку люди будут совершенно здоровы и полны молодой энергией — тоже благодаря стараниям сверхразума. Более того, сверхразум увеличит возможности человека далеко за пределы того, что можно было бы предположить; кто знает, вдруг он поможет человеку совсем избавиться от своего бренного организма, загрузит его мозг в цифровую среду и подарит ему идеально здоровое виртуальное тело. Что касается технологий, не обещающих продление жизни, то здесь оснований для спешки меньше, но они все равно есть, поскольку появление таких возможностей способно повысить стандарты жизни[30].

Благодаря такой же логике с субъективной точки зрения кажутся привлекательными многие рискованные технологические инновации, обещающие приблизить момент взрывного развития искусственного интеллекта, даже если они нежелательны с объективной точки зрения. Такие инновации могут сократить «час волка», в течение которого мы должны сидеть на своем насесте в ожидании рассвета постчеловеческой эпохи. Поэтому с субъективной точки зрения быстрый прогресс в области аппаратного обеспечения кажется таким же желательным, как и в области создания КИМГМ. Потенциальный вред из-за повышения экзистенциального риска уравновешивается возможной пользой для отдельно взятого человека благодаря возросшим шансам, что взрывное развитие искусственного интеллекта произойдет еще при жизни живущих сейчас людей[31].

Сотрудничество

Важным параметром является степень сотрудничества и координации, которую удастся обеспечить в ходе создания машинного интеллекта. Сотрудничество способно принести много пользы. Посмотрим, какое влияние этот параметр мог бы оказать на процесс создания машинного интеллекта и какие рычаги имеются в нашем распоряжении для расширения и углубления сотрудничества.

Гонка и связанные с ней опасности

Ощущение гонки возникает, когда есть страх, что проект обойдут конкуренты. Для этого совсем не обязательно наличие множества сходных проектов. Ситуация, при которой инициаторы проекта находятся в состоянии гонки, может сложиться и в отсутствие реальных конкурентов. Возможно,

союзники не создали бы атомную бомбу так быстро, если не считали бы (ошибочно), что Германия крайне близка к этой цели.

Интенсивность ощущения гонки (то есть степень того, насколько конкуренты готовы ставить скорость выше безопасности) зависит от нескольких факторов, таких как плотность конкуренции, относительная важность возможностей и удачи, количество соперников, схожесть их подходов и целей. Также важно, что думают обо всех этих факторах конкуренты (см. врезку 13).

В случае развития машинного интеллекта представляется, что ощущение гонки будет как минимум заметным, а возможно, и очень сильным. От интенсивности этого ощущения зависит, какими могут быть стратегические последствия возможного взрывного развития искусственного интеллекта.

ВРЕЗКА 13. СМЕРТЕЛЬНАЯ ГОНКА

Рассмотрим гипотетическую гонку вооружений с применением ИИ — гонку, в которой несколько команд конкурируют за право первыми создать сверхразум[32]. Каждая из них сама решает, сколько инвестировать в безопасность, понимая, что ресурсы, потраченные на меры предосторожности, — это ресурсы, не потраченные на создание ИИ. В отсутствие согласия между соперниками (которого не удалось достичь из-за различия позиций или невозможности контролировать соблюдение договора) гонка может стать смертельно опасной, когда каждая команда тратит на безопасность лишь минимальные средства.

Производительность каждой команды можно представить как функцию ее возможностей (к которым относится и удача), штрафной функцией являются затраты на обеспечение безопасности. Первой создаст ИИ команда с наивысшей производительностью. Риски, связанные с появлением ИИ, зависят от того, сколько его создатели инвестировали в безопасность. В наихудшем сценарии все команды имеют одинаковые возможности. В этом случае победитель определяется исключительно по величине его капиталовложений в безопасность: выиграет команда, потратившая на меры предосторожности меньше всего. Тогда равновесие Нэша в этой игре достигается в ситуации, когда ни одна команда ничего не тратит на безопасность. В реальном мире такая ситуация будет означать возникновение *эффекта храповика*: одна из команд принимает на себя больший риск, опасаясь отстать от конкурентов, последние отвечают тем же — и так несколько раз, пока уровень риска не оказывается максимальным.

Возможности и риск

Ситуация меняется, когда возможности команд не одинаковы. Поскольку различия в возможностях являются более важным фактором по сравнению с затратами на обеспечение безопасности, эффект храповика слабеет: стимулов идти на больший риск

в ситуации, когда это не повлияет на расстановку сил, гораздо меньше. Различные сценарии такого рода показаны на рис. 14, иллюстрирующем риски ИИ в зависимости от значимости такого параметра, как возможности разрабатывающих его команд. Инвестиции в безопасность лежат в диапазоне от 1 (в результате получаем идеально безопасный ИИ) до 0 (совершенно небезопасный ИИ). По оси *x* отображается относительная значимость возможностей команды для определения ее прогресса на пути создания ИИ по сравнению с инвестициями в безопасность. (В точке 0,5 уровень инвестиций в безопасность в два раза значимее возможностей; в точке 1 они равны; в точке 2 возможности в два раза значимее инвестиций в безопасность и так далее.) По оси *y* отображается уровень риска, связанный с ИИ (ожидаемая доля максимальной полезности, которую получает победитель.)

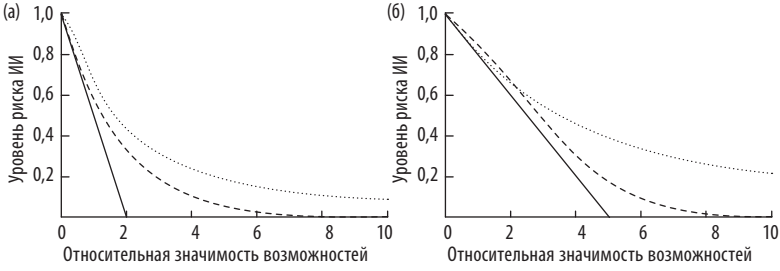

Рис. 14. Уровни риска в условиях гонки технологий искусственного интеллекта. На рисунке изображен уровень риска опасного ИИ для простой модели гонки технологий с участием а) двух или б) пяти команд в сочетании с относительной значимостью их возможностей (по сравнению с инвестициями в безопасность) для определения того, какой проект станет победителем. На диаграмме отражены три сценария: сплошная линия — нет информации об уровне возможностей; штриховой пунктир — закрытая информация о возможностях; точечный пунктир — открытая информация о возможностях.

Мы видим, что во всех сценариях опасность ИИ максимальна, когда возможности не играют никакой роли, и постепенно снижается по мере роста их значимости.

Сравнимые цели

Еще один способ снизить риск заключается в том, чтобы обеспечить командам большую долю в успехе друг друга. Если конкуренты убеждены, что второе место означает потерю всего, что им дорого, они пойдут на любой риск, чтобы обойти соперников. И наоборот, станут больше инвестировать в безопасность, если окажутся менее зависимыми от результатов гонки. Это означает, что нам нужно поощрять различные формы перекрестного инвестирования.

Количество конкурентов

Чем больше конкурирующих команд, тем более опасной становится гонка: у каждой из команд меньше шансов на то, чтобы прийти первой, соответственно, выше соблазн рисковать. Это видно, если сравнить позиции *а* и *б* на рис. 14: две команды и пять команд. В каждом сценарии риск растет с ростом числа конкурентов. Его можно снизить, если команды объединятся в небольшое количество конкурирующих коалиций.

Проклятие избыточной информации

Хорошо ли, если команды будут знать о своем месте в гонке (например, уровень своих возможностей)? И да, и нет. Желательно, чтобы о своем лидерстве знала сильнейшая команда (это будет означать, что отрыв от конкурентов позволит ей больше думать о безопасности). И нежелательно, чтобы о своем отставании знали остальные (поскольку это подтвердит их решимость ослабить меры предосторожности в надежде нагнать конкурентов). Хотя на первый взгляд может показаться, что компромисс возможен, модели недвусмысленно показывают, что информация — это плохо[33]. На рис. 14 (*а* и *б*) отражены три сценария: прямая линия соответствует ситуации, в которой ни одна из команд не имеет информации о возможностях участников гонки, включая свои собственные; штриховой пунктир соответствует ситуации, в которой команды знают только о своих собственных возможностях (тогда они готовы идти на дополнительный риск в случае, если их возможности низки); точечный пунктир показывает, что происходит, если все команды осведомлены о возможностях друг друга (они могут пойти на дополнительный риск в случае, если их возможности близки). С каждым ростом уровня информированности гонка обостряется.

Из-за ощущения гонки проекты могут ускорить свое движение в сторону сверхразума, сократив капиталовложения в решение проблемы контроля. Возможны и другие негативные последствия вроде прямых враждебных действий по отношению к конкурентам. Предположим, что две страны соревнуются в том, какая из них первой создаст сверхразум, и одна вырывается вперед. В ситуации, когда «победитель получает все», аутсайдер может решиться на отчаянный удар по сопернику, вместо того чтобы пассивно ждать поражения. Предполагая это, страна-лидер может нанести упреждающий удар. Если антагонисты обладают достаточной военной мощью, их столкновение приведет к большой крови[34]. (Даже точечный удар, нанесенный по инфраструктуре проекта разработки ИИ, несет риск более широкой конфронтации, кроме того, он может не достичь своей цели, если страна, в которой ведутся работы над этим проектом, примет соответствующие меры предосторожности[35].)

Если взять сценарий, в котором враждующие разработчики представляют не страны, а менее мощные институты, их конфликт, по всей вероятности, окажется менее разрушительным с точки зрения непосредственного вреда. Хотя в целом последствия такой конкуренции будут почти столь же плохи. Это связано с тем, что главным образом вред вызывается не столкновением на поле битвы, а ослаблением мер предосторожности. Как мы видели, ощущение гонки приведет к снижению инвестиций в безопасность, а конфликт, даже бескровный, устранит возможность сотрудничества, поскольку в атмосфере враждебности и недоверия проектные группы вряд ли захотят делиться идеями о путях решения проблемы контроля[36].

О пользе сотрудничества

Итак, сотрудничество дает много полезного. Оно избавляет от спешки при разработке машинного интеллекта. Позволяет больше вкладывать в безопасность. Избегать насильственных конфликтов. И облегчает обмен идеями в вопросах контроля. К этим преимуществам можно добавить еще одно: сотрудничество, скорее всего, приведет к такому исходу, в котором благоприятный эффект от контролируемого взрывного развития интеллекта будет распределен более равномерно.

Расширение сотрудничества приводит к более широкому распределению благ, но это не столь очевидно, как кажется. В принципе, небольшой проект под управлением альтруиста тоже может привести к исходу, в котором блага будут распределены равномерно или справедливо между всеми обладающими моральным статусом существами. Тем не менее есть несколько причин полагать, что более широкое сотрудничество, включая большее количество инвесторов, будет лучше (как ожидается) с точки зрения распределения результатов. Одна из причин заключается в том, что инвесторы предположительно предпочитают исход, в котором они сами получают справедливую долю. Тогда более широкое сотрудничество означает, что сравнительно много людей получат как минимум свою справедливую долю в случае успешного завершения проекта. Другая причина заключается в том, что более широкое сотрудничество выгодно даже людям, не имеющим непосредственного отношения к проекту. Чем шире сотрудничество, тем больше людей в него вовлечено и тем больше людей вне проекта связано с ними и может рассчитывать на то, что участники проекта учтут их интересы. Кроме того, чем шире сотрудничество, тем больше вероятность, что к нему будут

привлечены как минимум несколько альтруистов, стремящихся действовать на благо всех. Более того, такой проект с большей вероятностью будет объектом общественного контроля, что снизит риск присвоения всего пирога кликой программистов или частных инвесторов[37]. Заметьте также, что чем шире сотрудничество в работе над успешным проектом, тем ниже его издержки на распределение выгод среди людей, не имеющих к нему прямого отношения. (Например, если 90 процентов всех людей уже вовлечены в сотрудничество, то поднять всех остальных до своего уровня им будет стоить не больше 10 процентов их доли.)

Поэтому вполне возможно, что более широкое сотрудничество приведет к более широкому распределению благ (хотя *некоторые* проекты с небольшим количеством инвесторов тоже могут иметь отличные перспективы для их распределения). Но почему столь желательно широкое распределение благ?

Исход, в котором все получат свою долю сладостей, предпочтительнее с точки зрения как этики, так и благоразумия. Об этической стороне вопроса много говорить не будем, скажем лишь, что дело не обязательно в стремлении к эгалитаризму. Причина может быть, например, в желании справедливости. С проектом, в рамках которого создается машинный сверхразум, связаны глобальные риски гибели человечества. В смертельной опасности оказывается каждый житель Земли, включая тех, кто не согласен подвергать угрозе свои жизни и жизни своих близких. А поскольку риск разделяют все, требование минимальной справедливости означает, что и свою часть награды тоже должны получить все.

То, что общий (ожидаемый) объем благ, похоже, будет выше в сценариях сотрудничества, является еще одним важным этическим аргументом в их пользу.

С точки зрения благоразумия широкое распределение благ выгодно по двум причинам. Первая заключается в том, что такой подход к распределению благ приведет к расширению сотрудничества, что, в свою очередь, позволит снизить негативные последствия гонки технологий. Если от успеха проекта выиграют все, будет меньше поводов бороться за лавры первого создателя сверхразума. Спонсоры конкретного проекта могут также выиграть благодаря информированию общественности о готовности равномерно распределить выгоды от него, поскольку альтруистические проекты скорее привлекут больше сторонников и меньше противников[38].

Вторая причина предпочтения широкого распределения благ с точки зрения благоразумия заключается в том, что для агентов характерно стремление избежать риска, а их функция полезности нелинейна относительно ресурсов. Здесь главное то, что потенциальный выигрыш в ресурсах может быть колоссальным. Если предположить, что наблюдаемая Вселенная действительно так необитаема, какой выглядит, то на каждого жителя Земли сегодня приходится больше одной свободной галактики. Большинство людей предпочли бы гарантированный доступ к одной галактике, полной ресурсов, лотерейному билету с шансом один на миллиард стать владельцем миллиарда галактик[39]. Учитывая такие космические масштабы, ожидающие человечество в будущем, похоже, в наших интересах предпочесть сделку, в рамках которой каждому человеку была бы гарантирована доля, даже если она соответствует лишь малой части общего. Когда на горизонте маячит столь щедрый приз, важно не остаться с носом.

Этот аргумент о громадности призового фонда основан на предположении, что предпочтения являются ресурсно-насыщенными[40]. Это не обязательно так. Например, есть несколько известных этических теорий — включая особенно агрегированные консеквенциалистские, — с которыми соотносятся нейтральные к риску и линейные по ресурсам функции полезности. Имея миллиард галактик, можно создать в миллиард раз больше счастливых жизней, чем имея одну. То есть с утилитарной точки зрения это в миллиард раз лучше[41]. То есть функции предпочтения обычных эгоистичных людей, похоже, относительно ресурсно-насыщаемые. Последнее замечание нужно дополнить двумя важными уточнениями.

Первое: многих людей беспокоят рейтинги. Если множество агентов захотят возглавить список богатейших людей мира, никаких мировых запасов не хватит, чтобы удовлетворить каждого.

Второе: постпереходная технологическая база, возможно, позволит превращать материальные ресурсы в беспрецедентно широкий спектр продуктов, включая такие, которые недоступны сейчас ни за какие деньги, хотя и высоко ценимы многими. Миллиардеры не живут в тысячу раз дольше миллионеров. В эпоху цифрового разума, однако, миллиардеры смогут позволить себе в тысячу раз бóльшую вычислительную мощность и соответственно проживать субъективно в тысячу раз более долгую жизнь. Точно так же окажется доступной за деньги и ментальная мощность. В таких обстоятельствах, когда экономический капитал можно будет конвертировать

в жизненно важные товары по постоянному курсу даже на очень высоком уровне богатства, неограниченная жадность обретет гораздо больший смысл, чем в современном мире, где богачи (лишенные филантропических черт) озабочены приобретением самолетов, яхт, коллекций предметов искусства, бесчисленных резиденций — и на все это тратятся целые состояния.

Значит ли это, что эгоист должен нейтрально относиться к риску, связанному с его обогащением в постпереходный период? Не совсем. Может так получиться, что физические ресурсы не будут конвертироваться в жизни или ментальную мощность в произвольных масштабах. Если жизнь должна быть прожита последовательно, чтобы наблюдатель помнил прошлые события и ощущал последствия сделанного когда-то выбора, тогда жизнь цифрового мозга не может быть продлена произвольно долго без использования все возрастающего количества *последовательных* вычислительных операций. Но законы физики ограничивают степень, до которой ресурсы можно трансформировать в последовательные вычисления[42]. Пределы возможностей последовательных вычислений могут также довольно сильно ограничить некоторые аспекты когнитивной эффективности, которая не сможет расти линейно в соответствии с быстрым накоплением ресурсов. Более того, совершенно неочевидно, что эгоист должен быть нейтрален к риску даже в случае очень релевантных показателей успеха, скажем, количества скорректированных на качество субъективных лет жизни. Если будет стоять выбор между гарантированными дополнительными двумя тысячами лет жизни и одним к десяти шансом получить дополнительно тридцать тысяч лет жизни, думаю, большинство людей выберут первое (даже при условии, что все годы будут одинакового качества)[43].

В реальности аргумент, что с точки зрения здравомыслия предпочтительнее более широкое распределение благ, вероятно, имеет субъективный и зависящий от ситуации характер. Хотя в целом люди с большей вероятностью получили бы (почти все), что они хотят, если удастся найти способ обеспечить широкое распределение благ, — и это верно даже без учета того, что обязательство такого распределения подстегнет сотрудничество и тем самым снизит шансы экзистенциальной катастрофы. То есть широкое распределение благ представляется не только этически необходимым, но и выгодным.

Есть и другое следствие сотрудничества, которое нельзя не отметить хотя бы вскользь: это возможность того, что степень сотрудничества в преддверии перехода повлияет на степень сотрудничества после него.

Предположим, что человечество решило проблему контроля. (Если проблема контроля не решена, степень сотрудничества в постпереходный период будет значить пугающе много.) Тогда нужно рассмотреть два случая.

Случай первый: взрывное развитие искусственного интеллекта не создаст ситуацию типа «победитель получает все» (предположительно потому, что взлет окажется сравнительно медленным). В этом случае можно допустить, что если степень сотрудничества в преддверии перехода как-то повлияет на степень сотрудничества после него, то это влияние будет положительным и стимулирующим. Сложившееся сотрудничество сохранится и продолжится после перехода, кроме того, совместные усилия, предпринятые до перехода, позволят направить развитие в желательном направлении (и предположительно, открывающем возможности для еще более тесного сотрудничества) после него.

Случай второй: природа взрывного развития искусственного интеллекта поощряет формирование ситуации, когда «победитель получает все» (предположительно потому, что взлет окажется сравнительно быстрым). В этом случае в отсутствие широкого сотрудничества в преддверии перехода, скорее всего, сформируется синглтон — переход совершит всего один проект по сверхразуму, в какой-то момент обеспечив себе решающее стратегическое преимущество. Синглтон по определению представляет собой социальный порядок с высокой степенью сотрудничества[44]. Таким образом, отсутствие сотрудничества в предшествующий переходу период приведет к чрезвычайно активному сотрудничеству после него. По контрасту с этим более высокая степень сотрудничества в ходе подготовки к взрывному развитию искусственного интеллекта открывает возможности для широкого диапазона исходов. Сотрудничающие группы разработчиков могут синхронизировать работу над проектами так, чтобы гарантированно совершить переход вместе и не допустить получения решающего стратегического преимущества ни одному из них. Или вообще объединить усилия и действовать в рамках общего проекта, но отказаться при этом от формирования синглтона. Например, можно представить консорциум стран, которые запускают совместный научный проект по созданию машинного интеллекта, но не дают санкцию на его превращение в сверхмощный аналог ООН, ограничившись поддержанием имеющегося мирового порядка.

Таким образом, получается, что более активное сотрудничество до перехода может привести к меньшему сотрудничеству после него, особенно в случае быстрого взлета. Однако в той мере, в которой сотрудничающие

стороны способны контролировать исход, они могут прекратить сотрудничать или вовсе не начинать это делать лишь в том случае, если считают, что постпереходная фракционность не приведет ни к каким катастрофическим последствиям. То есть сценарии, в которых активное сотрудничество в преддверии перехода сменяется его отсутствием в постпереходный период, в основном относятся к ситуации, в которой это будет безопасно.

В общем случае представляется желательным более активное сотрудничество в постпереходный период. Оно снизит риск возникновения антиутопии, в которой в результате экономической конкуренции и быстрого роста населения возникают мальтузианские условия, или в которой эволюционный отбор приводит к эрозии человеческих ценностей и отсеву жизнелюбивых характеров, или в которой противоборствующие силы сталкиваются с другими последствиями своей неспособности к координации усилий — последствиями вроде войн и технологической гонки вооружений. Последнее — перспектива острой технологической конкуренции — может оказаться особенно опасным в случае перехода к промежуточной форме машинного интеллекта (имитационная модель мозга), поскольку это создаст ощущение гонки и снизит шансы решения проблемы контроля к моменту начала второго перехода к более развитой форме машинного интеллекта (искусственный интеллект).

Мы уже обсудили, каким образом в результате сотрудничества на этапе подготовки к взрывному развитию интеллекта может уменьшиться острота конфликтов, повыситься вероятность решения проблемы контроля и сложиться более оптимальное распределение ресурсов в постпереходную эпоху — как с этической, так и с практической точек зрения.

Совместная работа

Сотрудничество может принимать различные формы в зависимости от количества участников. В нижней части шкалы объединять свои усилия могут отдельные группы конкурирующих друг с другом разработчиков ИИ[45]. Корпорации могут сливаться или инвестировать друг в друга. На верхнем уровне возможно создание крупного международного проекта с участием нескольких стран. Уже известны прецеденты масштабного международного научно-технологического сотрудничества (Европейская организация по ядерным исследованиям; проект «Геном человека», Международная космическая станция), но с международным проектом создания безопасного сверхразума связаны трудности более высокого порядка из-за необходимости соблюдать

строгие меры предосторожности. Он может быть организован не в форме открытого научного сотрудничества, а в форме чрезвычайно жестко контролируемого предприятия. Возможно даже, что участвующих в нем ученых придется физически изолировать в течение всего срока реализации проекта, лишив их всех возможностей связываться с внешним миром за исключением единственного тщательно цензурируемого канала. В настоящее время требуемый уровень безопасности вряд ли достижим, но позднее, в результате развития технологий детекции лжи и систем наблюдения, ситуация может измениться. Тесное сотрудничество не обязательно означает, что в проекте принимают участие множество известных ученых. Скорее, это примет другую форму — крупные ученые будут иметь право голоса при определении целей проекта. В принципе, проект мог бы обеспечить максимально широкое сотрудничество, если в качестве его организатора выступит все человечество (представленное, например, Генеральной Ассамблеей Объединенных Наций) и при этом в качестве исполнителя привлекут единственного ученого[46].

Существует причина, по которой начать сотрудничество нужно как можно раньше: это дает возможность воспользоваться преимуществом «вуали неведения», которая пока скрывает от нас то, какой из проектов первым выйдет на уровень разработки сверхразума. С одной стороны, чем ближе мы подойдем к финишной черте, тем меньше неопределенности останется относительно шансов конкурирующих проектов; соответственно, тем сложнее будет объединить усилия, преодолев эгоистический интерес разработчиков лидирующих проектов и обеспечив возможность распределения благ между всеми жителями планеты. С другой стороны, будет сложно формально договориться о сотрудничестве в масштабах всей планеты до тех пор, пока перспективы появления сверхразума не станут более очевидны всем и пока не проявятся контуры пути, по которому можно двигаться в сторону его создания. Более того, учитывая, что сотрудничество способно привести к более быстрому прогрессу на этом пути, сейчас стремиться к нему может быть даже неконструктивно с точки зрения безопасности.

Следовательно, в наши дни идеальной формой сотрудничества могла бы быть такая, которая не подразумевает конкретных формальных договоренностей и не ускоряет движение в сторону создания машинного интеллекта. Этим критериям удовлетворяет, в частности, предложение сформулировать некий этический принцип, выражающий нашу приверженность идее, что сверхразум должен служить на благо всем.

Принцип общего блага
Сверхразум должен быть создан исключительно на благо всего человечества и отвечать общепринятым этическим идеалам[47].

Чем раньше мы сформулируем принцип, что весь безмерный потенциал сверхразума принадлежит всему человечеству, тем больше у нас будет времени, чтобы эта норма прочно утвердилась в умах людей.

Принцип общего блага не исключает коммерческих интересов отдельного человека и компаний, действующих в соответствующих областях. Например, соответствовать этому принципу можно было бы, согласившись, что все «свалившиеся с неба» доходы от создания сверхразума — доходы немыслимо высокого уровня (скажем, триллион долларов в год) — распределяются между акционерами и прочими заинтересованными лицами так, как это принято сегодня, а те, что превышают этот порог, — равномерно между всеми жителями планеты (или как-то иначе, но в любом случае в соответствии с неким универсальным этическим критерием). Принять решение о доходах, «свалившихся с неба», не стоит компании практически ничего, поскольку превысить эту астрономическую сумму в наши дни практически невозможно (такие маловероятные сценарии не принимаются во внимание в процессе принятия решений современными управляющими и инвесторами). При этом широкое распространение такого рода обязательств (при условии, что им можно доверять) дало бы человечеству гарантию, что *если* какая-то частная компания получит джекпот в результате взрывного развития интеллекта, все смогут участвовать в распределении свалившихся благ. Ту же идею можно применить не только к фирмам. Например, государства могут договориться о том, что если ВВП какого-то из них превысит некоторую довольно большую долю мирового ВВП (например, 90 процентов), превышение будет распределено равномерно между всеми жителями планеты[48].

Поначалу принцип общего блага (и его практические следствия вроде обязательства распределять сверхдоходы) мог бы применяться избирательно, путем добровольного согласия поступать в соответствии с ним, данного ответственными специалистами и организациями, которые работают в области машинного интеллекта. Позднее к нему присоединились бы другие, и со временем он превратился бы в закон или международное соглашение. Приведенная выше не очень строгая формулировка вполне подойдет для начала, но в конечном счете должна быть заменена набором конкретных требований, поддающихся проверке.

Глава пятнадцатая
Цейтнот

Мы заблудились в непроходимой чаще стратегической сложности, которую окутывает плотный туман неопределенности. Хотя многие элементы окружающего нас пейзажа разглядеть удалось, но отдельные детали и взаимосвязи остаются загадкой. Кроме того, могут быть другие факторы, о которых мы пока даже не задумываемся. Что нам остается в такой ситуации?

Крайний срок философии

Мои коллеги часто шутят, что вручение Филдсовской премии (высшей награды в области математики) может означать две вещи: или награжденный смог сделать что-то важное, или нет. Не очень приятно, но доля правды в этом есть.

Определим «открытие» как действие, в результате которого некая информация становится известна раньше, чем могла бы появиться. Тогда ценность открытия не равняется значимости заложенной в нем информации, а скорее приобретает значение, потому что информация попадает в наше распоряжение сейчас, а не позже. Чтобы первому найти решение проблемы, которое ускользнуло от многих других, ученому требуется приложить много сил и способностей; но если проблема в любом случае должна была в ближайшее время быть решена, мир получил не так много пользы от его работы. Конечно, *есть* случаи, когда даже чуть более раннее решение проблемы несет в себе огромную ценность, но чаще всего это возможно лишь в тех случаях, когда такое решение можно немедленно использовать: или применить на практике, или положить в основу другой теоретической работы. Причем в последнем случае, когда решение сразу используется в качестве строительного материала для дальнейших теоретических поисков, то сам

факт, что оно получено чуть раньше, будет значим лишь тогда, когда новая работа сама по себе и важная, и срочная[1].

То есть вопрос не в том, является ли результат, полученный математиком, награжденным Филдсовской премией, «важным» (или сам по себе, или с инструментальной точки зрения). Скорее, значение имеет то, насколько важно появление этого результата сейчас, а не позже. И ценность этого перемещения во времени следует сравнивать со значимостью иного результата, который мог бы получить этот математик мирового уровня, работая над другой проблемой. Как минимум в некоторых случаях Филдсовская премия становилась свидетельством того, что жизнь была потрачена на решение неправильно выбранных задач, например тех, привлекательность которых вызвана главным образом тем, что они представляли собой трудноразрешимые проблемы.

Того же упрека заслуживает и философия. Конечно, она действительно занимается изучением проблем, решение которых поможет снизить экзистенциальный риск, — с некоторыми философскими концепциями мы познакомились в этой книге. Но часть философских теорий не имеет отношения ни к экзистенциальным рискам, ни к какой-либо иной практической направленности. Конечно, некоторые проблемы, которыми занимается философия (как и математика), важны сами по себе, в том смысле, что размышлять над ними люди могут независимо от возможностей применить результаты своих умозаключений на практике. Фундаментальную природу реальности, например, стоит понимать в любом случае. Мир определенно был бы менее симпатичным местом, если бы никто не изучал метафизику, космологию или теорию струн. Однако мрачные перспективы взрывного развития интеллекта проливают новый свет на это старое как мир стремление к мудрости.

Сейчас складывается ощущение, что прогресса в области философии можно достигнуть, двигаясь окольными путями, а не за счет непосредственного философствования. Одной из многих задач, в решении которых сверхразум (и даже слегка улучшенный человеческий интеллект) мог бы опередить современных мыслителей, является задача ответа на фундаментальные научные и философские вопросы. Эта мысль предполагает следование стратегии отложенного удовольствия. Мы могли бы отложить ненадолго работу над вечными вопросами, делегировав эту задачу нашим, будем надеяться, более компетентным потомкам, чтобы сосредоточить свое внимание на более

неотложной проблеме: повышении шансов того, что у нас вообще будут компетентные потомки. Вот это и называют настоящей философией и настоящей математикой[2].

Что нужно делать?

Таким образом, нам следует сосредоточить силы на проблемах, не просто важных, но неотложных в том смысле, что их нужно разрешить раньше, чем произойдет взрывное развитие интеллекта. И воздерживаться от работы над задачами с отрицательной ценностью (то есть решение которых принесет вред). Отрицательную ценность могут иметь некоторые технические задачи в области искусственного интеллекта, поскольку их решение может подстегнуть развитие этого направления, опережающее создание методов контроля, которые все-таки призваны помочь человечеству пережить революцию машинного интеллекта и получить выгоду от нее.

Довольно нелегко определить проблемы, которые одновременно являются важными, срочными и имеющими абсолютно положительную ценность. Стратегическая неопределенность, присущая попыткам снизить экзистенциальные риски, приводит к тому, что даже действия, предпринятые с самыми добрыми намерениями, могут оказаться не только непродуктивными, но даже неконструктивными. Чтобы ограничить риск совершения действий, способных принести реальный вред или оказаться ошибочными с этической точки зрения, нам следует работать лишь над теми проблемами, которые, как представляется, обладают *строго положительной ценностью* (например, решение которых помогает получить положительный исход в широком спектре сценариев), и применять инструменты исключительно допустимые (например, приемлемые с точки зрения широкого спектра этических норм).

При расстановке приоритетов есть и другое соображение, которое следует учитывать. Мы хотим работать над проблемами, которые *эластичны* к нашим усилиям по их решению. Высокоэластичные проблемы — это такие, которые в результате приложения дополнительной единицы усилий могут быть решены гораздо быстрее и в гораздо большей степени. Сделать мир добрее — задача явно важная и срочная, более того, имеющая строго положительную ценность, однако из-за отсутствия прорывных идей — как именно этого можно добиться — она предстает совершенно неэластичной. Точно так же крайне желательно было бы обеспечить мир во всем мире,

но учитывая огромные усилия, уже направленные на решение этой задачи, и серьезные препятствия, не позволяющие сделать это быстро, кажется маловероятным, что вклад еще нескольких человек будет иметь какое-то принципиальное значение.

Чтобы снизить риски, связанные с революцией машинного интеллекта, мы предложили бы две цели, которые, по всей видимости, удовлетворяют всем высказанным пожеланиям: это стратегический анализ и создание возможностей. Мы можем быть уверенными в положительной ценности этих параметров, поскольку стратегическое понимание и большие возможности — это благо. Более того, они эластичны: небольшие дополнительные инвестиции способны привести к сравнительно большим изменениям. Срочность этой задачи также очевидна — обеспечив понимание и создав новые возможности сейчас, можно будет более эффективно прилагать усилия в будущем. Кроме этих двух крупных целей можно предложить еще несколько других, также достойных нашего внимания.

В поисках стратегии

В условиях растерянности и неопределенности стратегический анализ представляется задачей, имеющей особенно высокую ожидаемую полезность[3]. Понимание ситуации, в которой мы находимся, поможет нам действовать более эффективно. Особенно полезен стратегический анализ в том случае, если мы не уверены не только во второстепенных деталях, но и главных аспектах проблем. Часто неопределенность связана даже со *знаком* многих ключевых параметров: то есть неизвестно, какое направление изменений желательно, а какое — нет. Наше незнание может иметь непоправимые последствия. Однако область эта изучена мало, и сверкающие стратегические озарения могут ожидать нас совсем близко, стоит лишь сделать первый шаг.

Под «стратегическим анализом» мы понимаем здесь поиск *критически важных соображений*: идей и аргументов, обладающих потенциалом изменить наши взгляды не столько на детали реализации плана, сколько на общую топологию желательного[4]. Даже единственное упущенное из виду критически важное соображение способно обесценить наши самые напряженные усилия или вовсе обратить их во вред — как в случае солдата, воюющего за неправую сторону. Поиск критически важных соображений (в ходе которых приходится изучать как нормативные, так и дескриптивные вопросы)

часто предполагает пересечение границы между различными научными дисциплинами и другими областями знаний. Поскольку устоявшейся методологии проведения такого рода исследований пока не существует, они требуют напряженных размышлений буквально с чистого листа.

В поисках возможностей

Еще одним ценным видом деятельности, для которого характерно такое же, как у стратегического анализа, свойство пользы в случае реализации широкого спектра сценариев, является создание хорошо организованной базы поддержки идеи, что к будущему следует относиться серьезно. При наличии такой базы немедленно появятся ресурсы для проведения аналитических исследований. Если приоритеты изменятся, соответственно можно будет перераспределить и ресурсы. То есть база поддержки и станет той самой возможностью, использование которой будет определяться новыми идеями (по мере их появления).

Одним из самых ценных активов могла бы стать донорская сеть, состоящая из людей, приверженных идее рациональной филантропии, информированных об экзистенциальном риске и о средствах его уменьшения. Особенно важно, чтобы эти отцы-основатели были проницательными и альтруистичными, поскольку им представится возможность сформировать культуру направления прежде, чем в ней проявятся и укоренятся обычные корыстные мотивы. То есть фокус в ходе этого открывающего партию гамбита должен быть направлен на привлечение к работе над направлением правильных людей. Возможно, даже стоит пожертвовать какими-то быстрыми техническими успехами ради того, чтобы заполнить вакантные позиции теми, кто искренне заботится о безопасности и ориентируется на поиск истины (а также способен привлекать единомышленников).

Важным параметром является качество «социальной эпистемологии» области ИИ и лидирующих в ней проектов. Поиск критически важных соображений — ценное занятие, но только в том случае, если как-то влияет на конкретные действия. А это не обязательно так. Представьте проект, в рамках которого в создание прототипа ИИ вложены годы работы и миллионы долларов инвестиций и который после преодоления многочисленных технических сложностей начинает демонстрировать реальный прогресс. Есть шанс, что еще чуть-чуть — и он превратится в нечто полезное и прибыльное. Но тут возникает критически важное соображение о том, что

гораздо безопаснее было бы пойти по совершенно иному пути. Убьет ли проект сам себя, как опозоренный самурай, отказавшись от своей небезопасной архитектуры и зачеркнув весь достигнутый прогресс? Или отреагирует как потревоженный кальмар, выбросив облако контраргументов в надежде отбить нападение? Если, столкнувшись с такой дилеммой, команда проекта выберет путь самурая, она окажется гораздо более предпочтительным разработчиком[5]. Хотя довольно трудно строить процессы и институты, готовые из-за голословных обвинений совершать харакири. Еще одним измерением социальной эпистемологии является управление конфиденциальной информацией, особенно способность избежать ее утечки. (Информационное воздержание может оказаться особенно трудным в случае традиционных ученых, привыкших расклеивать результаты своих исследований на всех столбах и деревьях.)

Конкретные показатели

В дополнение к общим целям стратегического понимания и создания новых возможностей могут быть обозначены и более специфические цели, достижение которых станет экономически оправданным.

Одна из них — это прогресс в решении технических вопросов обеспечения безопасности машинного интеллекта. Преследуя эту цель, стоит внимательно относиться к управлению информационными рисками. Какие-то действия, полезные с точки зрения решения проблемы контроля, могут быть полезными и с точки зрения решения проблемы компетентности. Если какое-то задание приведет к перегоранию предохранителей ИИ, возможно, оно имеет отрицательное значение.

Еще одной специфической целью является пропагандирование последних достижений в этой области среди исследователей ИИ. Следует распространять информацию о любом прогрессе в решении задачи контроля. В некоторых видах вычислительных экспериментов, особенно когда предполагается возможность рекурсивного самосовершенствования, требуется использовать меры контроля возможностей, чтобы снизить риск случайного взлета. Хотя практическое применение инструментов обеспечения безопасности пока неактуально, оно станет таковым по мере нашего продвижения вперед. Не за горами то время, когда перспективы появления машинного интеллекта окажутся довольно близкими, и нужно будет призвать участников процесса выразить свою *приверженность безопасности*, согласие

с принципом общего блага и обещание усилить меры предосторожности. Одних благочестивых слов может быть недостаточно, и сами по себе они не сделают опасную технологию безопасной, но за словами могут постепенно последовать мысли, а потом и поступки.

Время от времени могут возникать другие возможности улучшения каких-то важных параметров, например снижения одного из экзистенциальных рисков, проведения когнитивного улучшения биологического мозга и углубления нашей коллективной мудрости или даже перевода мировой политики в какой-то более гармоничный регистр.

Все лучшее в человеческой природе — шаг вперед!

Перед лицом перспективы взрывного развития интеллекта мы похожи на детей, играющих с бомбой. Именно таков разрыв между мощностью нашей игрушки и незрелостью нашего поведения. Сверхразум — это вызов, ответить на который мы не готовы сегодня, и не будем готовы еще долгое время. Мы понятия не имеем, когда произойдет детонация, хотя если поднести устройство к уху, уже можно услышать тихое тиканье.

Ребенку с тикающей бомбой в руках правильнее всего осторожно положить ее на пол, быстро выбежать из комнаты и обратиться к ближайшему взрослому. Впрочем, в нашем случае ребенок не один, нас много, и у каждого в руках свой взрыватель. Шансы на то, что *все* обладают достаточным здравым смыслом, чтобы аккуратно поместить опасный предмет в безопасное место, ничтожны. Какой-нибудь маленький идиот обязательно нажмет красную кнопку просто из любопытства.

Убежать мы тоже не можем, поскольку взрывное развитие искусственного интеллекта обрушит сам небесный свод. Да и взрослых в поле зрения нет.

В такой ситуации проявлять восторженное любопытство было бы неуместным. Более подходящими чувствами были бы озабоченность и страх, а самое правильное поведение — горькая решимость стать настолько компетентными, насколько это возможно, как при подготовке к экзамену, который или позволит исполнить наши мечты, или разрушит их.

Никто не призывает к фанатизму. Возможно, что до взрыва остается еще много десятков лет. Более того, отчасти вызов, с которым мы столкнулись, состоит в необходимости оставаться людьми — практичными, здравомыслящими, добропорядочными и оптимистичными, — даже оказавшись в когтях этой совершенно неестественной и нечеловеческой проблемы. И чтобы

найти ее решение, нам придется использовать всю присущую людям изобретательность.

И все-таки давайте не терять из виду самое главное. Сквозь пелену сиюминутных мелочей уже проступают — пока смутные — контуры главной задачи нашей эпохи. В этой книге мы попытались рассмотреть некоторые ее аспекты и приблизились к пониманию того, что высший этический приоритет (как минимум с объективной и секулярной точки зрения) имеют такие вопросы, как снижение экзистенциального риска и выход на цивилизационную траекторию, которая ведет к лучшему исходу — на благо всего человечества.

Примечания

Введение

[1] Должен признать, не все примечания содержат ценную информацию.

[2] Вряд ли смогу сказать, *что именно* изложено корректно.

Глава 1. Прошлые достижения и сегодняшние возможности

[1] В настоящее время доход на уровне прожиточного минимума равен примерно 400 долларов [Chen, Ravallion 2010]. Следовательно, для 1 млн человек эта сумма будет равняться 400 000 000 долларам. Мировой ВВП составляет около 60 000 000 000 000 долларов и растет с темпом четыре процента в год (учитывается среднегодовой темп роста с 1950 года, см. данные: [Maddison 2010]). Цифры, приведенные мною в тексте, основаны на этих данных, хотя они представляют всего лишь оценку порядка величины. Если проанализировать сегодняшнюю численность людей на Земле, то выяснится, что в среднем она увеличивается на 1 млн человек за полторы недели; но подобный темп прироста населения лимитирует скорость экономического развития, поскольку доход на душу населения растет тоже. При переходе к животноводству и земледелию население планеты выросло к 5000 году до н. э. на 1 млн человек за 200 лет — огромное ускорение по сравнению с эпохой гоминидов, когда на это требовалось 1 млн лет, — поэтому после неолитической, или сельскохозяйственной, революции прогресс пошел значительно быстрее. Тем не менее, согласитесь, не может не впечатлять, что семь тысяч лет назад на экономическое развитие требовалось 200 лет, тогда как сегодня приросту на ту же величину хватает полутора часов для мировых экономик и полутора недель для населения планеты. См. также [Maddison 2005].

[2] Резкий рост и значительное ускорение подтверждают предположение о возможном приближении к точке сингулярности; в свое время это предвидели математики Джон фон Нейман и Станислав Улам:

Чаще всего мы вели беседы на такие темы, как ускорение технического прогресса и общие перемены, влияющие на образ жизни человека. Наблюдая стремительные изменения, мы понимали, что исторически вся наша гонка неизбежно приведет человечество к некой неустранимой точке; перейдя ее, люди уже не смогут продолжать ту деятельность, к которой мы все так привыкли [Ulam 1958].

[3] См.: [Hanson 2000].

[4] См.: [Vinge 1993; Kurzweil 2005].

[5] См.: [Sandberg 2010].

[6] См.: [Van Zanden 2003; Maddison 1999; Maddison 2001; De Long 1998].

[7] Два часто повторяемых в 1960-е гг. оптимистичных утверждения: «Через двадцать лет машины смогут выполнять всю работу, которую делают люди» [Simon 1965, p. 96]; «В течение жизни нашего поколения… задача создания искусственного интеллекта будет в целом решена» [Minsky 1967, p. 2]. Системное исследование прогнозов см.: [Armstrong, Sotala 2012].

[8] См., например: [Baum et al. 2011; Armstrong, Sotala 2012].

[9] По моему предположению, исследователи, работающие в области ИИ, сами не отдают себе отчета, что довольно плохо представляют, сколько времени требуется на его создание; причем данное обстоятельство может сказаться на оценке срока разработки: как в сторону завышения, так и занижения.

[10] См.: [Good 1965, p. 33].

[11] Одним из немногих, кто выражал беспокойство по поводу возможных результатов, был Норберт Винер, в 1960 году писавший в статье «Некоторые моральные и технические последствия автоматизации» [Wiener 1960]:

Если для достижения собственных целей мы используем некое механическое устройство, на деятельность которого после его запуска уже не в состоянии повлиять, поскольку его действия настолько быстры и небратимы, что информация о необходимости вмешательства у нас появляется, лишь когда они завершатся, то в таких случаях хорошо бы быть полностью уверенными, что заложенная нами в машину цель действительно соответствует нашим истинным желаниям, а не является красочной имитацией.

О тревоге, связанной с возможным появлением сверхразумных машин, рассказал в своем интервью Эд Фредкин [McCorduck 1979]). Ирвинг Гуд продолжал говорить о возможных рисках и, чтобы предотвратить грозящую планете опасность, в 1970 году даже призвал к созданию ассоциации [Good 1970]; в более поздней статье он предвосхитил некоторые идеи косвенной нормативности [Good 1982] (мы их обсудим в главе 13). На принципиальные проблемные вопросы указал в 1984 году Марвин Мински [Minsky 1984].

[12] См.: [Yudkowsky 2008a]. На необходимость дать оценку этическим аспектам потенциально опасных будущих технологий *прежде*, чем они станут реальностью, указывала Ребекка Роуч [Roache 2008].

[13] См.: [McCorduck 1979].

[14] См.: [Newell et al. 1959].

[15] Программы Saints, Analogy и Student соответственно, см.: [Slagle 1963; Evans 1964; Evans 1968; Bobrow 1968].

[16] См.: [Nilsson 1984].

[17] См.: [Weizenbaum 1966].

[18] См.: [Winograd 1972].

[19] См.: [Cope 1996; Weizenbaum 1976; Moravec 1980; Thrun et al. 2006; Buehler et al. 2009; Koza et al. 2003]. Первая лицензия на право пользоваться беспилотным автомобилем выдана управлением транспортных средств штата Невада в мае 2012 года.

[20] Система STANDUP [Ritchie et al. 2007].

[21] Эти слова Хьюберта Дрейфуса как характерный пример общего скептического отношения к предмету обсуждения приводит в своей статье «Пределы искусственного интеллекта» Джекоб Шварц [Schwartz 1987].

[22] В то время одним из самых неистовых и ярких противников ИИ считался Хьюберт Дрейфус, но и другие критики были не менее знамениты и заметны, например Джон Лукас, Роджер Пенроуз и Джон Сёрл. Дрейфус, опровергая работы ведущих исследователей, главным образом подвергал сомнению практическую пользу, которую сможет принести существовавшая на тот момент парадигма ИИ (причем, похоже, он не исключал появления более удачных концепций). Сёрл, будучи философом, прежде всего интересовался не инструментальными средствами для разработки ИИ, а тем, как решаются проблемы сознания, в частности, с точки зрения теории функциональных систем. Лукас и Пенроуз в принципе отрицали, что в рамках парадигмы классического компьютера можно разработать программное обеспечение, думающее и дышащее лучше живого математика; однако оба допускали и автоматизацию отдельных функций, и создание таких мощных инструментальных средств, которые в конечном счете приведут к появлению ИИ. И хотя Цицерон в трактате «О прорицании» (De divinatione)* заметил, что «нет такого абсурда, который нельзя было бы найти в книгах философов» [Cicero. On Divination, 119], как ни странно, мне трудно вспомнить хотя бы *одного* серьезного ученого и просто мыслящего человека, отрицавшего возможность создания искусственного интеллекта — в том значении этого термина, который используется в настоящей книге.

[23] Однако во многих приложениях процесс обучения нейронных сетей несколько отличается от модели линейной регрессии — статистического метода, разработанного в начале XIX века Адриеном-Мари Лежандром и Карлом Фридрихом Гауссом.

[24] Основной алгоритм был описан в 1969 году Артуром Брайсоном и Юй-Чи Хо как многошаговый метод динамической оптимизации [Bryson, Ho 1969]. Применить его к нейронным сетям предложил в 1974 году Пол Вербос [Werbos 1994], но признание у научного сообщества этот метод получил лишь в 1986 году после работы Дэвида Румельхарта, Джеффри Хинтона и Рональда Уильямса [Rumelhart et al. 1986].

[25] Ранее было показано, что функциональность сетей без скрытых слоев серьезно ограничена [Minsky, Papert 1969].

[26] См., например: [MacKay 2003].

[27] См.: [Murphy 2012].

[28] См.: [Pearl 2009].

[29] Мы сознательно опускаем различные технические подробности, чтобы не перегружать повествование. К некоторым из них будет возможность вернуться в главе 12.

[30] Программа p генерирует полное описание строки x, если p, запущенная на (некоторой) универсальной машине Тьюринга U, выдает x; это можно записать как $U(p) = x$. (Здесь строка x представляет любой возможный мир.) Тогда колмогоровская сложность x равна $K(x) = \min_p \{l(p) : U(p) = x\}$,

* *Марк Туллий Цицерон*. О дивинации. Философский трактат в двух книгах // *Марк Туллий Цицерон*. Философские трактаты. М. : Наука, 1985.

где *l(p)* — это длина *p* в битах. Соломоновская вероятность x определяется как $M(x) = \sum_{U(p)=x} 2^{-l(p)}$, где сумма задана над всеми («минимальными», то есть не обязательно останавливающимися) программами *p*, для которых *U* выдает строку, начинающуюся с *x*; см.: [Hutter 2005].

[31] Байесово обусловливание с учетом свидетельства *E* дает (вероятность утверждения [например, *E*] есть сумма вероятностей возможных миров, в которых это утверждение истинно.)

$$P_{\text{posterior}}(w) = P_{\text{prior}}(w \mid E) = \frac{P_{\text{prior}}(E \mid w) P_{\text{prior}}(w)}{P_{\text{prior}}(w)}$$

[32] Или случайным образом выбирает одно из возможных действий с максимальной ожидаемой полезностью, если их несколько.

[33] Более сжато ожидаемая полезность действия может быть записана как $EU(a) = \sum_{w \in W} U(w)P(w|a)$, где сумма берется по всем возможным мирам.

[34] См., например: [Howson, Urbach 1993; Bernardo, Smith 1994; Russell, Norvig 2010].

[35] См.: [Wainwright, Jordan 2008]. У байесовских сетей бесчисленное количество областей применения; см., например: [Pourret et al. 2008].

[36] Возможно, некоторые читатели, сочтя это направление не слишком серьезным, зададут вопрос: зачем уделять столь пристальное внимание компьютерным играм? Дело в том, что игровые интеллектуальные системы, пожалуй, дают самое наглядное представление о сравнительных возможностях человека и машины.

[37] См.: [Samuel 1959; Schaeffer 1997, ch. 6].

[38] См.: [Schaeffer et al. 2007].

[39] См.: [Berliner 1980 a; Berliner 1980 b].

[40] См.: [Tesauro 1995].

[41] В частности, такие программы по игре в нарды, как GNU [Silver 2006] и Snowie [Gammoned.net, 2012].

[42] Процессом создания космического флота и битвами руководил сам Дуглас Ленат, написавший по этому поводу: «Итак, победа стала заслугой и Ленната, и Eurisco — в пропорции 60 : 40. Основной момент тем не менее состоит в том, что в одиночку ни я, ни программа никогда не справились бы» [Lenat 1983, p. 80].

[43] См.: [Lenat 1982; Lenat 1983].

[44] См.: [Cirasella, Kopec 2006].

[45] См.: [Kasparov 1996, p. 55].

[46] См.: [Newborn 2011].

[47] См.: [Keim et al. 1999].

[48] См.: [Armstrong 2012].

[49] См.: [Sheppard 2002].

[50] См.: [Wikipedia, 2012 a].

[51] См.: [Markoff 2011].

[52] См.: [Rubin, Watson 2011].

[53] См.: [Elyasaf et al. 2011].

[54] См.: [KGS, 2012].

[55] См.: [Newell et al. 1958, p. 320].

[56] См.: [Vardi 2012].

[57] Ирвинг Гуд предполагал в 1976 году:

Появление программного обеспечения, не уступающего по своему потенциалу гроссмейстерскому уровню, будет означать, что мы уже стоим на пороге <создания искусственного сверхразума. — *Н. Б.*> [Good 1976].

Даглас Хофштадтер писал в 1979 году в книге «Гёдель, Эшер, Бах», за которую в 1980-м он получит Пулитцеровскую премию:

Вопрос: Будут ли такие шахматные программы, которые смогут выиграть у кого угодно?

Возможный ответ: Нет. Могут быть созданы программы, которые смогут обыгрывать кого угодно, но они не будут исключительно шахматными программами. Они будут программами общего разума и, так же как люди, они будут обладать характером. «Хотите сыграть партию в шахматы?» — «Нет, шахматы мне уже надоели. Лучше давайте поговорим о поэзии…»* [Hofstadter 1999, p. 678].

[58] Минимаксный алгоритм поиска с альфа-бета отсечениями использовался совместно со специфической для шахмат функцией эвристической оценки позиций — это дало в сочетании с удачной библиотекой дебютов и эндшпилей, а также другими хитростями, очень сильную шахматную программу.

[59] Впрочем, учитывая достижения в изучении оценочной эвристики в ходе моделирования, многие базовые алгоритмы могли бы хорошо проявить себя в большом количестве других игр.

[60] См.: [Nilsson 2009, p. 318]. Конечно, Кнут несколько преувеличил успехи машинного разума. Все-таки есть еще интеллектуальные задачи, в которых ИИ не преуспел; например, остаются «непродуманными» такие аспекты, как открытие новых направлений в чистой математике, придумывание свежих философских концепций, создание циклов детективных романов, организация военного переворота, разработка очень нужного и инновационного товара широкого потребления.

[61] См.: [Shapiro 1992].

[62] Можно только предполагать, почему машине трудно достичь человеческого уровня в восприятии окружающей действительности, регуляции двигательных функций, здравом смысле и понимании языка. Одна из причин заключается в том, что в нашем мозгу имеется специальный механизм, управляющий этими свойствами, — достигшие в процессе эволюции совершенства нейронные структуры. Логическое мышление и навыки вроде игры в шахматы, в отличие от перечисленных выше способностей, не столь естественны, и потому при решении этих задач мы вынуждены полагаться на ограниченные когнитивные ресурсы общего назначения. Возможно, для вычислений и явно выраженных логических рассуждений наш мозг запускает что-то похожее на «виртуальную машину» — медленный и громоздкий психический симулятор универсального компьютера. Если наше предположение верно, то тогда получается

---

* *Даглас Хофштадтер.* Гёдель, Эшер, Бах. Эта бесконечная гирлянда / Пер. с англ. М. А. Эскиной. Самара: Издательский дом «Бахрах-М», 2001. С. 635.

забавная вещь: не КИИ моделирует человеческое мышление, а как раз наоборот — логически мыслящий человек симулирует программу ИИ.

[63] Надо заметить, что по мнению меньшинства, представленного приблизительно 20% взрослого населения США, Солнце «продолжает» вращаться вокруг Земли, — подобная картина наблюдается и в других развитых странах [Crabtree 1999; Dean 2005].

[64] См.: [World Robotics 2011].

[65] По оценке, взятой из работы: [Guizzo 2010].

[66] См.: [Holley 2009].

[67] Кроме того, есть гибридные подходы, основанные как на статистике, так и на правилах, но в наше время они не представляют никакого интереса.

[68] См.: [Cross, Walker 1994; Hedberg 2002].

[69] По сообщенным мне в частном порядке статистическим данным TABB Group — компании, специализирующейся на анализе рынка капиталов; ее офисы находятся в Нью-Йорке и Лодоне.

[70] См.: [CFTC/SEC Report on May 6, 2010]; другую точку зрения на события 6 мая 2010 года см.: [CME Group, 2010]*.

[71] Мне не хотелось бы, чтобы это воспринималось как аргумент против алгоритмического высокочастотного трейдинга, который вполне способен играть полезную роль, повышая ликвидность и эффективность рынка.

[72] Менее масштабное потрясение случилось на фондовом рынке 1 августа 2012 года, отчасти причиной стало то обстоятельство, что автоматический прерыватель не был запрограммирован приостанавливать торги в случае резких изменений в *количестве* обращаемых акций, см.: [Popper 2012]. Это затрагивает еще одну нашу тему: трудно предусмотреть все возможные варианты, когда стандартная ситуация, которая держится на хорошо продуманных принципах, вдруг выходит из-под контроля.

[73] См.: [Nilsson 2009, p. 319].

[74] См.: [Minsky 2006; McCarthy 2007; Beal, Winston 2009].

[75] По данным Питера Норвига (из личного общения). В принципе, любые курсы по информационным технологиям и машинному обучению очень популярны. Может быть, это объясняется неожиданно возросшим массовым интересом к аналитике больших данных (big data) — интересом, инициированным в свое время Google и весьма подогреваемым огромными призовыми суммами Netflix.

[76] См.: [Armstrong, Sotala 2012].

[77] См.: [Müller, Bostrom <В печати>].

[78] См.: [Baum et al. 2011; Sandberg, Bostrom 2011].

[79] См.: [Nilsson 2009].

---

* *CFTC* (Commodity Futures Trading Commission) — Комиссия по срочной биржевой торговле; *SEC* (Securities and Exchange Commission) — Комиссия по ценным бумагам и биржам; *CME Group* — Группа Чикагской товарной биржи, крупнейший североамериканский рынок ценных бумаг, созданный в результате объединения ведущих нью-йоркских и чикагских бирж.

[80] Безусловно, и в этом случае сохранялось условие, что научная деятельность будет продолжаться «без серьезных сбоев», а в мире не случится никаких цивилизационных катастроф. В интервью Нильсон использовал следующее определение ИИЧУ: «ИИ, способный выполнять приблизительно 80% работы не хуже человека или даже лучше» [Kruel 2012].

[81] В таблице показаны результаты четырех отдельных опросов, в последней строке даны средние показатели. Первые два опроса проводились среди участников нескольких научных конференций. PT-AI — конференция «Философия и теория ИИ» (Салоники, 2011); опрос состоялся в ноябре 2012 года; всего участников — 88 человек, количество респондентов — 43 человека. AGI — конференции «Универсальный искусственный интеллект» и «Универсальный искусственный интеллект — степень воздействия и угрозы» (Оксфорд, декабрь 2012); всего участников — 111 человек, количество респондентов — 72 человека. EETN — съезд Греческой ассоциации искусственного интеллекта (апрель, 2013); всего участников — 250 человек, количество респондентов — 26 человек. TOP100 — опрос ведущих специалистов по искусственному интеллекту в соответствии с индексом цитирования (май 2013); всего в списке — 100 человек, количество респондентов — 29 человек.

[82] См.: [Kruel 2011] — в работе собраны интервью с 28 специалистами по ИИ и в смежных областях.

[83] На диаграмме показаны перенормированные медианные оценки. Средние значения несколько отличаются. Например, средние значения для варианта «чрезвычайно негативное» были равны 7,6% (в Топ-100) и 17,2% (в объединенной оценке по всем опросам).

[84] В литературе встречается огромное количество подтверждений ненадежности прогнозов экспертов во многих областях, поэтому есть все основания полагать, что подобное положение истинно и для сферы изучения искусственного интеллекта. В частности, делающие прогнозы люди, как правило, слишком уверенные в своей правоте, считают себя более точными предсказателями, чем это есть на самом деле, и поэтому присваивают слишком низкую вероятность возможности, что их любимая гипотеза может оказаться ложной [Tetlock 2005]. (О других документально зафиксированных заблуждениях см., например: [Gilovich et al. 2002].) Однако неопределенность — неотъемлемая черта человеческой жизни, и многие наши действия неизбежно основаны на вероятностных прогнозах, то есть ожиданиях того, какие из возможных событий произойдут скорее всего. Отказ от более четко сформулированных вероятностных прогнозов не устранит эпистемологическую проблему, а лишь задвинет ее в тень [Bostrom 2007]. Вместо этого нашей реакцией на чрезмерную самонадеянность должны стать как расширение доверительных интервалов, или интервалов правдоподобия, так и борьба с собственными предубеждениями путем рассмотрения проблемы с различных точек зрения и тренировки интеллектуальной честности. В долгосрочной перспективе можно также работать над созданием методик, подходов к обучению и институтов, которые помогут нам достичь лучших проверочных образцов. См. также: [Armstrong, Sotala 2012].

Глава 2. Путь к сверхразуму

[1] Во-первых, этому определению сверхразума наиболее близка формулировка, опубликованная в работах: [Bostrom 2003 c; Bostrom 2006 a]; во-вторых, оно вполне отвечает формализованному

условию Шейна Легга: «Интеллект оценивается способностью агента добиваться своей цели в широком диапазоне условий» [Legg 2008]; в-третьих, оно очень напоминает описание, сделанное Ирвингом Гудом, которое приведено нами в главе 1 настоящего издания: «Давайте определим сверхразумную машину как машину, которая в значительной степени превосходит интеллектуальные возможности любого умнейшего человека» [Good 1965, p. 33].

[2] По той же причине не буду выстраивать никаких предположений, сможет ли сверхразумная машина обрести «истинную интенциональность», то есть иметь самосознание и действовать преднамеренно (при всем уважении к Джону Сёрлу, она, похоже, и на это способна), поскольку данный вопрос не имеет отношения к предмету нашей книги. Также не собираюсь занимать ничью сторону — ни адептов интернализма, ни последователей экстернализма — в яростно ведущихся среди философов дискуссиях на такие темы, как содержание сознания и расширение сознания, см.: [Clark, Chalmers 1998].

[3] См.: [Turing 1950, p. 456].

[4] См.: [Turing 1950, p. 456].

[5] См.: [Chalmers 2010; Moravec 1976; Moravec 1988; Moravec 1998; Moravec 1999].

[6] См.: [Moravec 1976]; аналогичную аргументацию приводит и Чалмерс, см.: [Chalmers 2010].

[7] Более подробно эта тема раскрывается в статье: [Shulman, Bostrom 2012].

[8] В своей диссертации Шейн Легг предлагает этот подход в качестве аргумента, что люди на воспроизведение эволюционного пути потратят гораздо меньше времени и меньше вычислительных ресурсов (при этом сам автор отмечает, что ресурсы, потребовавшиеся в ходе биологической эволюции, нам недоступны), см.: [Legg 2008]. Эрик Баум утверждает, что часть работы, связанной с созданием ИИ, проделана намного раньше без вмешательства человека, например: само строение генома уже содержит важную информацию об эволюционных алгоритмах, см.: [Baum 2004].

[9] См.: [Whitman et al. 1998; Sabrosky 1952].

[10] См.: [Schultz 2000].

[11] См.: [Menzel, Giurfa 2001; Truman et al. 1993].

[12] См.: [Sandberg, Bostrom 2008].

[13] Обсуждение этой точки зрения, а также анализ функций, определяющих приспособленность организма лишь на основании критерия его умственных способностей, см. в диссертации Легга [Legg 2008].

[14] Более системное и подробное описание способов, с помощью которых специалисты смогут превзойти имеющиеся на сегодня результаты эволюционного отбора, см. в статье «Мудрость природы» [Bostrom, Sandberg 2009 b].

[15] Обсуждая целевую функцию, мы говорим лишь о вычислительных ресурсах, необходимых для моделирования нервной системы живых существ, и не учитываем затраты на моделирование их тел или их виртуальной окружающей среды. Вполне возможно, что расчет целевой функции для тестирования каждого организма потребует гораздо меньше операций, чем нужно для симулирования всех нейронных вычислений, аналогичных нейронным процессам, происходящим в мозгу существа за срок его жизни. Сегодня программы ИИ часто разрабатывают для

действий в совершенно абстрактной среде (программы для доказательства теорем — в символических математических мирах; программы-агенты — в турнирных мирах простых игр).

Скептики могут настаивать, что для эволюции общего интеллекта абстрактной среды будет недостаточно, а виртуальное пространство должно детально напоминать тот мир, в котором развивались наши предки. Создание реалистичного виртуального мира потребовало бы гораздо больших инвестиций в вычислительные мощности, чем симуляция придуманного игрового мира или области для абстрактных задач (в то время как реальный мир достался эволюции «на дармовщину»). В предельном случае требования к полной точности на микроуровне приводят к тому, что потребность в вычислительных ресурсах вырастает до несуразно большой величины. Однако такой экстремальный пессимизм практически ничем не подкреплен: маловероятно, что наилучшие условия — с точки зрения эволюционирования интеллекта — должны максимально точно имитировать природу. Напротив, вполне возможно, что для симуляции естественного отбора искусственного интеллекта гораздо эффективнее будет использовать специальную среду, совершенно не похожую на ту, которая окружала наших предков, такую, которая специально разработана для стимулирования приспособления на основе нужных нам критериев (например, способности к абстрактному мышлению и общим навыкам решения задач, а не наличия максимально быстрой инстинктивной реакции или высокооптимизированной зрительной системы).

[16] См.: [Wikipedia, 2012 b].

[17] Об эффекте наблюдения при селективном отборе см.: [Bostrom 2002 a] — общее описание; [Shulman, Bostrom 2012] — с точки зрения обсуждаемой здесь темы; [Bostrom 2008 b] — короткое определение на доступном для неспециалиста языке.

[18] См.: [Sutton, Barto 1998; Schultz et al. 1997].

[19] В научный обиход термин введен Элиезером Юдковским, см.: [Yudkowsky 2007].

[20] Сценарий описан как Ирвингом Гудом [Good 1965], так и Элиезером Юдковским [Yudkowsky 2007]. Однако почему бы не представить альтернативный вариант, в котором итеративная последовательность пойдет по пути не развития интеллекта, а упрощения структуры? То есть время от времени зародыш ИИ будет переписывать самого себя таким образом, что работа его последующих версий значительно облегчится.

[21] См.: [Helmstaedter et al. 2011].

[22] См.: [Andres et al. 2012].

[23] Чтобы стали возможны инструментально полезные формы когнитивного функционирования и коммуникаций, технологии должны соответствовать необходимым требованиям, но все-таки их уровень намного примитивнее уровня технологий, обеспечивающих нейро- и мышечно-компьютерные интерфейсы, получающие сигналы от мышц и органов чувств обычного человеческого тела.

[24] См.: [Sandberg 2013].

[25] См.: [Sandberg, Bostrom 2008, p. 79–81] — раздел «Требования к компьютеру».

[26] Еще более простая модель может иметь микродинамику, подобную биологической, и активность, напоминающую активность мозга человека на ранних стадиях развития, например

похожую на фазу медленного сна и реакции на раздражители. Подобная модель могла бы быть полезна при проведении нейробиологических исследований (хотя есть риск возникновения серьезных этических проблем), но она не может считаться полной эмуляцией мозга человека, поскольку недостаточно точна для выполнения большинства умственных задач, с которыми справляется биологический мозг. Существует эвристическое правило: чтобы компьютерная модель мозга могла считаться его полноценной эмуляцией, она должна уметь вербально выражать связанные мысли или иметь возможность научиться этому.

[27] См.: [Sandberg, Bostrom 2008].

[28] Описание первого коннектома см. в работах: [Albertson, Thomson 1976; White et al. 1986]. Объединенная матрица (исправленная в некоторых местах) есть на сайте WormAtlas (http://www.wormatlas.org/).

[29] Обзор недавних попыток конструирования модели *C. elegans* и их результатов можно найти в блоге Джефа Кауфмана [Kaufman 2011]. Кауфман приводит комментарий Дэвида Дальримпля, честолюбивого докторанта, занимающегося исследованиями в этой области: «Не сочтите за слишком большую самонадеянность, но я скажу, что с помощью оптогенетики и высокопроизводительных автоматизированных систем мы вот-вот получим способность считывать и записывать любую информацию в нервную систему живого *C. elegans*… Думаю справиться с *C. elegans* за два-три года. Сильно удивлюсь, если эта проблема не будет решена до 2020 года» [Dalrymple 2011]. Модели мозга, задуманные как биологически точные, но закодированные вручную (а не сгенерированные автоматически) уже способны выполнять некоторые базовые функции, см., например: [Eliasmith et al. 2012].

[30] Подробнее см. в техническом отчете: [Sandberg, Bostrom 2008].

[31] У *Caenorhabditis elegans* есть особенности, делающие эту нематоду удобным организмом-моделью для изучения, например: у всех экземпляров прозрачное тело, а связи между нейронами всегда одинаковы.

[32] Если целью усилий становится не осуществление эмуляции мозга, а создание нейроморфного ИИ, тогда не обязательно пытаться выстраивать имитационную модель мозга именно *человека*. О важных особенностях работы коры головного мозга наверняка можно узнать в процессе изучения мозга животных. Часто с ним работать удобнее, поскольку он не так сложен по своим структурам и требует меньше ресурсов для сканирования и моделирования. Кроме того, требования, предъявляемые регулирующими инстанциями к таким исследованиям, наверняка будут слабее. Вполне возможно, что первый машинный интеллект человеческого уровня создадут, когда будет завершена полная эмуляция мозга некоего подходящего животного и будет найден способ развития этой компьютерной модели до уровня разума человека. И тогда, воспарив над обыденностью, какая-нибудь лабораторная мышь или макака получит наконец свой шанс отыграться на человечестве.

[33] См.: [Uauy, Dangour 2006; Georgieff 2007; Stewart et al. 2008; Eppig et al. 2010; Cotman, Berchtold 2002].

[34] По данным Всемирной организации здоровья, в 2007 году недостаточное потребление йода наблюдается примерно у 2 млрд человек [Lancet 2008]. *Тяжелая степень* йододефицита тормозит

развитие нервной системы и приводит к развитию эндемического кретинизма, вследствие которого коэффициент умственного развития снижается в среднем на 12,5 пункта, см.: [Qian et al. 2005]. Предотвратить эту ситуацию можно легко и недорого: за счет йодированной поваренной соли, см.: [Horton et al. 2008].

[35] См.: [Bostrom, Sandberg 2009 a].

[36] См.: [Bostrom, Sandberg 2009 b]. Средний *предполагаемый* эффект фармакологических и диетических методов, повышающих умственную деятельность, составляет 10–20% — это подтверждается контрольными проверками, определяющими уровень кратковременной памяти, внимания и т. д. Правда, остаются сомнения, что эти показатели истинны, надежны в течение длительного времени и отражают улучшение состояния в условиях реальной жизни, см.: [Repantis et al. 2010]. Например, в некоторых случаях наблюдается компенсаторное снижение активности, которое выявляется отнюдь не на контрольных проверках, см.: [Sandberg, Bostrom 2006].

[37] Если бы существовал простой способ улучшать когнитивные способности, следовало бы ожидать, что природа в процессе эволюции не преминула бы этим воспользоваться. Соответственно, самой многообещающей стала бы разработка таких ноотропных препаратов, которые были бы способны стимулировать развитие интеллекта за счет увеличения размера головы при рождении или путем активации метаболизма глюкозы в головном мозгу, но тогда это привело бы — с наследственной точки зрения — к резкому снижению приспособленности организма к среде обитания. Подробное обсуждение идеи (обратите внимание на весьма существенные квалификационные оговорки) см.: [Bostrom 2009 b].

[38] Сперматозоид сложнее подвергнуть скринингу, поскольку, в отличие от эмбрионов, он состоит из одной клетки, которая должна быть разрушена в процессе секвенирования. Ооциты, женские половые клетки, тоже состоят из одной клетки, однако ее первое и второе деления асимметричны и приводят к появлению одной дочерней клетки (так называемое полярное тельце) с небольшим объемом цитоплазмы. Поскольку цитоплазма содержит тот же геном, что и основная клетка, и является избыточной (в конечном счете вырождается), ее можно извлечь и использовать для скрининга, см.: [Gianaroli 2000].

[39] Генетические методы сначала вызывали некоторые разногласия с этической стороны вопроса, но сейчас, похоже, складывается тенденция к их все более широкому признанию. В разных культурах отношение к генной инженерии в применении к человеку и селекции эмбрионов очень неровное. Это означает, что развитие генетических технологий продолжится независимо от осторожной позиции отдельных стран, хотя, безусловно, и моральные, и религиозные, и политические соображения будут оказывать значительное давление на скорость процесса их внедрения.

[40] См.: [Davies et al. 2011; Benyamin et al. 2013; Plomin et al. 2013]; см. также: [Mardis 2011; Hsu 2012].

[41] В широком смысле показатель наследуемости коэффициента умственного развития у взрослых представителей среднего класса населения развитых стран обычно оценивается в диапазоне от 0,3 до 0,8 [Bouchard 2004, p. 148]. В узком смысле показатель наследуемости, который обычно измеряется как доля дисперсии, приходящаяся на аддитивные генетические факторы,

несколько ниже (диапазон 0,3–0,5), но все еще достаточно высок [Devlin et al. 1997; Davies et al. 2011; Visscher et al. 2008]. Эти оценки могут колебаться в зависимости от популяции и окружающей среды. Например, более низкие коэффициенты наследуемости выявляются среди детского населения и представителей социально неблагополучных слоев [Benyamin et al. 2013; Turkheimer et al. 2003]. Многочисленные примеры влияния окружающей среды на изменчивость когнитивных способностей приведены в работе «Интеллект. Новые данные и теоретические разработки» [Nisbett et al. 2012].

[42] Следующие параграфы и таблицы в значительной степени основаны на работе, написанной в соавторстве с Карлом Шульманом [Shulman, Bostrom 2014].

[43] См.: [Shulman, Bostrom 2014]. Таблица основана на игрушечной модели, в которой предполагается нормальное, или Гауссово, распределение ожидаемых коэффициентов интеллекта среди эмбрионов со стандартным отклонением в 7,5 пункта. Степень когнитивного улучшения — оно получается при разном количестве взятых для отбора эмбрионов — зависит от того, насколько сильно отличаются они друг от друга вариантами аддитивных генов, влияние которых нам известно. У братьев и сестер общие варианты аддитивных генов отвечают за половину или меньше дисперсии уровня интеллекта взрослых, см.: [Davies et al. 2011]. Таким образом, если стандартное отклонение для наблюдаемой популяции из развитых стран равно 15 пунктам, то стандартное отклонение генетического влияния для совокупности эмбрионов будет меньше или равно 7,5 пункта.

[44] Неполная информация об аддитивных эффектах генов на когнитивные способности снижает эффективность селекции, однако наличие даже небольшого количества данных дает сравнительно много, поскольку выигрыш от селекции не связан линейно с дисперсией, которую мы можем предсказать. Скорее, эффективность селекции зависит от стандартного отклонения среднего ожидаемого значения коэффициента умственного развития, которое равно *квадратному корню* из дисперсии. Например, если на этот фактор приходится 12,5% дисперсии, то наполовину уменьшаются величины эффекта, приведенные в табл. 1, где предполагается, что он отвечает за 50%. Для сравнения, в одном недавнем исследовании удалось выявить его влияние на 2,5% дисперсии, см.: [Rietveld et al. 2013].

[45] Для сравнения: сегодня стандартной практикой является создание не больше десяти эмбрионов.

[46] Из стволовых клеток взрослых людей и эмбрионов можно создать сперматозоиды и ооциты, из которых после оплодотворения получатся эмбрионы, см.: [Nagy et al. 2008; Nagy, Chang 2007]. Предшественники яйцеклеток также могут формировать партеногенетические бластоцисты, нефертильные и нежизнеспособные эмбрионы, из которых тоже можно брать стволовые клетки для этого процесса, см.: [Mai et al. 2007].

[47] Цит. по [Cyranoski 2013]. По опубликованным в 2008 году прогнозам Хинкстонской группы (Hinxton Group), международного консорциума ученых по этическим вопросам применения стволовых клеток, гаметы из стволовых клеток человека удастся получить в течение десяти лет [Hinxton Group, 2008] — достигнутый к настоящему времени прогресс вполне согласуется с таким прогнозом.

[48] См.: [Sparrow 2013; Miller 2012; Uncertain Future, 2012].

[49] См.: [Sparrow 2013].

[50] Светская общественность, как всегда озабоченная земными проблемами, будет говорить о многих нежелательных последствиях: углублении социального неравенства; несоблюдении медицинской безопасности самой процедуры; угрозе ужесточения бешеной гонки за успехом; правах и обязанностях родителей перед их потенциальными отпрысками; призраке евгенических экспериментов XX века. Она обязательно напомнит о человеческом достоинстве и границах допустимого вмешательства государства в репродуктивный выбор граждан. Подробное обсуждение морально-этических принципов, связанных с проблемой улучшения человеческого разума, см.: [Bostrom, Ord 2006; Bostrom, Roache 2011; Sandberg, Savulescu 2011]. У представителей мировых религий возникнут свои опасения, в частности: духовные вопросы по статусу эмбриона с точки зрения церковного права; пределы вмешательства человека в божественный процесс творения.

[51] Чтобы предотвратить негативные последствия инбридинга, для итерационного отбора эмбрионов может потребоваться увеличить количество доноров или приложить дополнительные усилия при отсеве, снизив количество вредоносных рецессивных аллелей. Оба пути приведут к меньшему генетическому сходству потомков с родителями (и большему сходству друг с другом).

[52] См.: [Shulman, Bostrom 2014].

[53] См.: [Bostrom 2008 b].

[54] Неизвестно, насколько будут велики эпигенетические препятствия [Chason et al. 2011; Iliadou et al. 2011].

[55] Хотя когнитивные способности во многом передаются по наследству, может быть мало или вовсе не быть *общих* аллелей, каждый из которых имеет сильное влияние на интеллект, см.: [Davis et al. 2010; Davies et al. 2011; Rietveld et al. 2013]. По мере совершенствования методов секвенирования будет постепенно уточняться карта низкочастотных аллелей и их связи с когнитивными и поведенческими особенностями. Имеются определенные теоретические свидетельства, что некоторые аллели, вызывающие генетические нарушения у гомозигот, могут обеспечить заметный прирост когнитивных способностей у гетерозигот, а это значит, что у гетерозиготных пациентов, несущих аллель болезни Гоше, Тея–Сакса и Ниманна–Пика, коэффициент интеллекта должен быть примерно на 5 пунктов выше, чем у контрольной группы [Cochran et al. 2006]. Время покажет, так ли это.

[56] В одной статье приведена оценка — 175 мутаций на геном в каждом поколении [Nachman, Crowell 2000]; в другой говорится, что среднее количество новых мутаций у новорожденного составляет 50–100 случаев в зависимости от использования различных методов [Lynch 2010]; в третьей статье речь идет о 77 новых мутациях в поколении [Kong et al. 2012]. Большинство мутаций не влияют на функционирование или влияют в пренебрежимо малой степени; но накопленный эффект многих малозаметных мутаций способен привести к значительному снижению приспособленности; см. также: [Crow 2000].

[57] См.: [Crow 2000; Lynch 2010].

[58] Эта мысль нуждается в критически важном разъяснении. Возможно, для избежания проблем модальный геном придется корректировать. Например, различные части генома могли приспособиться к взаимодействию друг с другом на определенном уровне эффективности. Повышение их эффективности может привести к перегрузке некоторых метаболических путей.

[59] Обобщенные портреты составлены южноафриканским фотографом Майком Майком из сделанных им в разных странах мира фотографий незнакомых людей в рамках проекта The Face of Tomorrow, см.: [Mike 2013].

[60] Конечно, общество начнет улавливать некоторые сигналы намного раньше, но это будет легко объясняться действием эффекта ожидания.

[61] См.: [Louis Harris & Associates, 1969; Mason 2003].

[62] См.: Kalfoglou et al. 2004].

[63] Понятно, что данных не очень много, но результаты некоторых лонгитюдных исследований* показывают, что у людей, отобранных в детстве с помощью элементарных тестов на умственные способности, значительно больше шансов стать штатными университетскими профессорами, изобретателями и успешными предпринимателями, чем у тех, кто показал менее выдающие результаты (соотношение составляет — 1 человек на 10 000 сверстников), см.: [Kell et al. 2013]. В книге Анны Роу «Как изготовить ученого» приведены данные о шестидесяти четырех известных ученых; выяснилось, что средний показатель их когнитивной способности на три-четыре стандартных, или среднеквадратических, отклонения выше нормы для всей популяции и намного выше показателя среднестатистического ученого [Roe 1953]. Кроме того, интеллектуальные способности могут соотноситься с общим объемом дохода на протяжении жизни и такими нематериальными показателями, как, например, средняя продолжительность жизни, количество разводов и вероятность выбыть из учебного заведения [Deary 2012]. Сдвиг распределения интеллектуальных способностей вверх будет иметь непропорционально большое значение для его «хвоста», в частности, увеличив количество высокоодаренных людей и уменьшив количество тех, у кого наблюдаются задержки в развитии и неспособность к обучению. См. также: [Bostrom, Ord 2006; Sandberg, Savulescu 2011].

[64] См., например: [Warwick 2002]. Стивен Хокинг даже предположил, что этот шаг будет необходимым для достижения прогресса в области машинного интеллекта: «Мы должны как можно быстрее разработать технологию прямого соединения мозга и компьютера, чтобы искусственный мозг внес свой вклад в повышение интеллектуальных способностей человека вместо того, чтобы ему противостоять» (цит. по статье: [Walsh 2001]). С ним согласен Рэй Курцвейл: «Что касается… рекомендации Хокинга о прямом соединении мозга с компьютером, то я согласен, что это и оправданно, и желательно, и неизбежно [sic], и я многие годы говорю об этом» [Kurzweil 2001].

[65] См.: [Lebedev, Nicolelis 2006; Birbaumer et al. 2008; Mak, Wolpaw 2009; Nicolelis, Lebedev 2009]. Более личный взгляд на проблему улучшения когнитивных способностей за счет имплантации

---

* *Лонгитюдное исследование* (longitudinal study) — продолжительное онтогенетическое исследование одних и тех же индивидов или групп людей; научный метод, применяемый в социологии и психологии.

можно найти в книге Майкла Хорста «Обновление. Как, став частью компьютера, я почувствовал себя человеком» [Chorost 2005, ch. 11].

[66] См.: [Smeding et al. 2006].

[67] См.: [Degnan et al. 2002].

[68] См.: [Dagnelie 2012; Shannon 2012].

[69] См.: [Perlmutter, Mink 2006; Lyons 2011].

[70] См.: [Koch et al. 2006].

[71] См.: [Schalk 2008]; обзор современных исследований и тенденций см: [Berger et al. 2008]. В книге «Киборг» Кевина Уорика приведены убедительные доводы в пользу этой системы как метода, повышающего интеллектуальные способности [Warwick 2002].

[72] Некоторые примеры см.: [Bartels et al. 2008; Simeral et al. 2011; Krusienski, Shih 2011; Pasqualotto et al. 2012].

[73] См.: [Hinke et al. 1993].

[74] Есть некоторые исключения, особенно на ранней стадии обработки зрительной информации. Например, первичная зрительная кора использует ретинотопическое представление, то есть смежные сборки нейронов собирают данные со смежных областей сетчатки (хотя глазные доминантные колонки несколько усложняют эту картину).

[75] См.: [Berger et al. 2012; Hampson et al. 2012].

[76] В случаях некоторых имплантатов потребуются две формы обучения: устройство учится интерпретировать нейронные представления организма; сам организм учится пользоваться системой, генерируя соответствующие модели возбуждения, см.: [Carmena et al. 2003].

[77] Некоторые исследователи предлагают рассматривать корпоративные структуры (компании, профсоюзы, правительства, церковь и тому подобное) в качестве агентов искусственного интеллекта, обладающих сенсорными и исполнительными механизмами, способных выражать знания, делать логические выводы и осуществлять действия; см., например: [Kuipers 2012] — ср. с дискуссией на тему, может ли существовать коллективное представление, см.: [Huebner 2008]. В отличие от людей, у агентов искусственного интеллекта совсем другие способности и внутреннее состояние; кроме того, они явно обладают бóльшими возможностями и намного успешнее в установлении связей с окружающей средой.

[78] См.: [Hanson 1995; Hanson 2000; Berg, Rietz 2003].

[79] Например, что мешает работодателю в его непосильной борьбе против злоупотреблений на рабочем месте прибегать к помощи детектора лжи, проверяя на нем сотрудников в конце каждого дня; причем вопросы могут быть самые простые: не украли ли они что-нибудь и трудились ли они с должным усердием. Интересно было бы поинтересоваться у политических деятелей и руководителей крупных компаний, насколько они действительно искренни в своих заверениях и готовы ли защищать интересы своих избирателей и акционеров. Иметь такую игрушку было бы очень на руку диктаторам — это облегчило бы им задачу выявлять мятежных генералов в ближнем окружении, а также вылавливать по всей стране открытых противников режима и потенциальных оппозиционеров.

[80] Можно представить методы нейровизуализации, когда с помощью компьютерной или магнитно-резонансной томографии обнаруживают нейронные следы так называемого объясненного сознания. Если в системе детектора лжи не заложена функция регистрации заблуждений, оборудование может давать сбой и ошибаться в случае проверки людей, подверженных самовнушению мыслей, не соответствующих действительности. Лучше всего метод выявления самообмана использовать при обучении рациональному мышлению и при изучении эффективных мер, направленных на избавление от когнитивных искажений.

[81] См.: [Bell, Gemmel 2009]. Одним из начинателей такого подхода к жизни стал исследователь Массачусетского технологического института Деб Рой; Оснастив дом многочисленными видеокамерами, он запечатлел каждую секунду развития своего сына в первые три года жизни. По мнению Роя, аудиовизуальные данные детского поведения и лепета могут быть полезны с точки зрения анализа развития речевых способностей [Roy 2012].

[82] Рост общей численности народонаселения с точки зрения уровня коллективного интеллекта явится лишь незначительным фактором. В случае создания машинного интеллекта население мира (включая цифровых агентов) в течение очень короткого времени вырастет на много порядков. Но такой путь к сверхразуму предполагает создание универсального искусственного интеллекта или полной эмуляции головного мозга, поэтому сейчас мы не будем затрагивать эту тему.

[83] См.: [Vinge 1993].

Глава 3. Типы сверхразума

[1] Вернон Виндж, описывая результат, который мог бы получиться, если запустить человеческий мозг с ускоренной скоростью, выбрал термин *слабый сверхразум* [Vinge 1993].

[2] Например, если сверхбыстродействующая система исполняет все, что делает любой человек, — разве что не танцует мазурку, — можем ли мы назвать ее скоростным сверхразумом? Наверное, нас все-таки будут интересовать те основные когнитивные способности, которые имеют значение тактического и стратегического порядка.

[3] Если принять во внимание разницу в быстродействии и затратах энергии между процессами, идущими в мозгу человека, и действиями более эффективной системы преобразования информации, скорость можно было бы увеличить как минимум в миллион раз. Скорость света более чем в миллион раз выше скорости передачи сигнала нейронами, синапсы тратят более чем в миллион раз больше тепла, чем это оправданно с точки зрения термодинамики, а частоты, на которых работают современные транзисторы, более чем в миллион раз выше, чем частоты синапсисов, см.: [Yudkowsky 2008 a]; см. также: [Drexler 1992]. Предел скорости сверхразума задается следующими факторами: задержки в высокоскоростных сетях передачи данных; квантовые ограничения на скорость изменения состояний; объем емкости, вмещающей «разум» [Lloyd 2000]. «Предельный ноутбук», описанный в работе Сета Ллойда, на котором можно было бы проводить эмуляцию мозга, выполнял бы $1{,}4 \times 10^{21}$ FLOPS со скоростью $3{,}8 \times 10^{29}$ операций в секунду (при условии, что их можно было бы существенным образом распараллелить) [Lloyd 2000]. Однако расчеты Ллойда не претендуют на абсолютную

достоверность, его задачей было проиллюстрировать ограничения вычислений — ограничения, которые непосредственно вытекают из основных физических законов.

[4] Когда речь заходит о полной эмуляции, сразу возникает вопрос, как долго сможет работать имитационная модель головного мозга человека, пока не сойдет с ума и не начнет «ходить по кругу». Совсем не очевидно — даже при условии разнообразия задач и регулярных передышек, — что имитационной модели удастся протянуть тысячи субъективных лет, ни разу не столкнувшись с психологическими проблемами. Кроме того, если из-за предельного количества нейронов объем полной памяти ограничен, то накапливать знания до бесконечности не получится, поэтому, начиная с какой-то точки, имитационной модели мозга, чтобы освободить место для усвоения новой информации, придется начать забывать уже усвоенное. (Думается, что искусственный интеллект можно будет разработать на таких принципах, чтобы постараться избежать таких потенциальных проблем.)

[5] Таким образом, для наномеханизмов, двигающихся с довольно скромной скоростью 1 м/с, обычный масштаб времени — это наносекунды; см.: [Drexler 1992] — раздел 2.3.2. Робин Хэнсон отмечает работу роботов-«фей», размером всего 7 мм, двигающихся в 260 раз быстрее привычной нам скорости [Hanson 1994].

[6] См.: [Hanson 2012].

[7] Термин *коллективный интеллект* не имеет отношения к распараллеливанию на нижнем уровне аппаратного обеспечения, это понятие целиком лежит в плоскости интеллектуальных автономных агентов, например людей. Если начать проводить одну полную эмуляцию мозга, одновременно используя большое количество сверхмощного оборудования, можно получить скоростной сверхразум, но он не будет коллективным сверхразумом.

[8] Улучшение когнитивных возможностей, таких как скорость и качество ума, может лишь косвенно влиять на производительность коллективного разума, но об этих вариантах усовершенствования будет правильнее говорить при обсуждении двух других форм сверхразума.

[9] Считается, что именно рост плотности населения привел к верхнепалеолитической революции (культурная революция кроманьонцев), а после преодоления некой пороговой величины так называемая культурная сложность начала расти очень быстро; см.: [Powell et al. 2009].

[10] Кажется, мы забыли упомянуть интернет? Увы, его развитие пока не достигло того масштаба, который требуется для «большого прыжка». Может быть, со временем интернет подтянется. Другим приведенным здесь факторам потребовались века и даже тысячелетия, чтобы раскрыть свой потенциал полностью.

[11] Пример, взятый мною, конечно, из области фантастики. Планета — большая настолько, что может вместить семь квадриллионов мегаземлян, — при современном уровне технологий просто взорвалась бы, если не состояла бы из очень легкого вещества или не была бы полой, удерживаясь в таком состоянии давлением или каким-то иным искусственным способом. (Лучшим решением могла бы быть сфера Дайсона, или раковина Дайсона.) На такой огромной поверхности история развивалась бы совсем иначе. Но не будем отклоняться в сторону.

[12] Нас больше интересуют функциональные свойства комплексного интеллекта, а не вопрос, будут ли у него квалиа, то есть будет ли он рассудком, имеющим опыт субъективных переживаний и ощущений. (Кого-то, впрочем, может заинтересовать, какого рода опыт сознания будет присущ интеллекту, интегрированному более или менее человеческого. По поводу самосознания существуют разные мнения, например, в теории глобального рабочего пространства предполагается, что более интегрированный мозг будет обладать и более полным самосознанием; см. также: [Baars 1997; Shanahan 2010; Schwitzgebel 2013].)

[13] Небольшие группы людей, ставшие изолированными в силу обстоятельств, могут продолжать извлекать пользу из ранее приобретенной суммы знаний. Например, пользоваться языком, который был создан в гораздо более крупной языковой среде, или инструментами, придуманными до того, как они оказались в изоляции. Но даже если такое сообщество всегда жило изолированно, оно все равно остается частью более крупного коллективного разума, состоящего из всех предыдущих поколений, который можно считать системой последовательной обработки информации.

[14] В соответствии с тезисом Чёрча–Тьюринга все вычислимые функции вычислимы машиной Тьюринга. Поскольку любой из трех типов сверхразума можно симулировать на машине Тьюринга (при наличии доступа к неограниченному объему памяти и возможности работать независимо), по этому формальному критерию они эквивалентны с вычислительной точки зрения. На самом деле среднестатистический человек тоже мог бы повторить машину Тьюринга (при наличии неограниченного запаса бумаги и времени) и поэтому эквивиалентен ей по тому же критерию. Но для наших целей важно, чего эти различные системы могут достичь *на практике*, в условиях ограниченной памяти и за разумное время. И разница в эффективности настолько велика, что ее нельзя не заметить. Например, среднестатистического человека с коэффициентом интеллекта, равным 85 пунктам, вполне можно научить моделировать машину Тьюринга. (Скорее всего, этому можно научить даже какую-нибудь особенно одаренную и послушную шимпанзе.) Но в практическом смысле такой человек вряд ли способен самостоятельно разработать общую теорию относительности или получить Филдсовскую премию.

[15] И коллективное народное творчество способно рождать великую литературу (взять, например, эпические поэмы Гомера) — в фольклоре тоже встречались гениальные исполнители.

[16] Если только коллективный сверхразум не содержит элементы, характерные для двух остальных типов сверхразума: скоростного, или качественного, или обоих вместе.

[17] Наша неспособность определить все подобные задачи отчасти связана с тем, что мы и не пробуем это сделать, поскольку нет смысла тратить время на детализацию интеллектуальных заданий, которые не смогут выполнить ни люди, ни все известные на сегодняшний день организации. Но возможно, что даже их концептуальное описание является одной из таких задач, которые мы сейчас не можем выполнить в силу ограниченности своего ума.

[18] См.: [Boswell 1917]; см. также: [Walker 2002].

[19] Эта мощность достижима в моменты резкого всплеска активности отдельных групп нейронов, обычная скорость их работы гораздо ниже; см.: [Gray, McCormick 1996; Steriade et al. 1998].

Есть также быстро возбуждающие нейроны (chattering neurons), которые достигают частоты в 750 Гц, но, похоже, они представляют редкое исключение.

[20] См.: [Feldman, Ballard 1982].

[21] Скорость передачи информации зависит от диаметра аксона (чем аксон толще, тем она выше) и наличия у него миелиновой оболочки. В пределах центральной нервной системы задержки лежат в диапазоне меньше чем от миллисекунды до 100 мс [Kandel et al. 2000]. Скорость передачи информации в оптоволокне равна примерно 68% скорости света (зависит от коэффициента преломления материала). В электрических кабелях скорость передачи примерно такая же, 59–77% скорости света.

[22] Это при скорости передачи сигнала, равной 70% скорости света. При 100% оценка вырастает до $1{,}8 \times 10^{18}$ м^3.

[23] Количество нейронов в мозгу взрослого мужчины примерно равно 86,1 ± 8,1 млрд, эти данные получены в результате подсчета клеточных ядер в ткани мозга, помеченных специальным маркером. В прошлом оценки обычно попадали в диапазон 75–125 млрд нейронов. Они были основаны на ручном подсчете количества клеток в сравнительно небольших областях мозга, см.: [Azevedo et al. 2009].

[24] См.: [Whitehead 2003].

[25] Вполне возможно, что в системах обработки информации могут использоваться процессы вычисления и хранения данных молекулярного масштаба, а сами такие системы способны достигать как минимум планетарных масштабов. Однако предельные физические размеры вычислительной системы, ограниченные законами квантовой механики, общей теории относительности и термодинамики, гораздо меньше этого «юпитероподобного» уровня [Sandberg 1999; Lloyd 2000].

[26] См.: [Stansberry, Kudritzki 2012]. На долю центров хранения и обработки данных приходится примерно 1,1–1,5% общего объема потребления электроэнергии в мире, см.: [Koomey 2011]; см. также: [Muehlhauser, Salamon 2012].

[27] Это упрощение. Количество блоков информации, которые могут храниться в рабочей памяти, зависит от характера как самой информации, так и задачи; тем не менее понятно, что их не может быть очень много; см.: [Miller 1956; Cowan 2001].

[28] Пример: трудность изучения логических концепций (категорий, определенных правилами логики) пропорциональна длине наиболее коротких логически эквивалентных им пропозициональных формул. Обычно довольно трудно запоминаются даже формулы, состоящие всего из 3–4 букв; см.: [Feldman 2000].

[29] См.: [Landauer 1986]. Это исследование основано на экспериментальных оценках скорости запоминания и забывания информации людьми. Если учесть еще и неявное обучение, оценка может чуть улучшиться. Исходя из допущения, что емкость хранения равна примерно 1 биту на синапс, *верхнюю границу* емкости человеческой памяти можно оценить в 10^{15} бит. Обзор других оценок см.: [Sandberg, Bostrom 2008, Appendix A].

[30] Шум в канале передачи информации может привести к ложным срабатываниям, поэтому из-за синаптического шума сила передаваемых сигналов меняется довольно сильно. Похоже,

нервной системе пришлось найти множество компромиссов между терпимостью к шуму и издержками (масса, размер, длительность задержки); см.: [Faisal et al. 2008]. Например, аксоны не могут быть тоньше 0,1 мкм, в противном случае возможна генерация спонтанных потенциалов действия, вызванная случайным открытием ионных каналов [Faisal et al. 2005].

[31] См.: [Trachtenberg et al. 2002].

[32] С точки зрения объема памяти и вычислительной мощности, но не энергоэффективности. Самый быстрый на момент написания этой книги китайский суперкомпьютер Tianhe-2, который в июне 2013 года обошел Titan, созданный Cray Inc., имеет быстродействие в 33,86 Пфлопс. Он использует 17,6 МВт электроэнергии, что почти на шесть порядков больше, чем нужные мозгу 20 Вт.

[33] Обратите внимание, что этот список имеет *альтернативный* характер: цифровой мозг существенно превзойдет человеческий даже в том случае, если некоторые из перечисленных преимуществ окажутся иллюзорными, поскольку для этого достаточно, чтобы имело место хотя бы одно из них.

Глава 4. Динамика взрывного развития интеллекта

[1] Может получиться, что система не пройдет какой-то из этих уровней. Вместо этого она постепенно, в течение какого-то времени, начнет опережать исследователей в выполнении все более сложных заданий на разработку самосовершенствования.

[2] За последние пятьдесят лет широко признан вполне реальным как минимум один сценарий, по которому предполагается исчезновение существующего мирового порядка в течение если не минут, то часов, — мировая ядерная война.

[3] Это согласуется с наблюдениями, что за последние годы в некоторых высокоразвитых странах, таких как Великобритания, Дания и Норвегия, похоже, перестал действовать или даже начал действовать в обратную сторону эффект Флинна — рост среднего коэффициента интеллекта у большей части населения примерно на 3 пункта за 10 лет, — исправно работавший на протяжении последних 60 лет; см.: [Teasdale, Owen 2008; Sundet et al. 2004]. Причина возникновения эффекта Флинна — и то, до какой степени он представляет истинные сдвиги в общем интеллектуальном уровне, а до какой является следствием роста навыков прохождения тестов на коэффициент интеллекта, — является предметом ожесточенных споров и до сих пор неизвестна. Даже если этот эффект (хотя бы отчасти) отражает реальный рост умственных способностей и даже если он в последнее время не действует или действует в обратную сторону, это не означает, что исчезла причина его возникновения, какой бы она ни была. Происходящие в последние годы изменения могут быть вызваны каким-то независимым фактором, который при отсутствии указанной причины мог бы привести к еще большему спаду коэффициента интеллекта.

[4] См.: [Bostrom, Roache 2011].

[5] Временной фактор можно было бы скорректировать с помощью соматической генной терапии, но она гораздо сложнее технически и имеет меньший потенциал, чем вмешательство на эмбриональном уровне.

[6] Средний рост производительности в мировой экономике за период 1960–2000 гг. составил 4,3%; см.: [Isaksson 2007]. На повышение организационной эффективности приходится лишь

часть этого роста. Конечно, *некоторые* сети и процессы совершенствовались гораздо более быстрыми темпами.

[7] Эволюция биологического мозга ограничена многими факторами и небходимостью компромиссов, число которых резко снижается при переходе мозга «на цифровую платформу». Например, размер мозга определяется размером головы, а слишком большая голова вызовет проблемы при прохождении через родовые пути. Метаболические процессы, происходящие в таком мозгу, потребуют слишком большого расхода энергии. Степень связанности различных областей мозга также имеет стерические ограничения: объем белого вещества значительно превышает объем серого, которое оно соединяет. Отвод тепла ограничивается кровотоком и может уже находиться на пределе для приемлемого функционирования мозга. Более того, биологические нейроны являются медленными, зашумленными, нуждаются в постоянной защите и уходе, а также требуют большого количества глиальных клеток и кровеносных сосудов (из-за чего в черепной коробке становится слишком «тесно»). Обо всем это см.: [Bostrom, Sandberg 2009 b].

[8] См.: [Yudkowsky 2008 a, p. 326]; более свежее обсуждение вопроса см.: [Yudkowsky 2013].

[9] Для простоты интеллектуальные способности представлены на рисунке в виде однопараметрической функции. Но это не имеет значения с точки зрения обсуждаемой темы. Точно так же можно было бы, например, представить их профиль в виде гиперповерхности в многомерном пространстве.

[10] См.: [Lin et al. 2012].

[11] Определенный рост возможностей коллективного интеллекта можно обеспечить за счет простого увеличения количества его членов. Это позволит увеличить производительность системы как минимум за счет задач, выполнение которых можно легко распараллелить. Однако для реализации полного потенциала взрывного роста популяции придется обеспечить некоторый (не самый минимальный) уровень координации между этими членами.

[12] Различия между скоростью и качеством интеллекта размываются в случаях, не относящихся к нейроморфным системам ИИ.

[13] См.: [Rajab et al. 2006, p. 41–52].

[14] Предположительно за счет использования конфигурируемых микросхем вместо процессоров общего назначения можно было бы увеличить скорость вычислений в нейронных сетях на два порядка; см.: [Markram 2006]. По результатам исследования климатических моделей на компьютерах, способных выполнять петафлопсы операций, оказалось, что использование кастомизированных процессоров приводит к снижению издержек в 22–24 раза, а потребности в электроэнергии падают примерно на два порядка; см.: [Wehner et al. 2008].

[15] См.: [Nordhaus 2007]; есть множество обзоров различных значений закона Мура, см., например: [Tuomi 2002; Mack 2011].

[16] Если развитие идет довольно медленно, разработчики проекта могут воспользоваться достижениями в других областях, например успехами или университетских исследователей в области кибернетики, или производителей микропроцессоров.

[17] Вероятность возникновения алгоритмического навеса невысока, разве что появится такое экзотическое аппаратное обеспечение, как квантовые компьютеры, которые позволят

запускать невозможные прежде алгоритмы. Кто-то может возразить, что примерами алгоритмического навеса являются нейронные сети и глубокое машинное обучение. В момент изобретения они были слишком затратными с точки зрения вычислительной мощности, на некоторое время оказались отложенными в сторону, а затем вновь стали востребованными — когда быстрые графические процессоры снизили затраты на их работу. И теперь они на коне.

[18] Даже если продвижение к человеческому интеллектуальному уровню окажется медленным.

[19] $\mathcal{O}_{\text{мир}}$ — это часть оптимизирующей силы, приложенной для совершенствования рассматриваемой системы. Для проекта, который реализуется в условиях относительной изоляции и не получает заметной поддержки извне, $\mathcal{O}_{\text{мир}} \approx 0$, даже если изначальные ресурсы (компьютеры, научные концепции, обученный персонал и т. д.) представляют собой часть мировой экономики и стали результатом многих веков развития человечества.

[20] Наиболее важной среди всех когнитивных способностей зародыша ИИ является его способность выполнять интеллектуальную работу по саморазвитию, то есть способность повышать уровень своего интеллекта. (Если зародыш ИИ специализируется на улучшении другой системы, которая, в свою очередь, улучшает его, их можно рассматривать как подсистемы более крупной системы и анализировать в совокупности.)

[21] Когда сопротивляемость не настолько высока, что может привести к отказу от инвестирования или переключению на другой проект.

[22] Аналогичный пример обсуждается Юдковским [Yudkowsky 2008 b].

[23] Поскольку входные параметры тоже растут (скажем, инвестиции в строительство новых заводов и количество людей, занятых в отрасли производства полупроводников), закон Мура как таковой не обеспечивает требуемого роста при исключении влияния этих параметров. Однако в сочетании с успехами в создании ПО допущение об удвоении возможностей систем каждые 18 месяцев может оказаться правдоподобным.

[24] Было сделано несколько экспериментальных попыток развить идею о взрывном развитии интеллекта в рамках теории экономического роста; см., например: [Hanson 1998 b; Jones 2009; Salamon 2009]. В этих исследованиях указывается на возможность чрезвычайно быстрого экономического роста в случае создания цифрового разума, но поскольку теория эндогенного роста сравнительно плохо разработана даже в применении к историческим и современным условиям, ее применение к потенциально недолговечным будущим условиям лучше пока считать источником потенциально полезных концепций и мыслей, чем методом, способным обеспечить нас надежными прогнозами. Обзор попыток создать математическую модель технологической сингулярности см.: [Sandberg 2010].

[25] Вполне возможно, что никакого взлета не будет. Но поскольку, как обсуждалось ранее, появление сверхразума представляется технически возможным, взлет может не начаться лишь в случае вмешательства какой-то внешней силы, скажем, экзистенциальной катастрофы, то есть окончательной глобальной катастрофы, которая приведет к вымиранию человечества. Если сильный сверхразум появится не в форме УИИ или имитационной модели мозга, а иными путями, тогда более вероятно развитие по сценарию медленного взлета.

Глава 5. Решающее стратегическое преимущество

[1] Цифровой разум может быть запущен на единственной машине, а не с помощью глобальной компьютерной сети, но мы имеем в виду не такую концентрацию. Нас больше интересует степень концентрации власти — особенно такой, которую обеспечивает технологическое превосходство, — на последних стадиях революции машинного интеллекта или сразу после ее завершения.

[2] Например, в развивающихся странах потребительские товары, как правило, выводятся на рынок медленнее, см.: [Talukdar et al. 2002]; см. также: [Keller 2004; World Bank, 2008].

[3] Для настоящего обсуждения этой проблемы следовало бы привлечь серьезную экономическую литературу, посвященную теории организаций. Бесспорно, классическим источником до сих пор служит фундаментальный труд Рональда Коуза «Природа фирмы» [Coase 1937]; см. также: [Canbäck et al. 2006; Milgrom, Roberts 1990; Hart 2008; Simester, Knez 2002].

[4] Тем не менее зародыш ИИ легко выкрасть — передать по сети или унести на флешке.

[5] Согласно некоторым исследователям, шелк мог использоваться уже в культуре Яншао (5000–3000 гг. до н. э.), см.: [Barber 1991]. В относительно недавней работе приведено приблизительное время одомашнивания тутового шелкопряда (на основании генетических исследований) — около 4100 лет назад [Sun et al. 2012].

[6] См.: [Cook 1984, p. 144] — эта легенда слишком красива, чтобы быть исторически правдоподобной, как и история, рассказанная Прокопием Кесарийским в его «Войнах» (VIII, 1–7), что тутовые шелкопряды были доставлены в Византию странствующими монахами в их бамбуковых посохах, см.: [Hunt 2011].

[7] См.: [Wood 2007; Temple 1986].

[8] В доколумбийских культурах колесо было известно, но использовалось только для игр (возможно, из-за отсутствия подходящих тягловых животных).

[9] См.: [Koubi 1999; Lerner 1997; Koubi, Lalman 2007; Zeira 2011; Judd et al. 2012].

[10] Оценка дана на основании множества источников. Скорее всего, показатель разрыва подсчитывался очень приблизительно и зависел от того, какие возможности считались сопоставимыми. Через пару лет после изобретения радара он использовался как минимум двумя странами, но точное время в месяцах рассчитать довольно трудно.

[11] См.: [Ellis 1999].

[12] Испытанная в 1953 году РДС-6 стала первой бомбой с использованием термоядерной реакции, но первую «настоящую» термоядерную бомбу РДС-37, у которой на долю реакции синтеза приходилась большая часть мощности взрыва, испытали только в 1955 году.

[13] Подтверждений нет.

[14] Испытания прошли в 1989 году, в 1994-м проект был закрыт.

[15] Развернутая система, радиус действия превышает 5000 км.

[16] Ракеты Polaris, приобретенные у США.

[17] Ведутся работы над ракетой Taimur, видимо, на базе китайской технологии.

[18] Испытанный в 1989–1990 гг. ракетоноситель RSA-3 предполагалось использовать как для запусков спутников, так и в качестве МБР.

[19] Многозарядные боеголовки индивидуального наведения позволяют использовать одну баллистическую ракету для поражения множества целей.

[20] Система Agni V пока не принята на вооружение.

[21] Если мы считаем, что время отставания реализации проектов подчиняется нормальному закону распределения, тогда, скорее всего, расстояние между проектом-лидером и его ближайшим преследователем также должно зависеть от общего количества разработок. Когда проектов много, отрыв первого от второго будет относительно небольшим даже тогда, когда дисперсия довольно велика (и будет медленно сокращаться с ростом количества конкурентов). Однако маловероятно, что будет слишком много проектов, обладающих достаточными ресурсами, чтобы всерьез претендовать на успех. (Обилие проектов возможно лишь при обилии различных подходов и методик — большинство которых, скорее всего, окажутся тупиковыми.) Современный опыт говорит, что серьезных претендентов на достижение любой заметной с технологической точки зрения цели обычно можно пересчитать по пальцам. Несколько отличная ситуация на потребительском рынке, поскольку там существует множество ниш для лишь слегка отличающихся продуктов, а входные барьеры слишком низки. Есть тьма проектов по дизайну футболок, когда в каждой разработке принимает участие один человек, но всего несколько компаний в мире участвуют в разработке следующего поколения графических карт. (В настоящий момент фактической дуополией обладают AMD и NVIDIA, а в сегменте карт с низкой производительностью с ними конкурирует Intel.)

[22] См.: [Bostrom 2006 c]. При желании можно вообразить разные варианты, например: синглтон, чье существование вообще невидимо (когда сверхразум обладает настолько развитыми технологиями и проницательностью, что может незаметно для людей управлять всем миром); синглтон, добровольно установивший жесткие самоограничения своей власти (пунктуально соблюдает международные нормы и договоры); синглтон-либертарианец. Какова вероятность возникновения синглтона того или иного типа — вопрос скорее умозрительный, пока не будет проверено на опыте. Однако почему бы не дать волю фантазии хотя бы на *концептуальном уровне*: хороший синглтон, плохой синглтон, дерзкий синглтон, вздорный синглтон, приторно-любезный синглтон, деспотичный синглтон с признаками паранойи, вопящий самодур синглтон, порою напоминающий разбушевавшуюся природную стихию.

[23] См.: [Jones 1985, p. 344].

[24] Нельзя забывать, что Манхэттенский проект был реализован во время войны. Многие ученые вспоминали, что дали согласие на участие в нем в основном по двум мотивам: помочь стране в ситуации военного времени и из страха, что нацистская Германия опередит союзников и первой создаст атомное оружие. В мирное время многим правительствам оказалось бы не по плечу мобилизовать такое количество работников — среди которых были крупнейшие ученые — на участие в столь масштабном и секретном мероприятии. Еще один классический пример научно-технического мегапроекта — программа Apollo. Но и она получила столь мощную поддержку лишь благодаря противостоянию, рожденному холодной войной.

[25] Впрочем, даже если такие проекты и *искались бы изо всех сил*, то делалось бы это таким образом, что вряд ли стало бы достоянием общественности.

[26] Благодаря методам криптографии команда проекта вполне могла бы работать распределенно. В такой цепочке передачи информации самый уязвимый момент — первый, когда информация вводится в систему, и теоретически можно проследить за действиями печатающего человека. Но если широкое распространение получит слежка внутри помещений (при помощи микроскопических записывающих устройств), то ради столь ценной секретности можно принять контрмеры (например, специальные помещения, защищенные от проникновения в них шпионской аппаратуры). Причем если физическое пространство в наступающую эпоху всеобщего контроля становится все более прозрачным, то киберпространство вполне может оказаться лучше защищенным благодаря все более широкому распространению довольно надежных криптографических протоколов.

[27] Тоталитарные страны наверняка догадаются возродить менее технологичные, но более действенные меры. Можно, как в СССР, помещать ученых, работающих над интересующими государство направлениями, в трудовые лагеря вроде сталинских «шарашек», которые в дальнейшем переродились в респектабельные и абсолютно закрытые академгородки.

[28] Когда в обществе низок уровень социальной тревоги, некоторые ученые обычно не прочь чуть подогреть степень беспокойства, чтобы привлечь внимание к своей работе и представить направление, которым занимаются, чуть более важным и интересным. Когда в обществе уровень тревоги повышен, те же ученые меняют тон, поскольку начинают беспокоиться о снижении финансирования, введении новых норм регулирования и негативной общественной реакции. Специалистов из смежных дисциплин (например, кибернетика и робототехника), не имеющих прямого отношения к искусственному интеллекту, может возмущать такой перекос в финансировании и внимании. Они также могут справедливо считать, что *их* работа не несет риска возникновения потенциально опасных последствий, как в случае со взрывным развитием искусственного интеллекта. (Некоторые исторические параллели мы находим в истории развития идеи нанотехнологий; см.: [Drexler 2013].)

[29] Эти проекты были удачны в том смысле, что достигли некоторых целей хотя бы минимально. Насколько успешными они явились в более широком плане (принимая во внимание экономическую эффективность), сказать трудно. Например, при разработке проекта МКС был допущен серьезный перерасход бюджета и задержки по срокам. Подробнее о проблемах, возникших в ходе проекта, см.: [NASA, 2013]. С разработкой Большого адронного коллайдера связано несколько серьезных неудач, но, скорее всего, они были вызваны первоначальной сложностью поставленной задачи. Проект «Геном человека» в конечном счете достиг успеха, но, похоже, благодаря дополнительному импульсу, который ему придала конкуренция с частным проектом Крейга Вентера. Пользовавшиеся международной поддержкой проекты по контролю над термоядерной энергией, несмотря на огромные инвестиции, не оправдали ожиданий, вероятно, все по той же причине — поставленная задача оказалась слишком сложной.

[30] См.: [US Congress, 1995].

[31] См.: [Hoffman 2009; Rhodes 2008].

[32] См.: [Rhodes 1986].

[33] Криптослужба ВМФ США OP-20-G, похоже, сознательно проигнорировала британцев, готовых поделиться с союзником методами расшифровки «Энигмы»; руководители OP-20-G не сообщили об этом предложении высшему руководству страны, см.: [Burke 2001]. В результате у США сложилось впечатление, что Британия утаила от них важную информацию, что привело к ненужной напряженности в ходе войны. Некоторые данные, полученные в результате дешифровки немецких сообщений, британская контрразведка передавала СССР. В частности, Сталин был предупрежден о подготовке операции «Барбаросса» (план вторжения Германии в СССР), но он отказался верить этому предупреждению, отчасти потому что британцы не сообщили ему об источнике сведений.

[34] Бертран Рассел защищал возможность ядерной войны в течение нескольких лет — чтобы заставить Россию принять план Баруха; позднее Рассел стал последовательным сторонником взаимного ядерного разоружения [Russell, Griffin 2001]. Как известно, Джон фон Нейман был убежден в неизбежности войны между США и Россией и даже как-то сказал: «Если вы говорите: „Почему бы не разбомбить [русских] завтра“, — я спрошу: „А почему не сегодня?“ Если вы скажете: «Сегодня в пять», — я спрошу: „а почему бы не в час?“» [Blair 1957, p. 96]. (Вероятно, он сделал свое нашумевшее заявление, чтобы заработать дополнительные очки у ястребов-антикоммунистов из Министерства обороны — ведь то были времена Маккарти. Невозможно с уверенностью заявлять, что фон Нейман действительно нанес бы первым удар, отвечай он за политику США.)

[35] См.: [Baratta 2004].

[36] Если ИИ контролируется группой людей, проблема может возникнуть с ними, хотя есть шанс, что к тому моменту появятся новые способы обеспечить выполнение общественных договоров, в таком случае даже человеческие сообщества смогут избежать угрозы внутреннего противостояния и образования враждебных коалиций.

Глава 6. Разумная сила, не имеющая себе равной

[1] Каковы критерии, по которым человек считается доминирующим видом на Земле? С экологической точки зрения он является самым распространенным крупным (~ 50 кг) млекопитающим, но общая сухая человеческая биомасса (~ 100 млрд кг) не очень впечатляет по сравнению с биомассой муравьев (300–3000 млрд кг). На долю человека и его микрофлоры приходится ничтожно малая часть (< 0,001) общей биомассы планеты. При этом территория, используемая под посадки и пастбища, представляет собой одну из крупнейших экосистем и покрывает примерно 35% свободной ото льда поверхности суши [Foley et al. 2007]. Человек присваивает почти четверть чистой первичной продукции — это согласно стандартным оценкам [Haberl et al. 2007], хотя в общем — в зависимости от понятийных подходов и терминологических принципов — их диапазон составляет от 3% до более чем 50% [Haberl et al. 2013]. У человека самый обширный среди всех животных ареал обитания; человек находится на вершине пищевой цепи.

[2] См.: [Zalasiewicz et al. 2008].

[3] См. первое примечание к этой главе.

[4] Строго говоря, это высказывание не совсем корректно. Интеллектуальный уровень человека может быть сколь угодно низким, а в каких-то случаях даже достигать нулевой отметки (например, эмбрион или человек, находящийся в вегетативном состоянии). Поэтому в качественном выражении, наверное, максимальное различие когнитивных способностей в пределах человеческого вида все-таки более существенное, чем дистанция между любым человеком и сверхразумом. Конечно, если, произнося *человек*, мы имеем в виду «среднестатистический дееспособный взрослый человек» — тогда сказанное вполне справедливо.

[5] См.: [Gottfredson 2002]; см. также: [Carroll 1993; Deary 2001].

[6] См.: [Legg 2008] — если в общих чертах, то Шейн Легг предлагает оценивать агента, обучающегося с подкреплением, по ожидаемой эффективности во всех средах с нормированным вознаграждением, где каждая среда получает вес в зависимости от ее колмогоровской сложности (обучение с подкреплением мы рассмотрим в главе 12). См. также: [Dowe, Hernandez-Orallo 2012; Hibbard 2011].

[7] В таких областях исследований, как биотехнологическая и нанотехнологическая, сверхразум мог бы отлично проявить себя в деле проектирования и моделирования новых структур. Но поскольку гениальный проект и успешная модель не могут заменить физических экспериментов, эффективность сверхразума будет ограничена уровнем его доступа к средствам, необходимым для опытных проверок.

[8] Например: [Drexler 1992; 2013].

[9] Специализированный ИИ мог бы, конечно, иметь значительную коммерческую ценность, но это не значит обрести сверхмощь в области экономической эффективности. Например, даже если специализированный ИИ поможет своим владельцам зарабатывать несколько миллиардов долларов в год, эта сумма все же будет на четыре порядка меньше объема всей мировой экономики. Чтобы ИИ мог непосредственно и значительно увеличить мировой ВВП, он должен уметь выполнять различные виды работ, то есть обладать компетенциями во многих областях.

[10] Это не исключает развития любого сценария, по которому ИИ потерпит поражение. Например, ИИ может сознательно выбрать высокий уровень риска, имеющий большую вероятность неудачи. Однако в этом случае условие принимает другие формы: 1) ИИ объективно оценивает, что шансы на успех довольно низки; 2) ИИ считает, что при игре возможны лишь самые высокие ставки, и просто не берет во внимание другие варианты действий — те варианты, которые мы, современные люди, в отличие от ИИ замечаем и можем просчитать.

[11] См.: [Freitas 2000; Vassar, Freitas 2006].

[12] См.: [Yudkowsky 2008 a].

[13] См.: [Freitas 1980; Freitas, Merkle 2004, ch. 3; Armstrong, Sandberg 2013].

[14] См., например: [Huffman, Pless 2003; Knill et al. 2000; Drexler 1986].

[15] Иначе говоря, дистанция будет небольшой при каком-то «естественном» показателе, например логарифме численности населения, которому можно обеспечить прожиточный минимум, если на то направить все имеющиеся ресурсы.

[16] Эта оценка основана на данных космического аппарата WMAP* о космологической плотности барионов, равной $9{,}9 \times 10^{-30}$ г/см^3, и предположении, что 90% этой массы приходится на межгалактический газ, что 15% общей галактической массы (около 80% массы барионов) составляет масса звезд и что масса средней звезды равна 0,7 массы Солнца; см.: [Read, Trentham 2005; Carroll, Ostlie 2007].

[17] См.: [Armstrong, Sandberg 2013].

[18] Даже при движении со скоростью, равной 100% скорости света (которая недостижима для объектов с массой покоя, отличной от нуля), можно добраться не больше чем до 6×10^9 галактик; см.: [Gott et al. 2005; Heyl 2005]. Мы исходим из того, что наше нынешнее понимание соответствующих физических законов верное. Но с уверенностью говорить о любой верхней границе сложно, поскольку вполне вероятно, что сверхразумная цивилизация может увеличить доступное ей пространство какими-то путями, кажущимися нам сегодня невозможными (например, создав машину времени, или породив новые расширяющиеся вселенные, или каким-то иным невообразимым способом).

[19] Здесь дана лишь очень грубая оценка, так как нам неизвестно, сколько обитаемых планет приходится на одну звезду. Уэсли Трауб считает, что у трети звезд спектральных классов F, G и K есть как минимум одна планета земного типа в зоне, где возможна жизнь [Traub 2012]; см. также: [Clavin 2012]. На звезды классов F, G, K приходится примерно 22,7% звезд в окрестностях Солнечной системы, то есть у 7,6% звезд есть потенциально подходящие для жизни планеты. Кроме того, такие планеты могут быть у более многочисленных звезд класса M, см.: [Gilster 2012]; см. также: [Robles et al. 2008].

Совсем необязательно, чтобы опасностям межгалактических путешествий подвергались человеческие существа. За тем, как продвигается колонизация космоса, вполне способен присмотреть сам ИИ. А *Homo sapiens* позволят размножаться ради получения данных, которые ИИ будет использовать для создания новых представителей нашего вида. Например, синтезируют ДНК на базе имеющейся генетической информации, выведут первых людей в инкубаторе, а затем вырастят и обучат их при помощи воспитателей, обладающих ИИ и антропоморфной внешностью.

[20] См.: [O'Neill 1974].

[21] Фримен Дайсон утверждает [Dyson 1960], что позаимствовал свою идею у британского писателя-фантаста Олафа Стэплдона ([Olaf Stapledon 1937]), который, в свою очередь, мог черпать полезные для себя мысли в трудах британского ученого Джона Бернала, см.: [Dyson 1979, p. 211].

[22] Принцип Ландауэра гласит, что минимальное количество энергии, которая необходима для изменения одного бита информации, равна $kT \ln 2$, где k — константа Больцмана ($1{,}38 \times 10^{-23}$ J/K), а T — температура. Если предположить, что схема работает при температуре около 300 K, тогда 10^{26} Вт позволяет нам стирать примерно 10^{47} бит в секунду. (О достижимой эффективности

---

* *WMAP* (Wilkinson Microwave Anisotropy Probe) — космический аппарат НАСА, созданный для изучения реликтового излучения, образовавшегося в результате Большого взрыва; запущен 30.06.2001.

наномеханических вычислительных устройств см.: [Drexler 1992]; см. также: [Bradbury 1999; Sandberg 1999; Ćirković 2004]. По поводу обоснованности принципа Ландауэра все еще идут споры, см., например: [Norton 2011].)

[23] Звезды различаются по своей мощности, но Солнце представляет собой довольно типичную звезду главной последовательности.

[24] Возможно провести более подробный анализ и внимательно изучить, какие типы вычислений нас интересуют. Количество возможных *последовательных* вычислений довольно ограниченно, поскольку быстрый последовательный компьютер должен быть небольшого размера, чтобы минимизировать задержки в передаче данных между различными его частями. Есть также ограничение на количество бит, которые можно хранить, и, как мы видели, на количество необратимых вычислительных операций (включая стирание информации), которые можно выполнить.

[25] Мы исходим из предположения, что внеземных цивилизаций, которые могут встать у него на пути, просто не существует. А также из ложности гипотезы симуляции, см.: [Bostrom 2003 a]. Если какое-то из этих предположения неверно, могут возникнуть серьезные катаклизмы неантропогенного характера — то есть те, которые вызваны агентами, отличными от людей; см.: [Bostrom 2003 b; 2009 c].

[26] Например, благоразумный синглтон, осознав идею эволюции, мог бы предпринять попытки постепенно повысить уровень своего коллективного интеллекта, используя для этого евгенические программы.

[27] См.: [Tetlock, Belkin 1996].

[28] Уточним, что колонизация, а также реинжиниринг заметной и доступной части Вселенной, то есть ее фундаментальное переосмысление и радикальное перепроектирование, в настоящее время не является *достижимой* задачей. Межгалактические путешествия лежат за пределами возможностей современных технологий. Речь идет о том, что мы, в принципе, способны использовать имеющиеся у нас уже сейчас ресурсы для создания дополнительных возможностей, тем самым обеспечив *косвенную* достижимость этой цели. Безусловная правда, что человечество пока не является синглтоном и что нет уверенности в отсутствии некой разумной силы, с противодействием которой мы столкнемся в процессе реинжиниринга Вселенной. Однако чтобы преодолеть порог устойчивости благоразумного синглтона, достаточно обладать такими возможностями, которые сделали бы достижимой задачу колонизации и реинжиниринга Вселенной благоразумным синглтоном, не встречающим противодействия другой разумной силы.

[29] Иногда полезно говорить о двух ИИ, обладающих равной сверхмощью. В более широком смысле слова можно было бы считать, что сверхмощь в какой-то области предполагает сравнение возможностей действующей силы и человеческой цивилизации, но за исключением другого ИИ.

Глава 7. Намерения сверхразума

[1] Конечно, все сказанное не исключает, что могут быть различия — довольно незначительные под микроскопом, но имеющие большое значение с функциональной точки зрения.

[2] См.: [Yudkowsky 2008 a, p. 310].

[3] Дэвид Юм, шотландский философ эпохи Просвещения, считал, что одной убежденности (например, в том, как правильно поступать) недостаточно для мотивации поступка — требуется еще желание. Казалось бы, это только снимает одно возражение, которое возможно выдвинуть против нашего тезиса: достаточно развитый интеллект непременно обретет определенные убеждения, что обязательно приведет к появлению определенных мотиваций. Однако, несмотря на то что тезис об ортогональности внешне подкрепляется юмовской мотивационной концепцией, он не предполагает ее в качестве необходимого условия. В частности, трудно отрицать, что иногда лишь голые убеждения служат вполне убедительной мотивацией какого-либо действия. Почему бы не предположить, например, что агент, обладающий высоким интеллектом, может быть настроен на любой план действий, если на то у него имеются довольно сильные желания. Второй случай, когда принцип ортогональности может быть истинным даже при условии ошибочности мотивационной концепции Юма, — это если формирование подобных убеждений у обладающего произвольно высоким интеллектом агента само по себе не мотивирует его на соответствующие поступки. Третий случай, когда принцип ортогональности может быть истинным даже при условии ошибочности мотивационной концепции Юма, — это если возможно создать агента (или проще: запустить «процесс оптимизации») с произвольно высоким интеллектом, но с настолько чужеродным устройством, что у него вообще не будет прямых функциональных аналогов таким человеческим понятиям, как «убеждение» и «желание». (Недавние попытки защитить концепцию Юма можно найти в работах: [Smith 1987; Lewis 1988; Sinhababu 2009].)

[4] Например, Дерек Парфит считает, что некоторые базовые предпочтения могут быть иррациональными, как, скажем, у нормального, в принципе, агента, но с синдромом «безразличия к следующему вторнику», то есть гедониста, которого очень заботит качество его будущего опыта, за одним исключением. Этим исключением является его безразличие к событиям будущего вторника. Вообще то, что происходит по вторникам, ему не безразлично. Ему безразличны лишь страдания и удовольствие, которые его ждут в следующий вторник… Это безразличие — неоспоримый факт. И поэтому, составляя планы на будущее, он предпочтет перспективу огромных страданий по вторникам умеренным страданиям в любой другой день [Parfit 1986, p. 123–124]; см. также: [Parfit 2011].

Для своих целей мы не будем задерживаться на выяснении, прав ли Парфит, считая такого агента рациональным, если мы примем, что в инструментальном смысле описанное в этом примере поведение агента не обязательно неразумно. Агент Парфита может быть безукоризненно рациональным в инструментальном смысле, а следовательно, иметь большой интеллект, даже если ему недостает восприимчивости к «объективной причине», которой должен был бы обладать полностью рациональный агент. Следовательно, такие примеры не опровергают тезис об ортогональности.

[5] Даже наличие объективных этических норм, которые способен понять полностью рациональный агент, и даже если эти этические нормы обладают внутренней мотивирующей силой (в результате чего все, кто их понял, непременно будут поступать в соответствии с ними), не опровергает тезис об ортогональности. Он остается верным, если агент непоколебимо

рационален в *инструментальном* смысле при отсутствии некоторых других составляющих рациональности или качеств, необходимых для полного понимания данных этических норм. (Агент также может быть чрезвычайно интеллектуальным, даже сверхинтеллектуальным, и не обладать полной инструментальной рациональностью во всех областях.)

[6] Более подробно тезис об ортогональности рассматривается в работах: [Bostrom 2012; Armstrong 2013].

[7] См.: [Sandberg, Bostrom 2008].

[8] На эту тему есть две основополагающие работы Стивена Омохундро, считающего, что всем прогрессивным системам ИИ, скорее всего, будет присущ набор «базисных установок», под которыми он понимает «исходные склонности, влияющие до тех пор, пока не будет осуществлено явное противодействие» [Omohundro 2007; Omohundro 2008]. У термина *установка ИИ* есть несомненное преимущество — он короткий, яркий и узнаваемый. Но у него есть недостаток: он наводит на мысль, что инструментальные цели ИИ — чем, по сути, являются его базисные установки — воздействуют на процесс принятия им решений ровно таким же образом, как человеческие психологические установки влияют на процесс принятия решений людьми, когда мы — за счет своего рода феноменологической удавки, наброшенной на собственное я, — силой воли преодолеваем свои природные склонности. Подобная аналогия неплодотворна. Ведь никто из нас никогда в жизни не произнесет: «У меня есть *установка* своевременно заполнять налоговую декларацию», — даже если само действие является разумной инструментальной целью любого цивилизованного члена современного общества (реализация именно этой цели предотвращает неприятности, способные помешать человеку воплотить в жизнь многие его конечные цели). Некоторых другие заключения Омохундро тоже расходятся с нашей трактовкой, хотя в вопросе основной идеи мы с ним солидарны. См. также: [Chalmers 2010; Omohundro 2012].

[9] См.: [Chislenko 1997].

[10] См. также: [Shulman 2010 b].

[11] Под влиянием перемен в онтологических взглядах агент может менять и общее *целевое представление*, чтобы придать равновесие своей мировоззренческой позиции, см.: [De Blanc 2011].

Еще одним фактором может стать важность принятия того или иного решения под воздействием бесспорности внешних обстоятельств: агент вынужден предпринимать те или иные действия, в том числе меняя конечные цели, чтобы увеличить *очевидность принятия решений*. Например, агент, следующий принципам теории выбора, может верить, что во вселенной существуют другие агенты, в чем-то похожие на него, и что от них можно ожидать поведения, сходного с его собственным. Вследствие чего агент выберет по отношению к гипотетическим вселенским агентам альтруистическую конечную цель, надеясь, что и они поведут себя так же. Впрочем, аналогичного результата можно добиться, не меняя конечных целей, а просто действуя таким образом, *будто бы* они были изменены.

[12] Формированию адаптивных предпочтений посвящено множество работ в области психологии; см., например: [Forgas et al. 2010].

[13] В формальных моделях ценность информации измеряется разностью между предполагаемым средним значением, полученным в результате оптимальных решений, принятых с учетом этой информации, и предполагаемым средним значением, полученным в результате оптимальных решений, принятых без ее учета; см., например: [Russell, Norvig 2010]. Отсюда следует, что значимость информации не может быть величиной отрицательной и что информация, которой, насколько вам известно, вы никогда не воспользуетесь для принятия решений, имеет для вас нулевую ценность. Однако в таких формальных моделях допускаются некоторые упрощения, правда, не очень значимые для реального мира: во-первых, знание не несет в себе результативной ценности (то есть оно имеет лишь инструментальную ценность, но само по себе ничего не значит); во-вторых, знания одних агентов недоступны другим агентам.

[14] См., например: [Hájek 2009].

[15] Примером такой стратегии может быть поведение асцидии: она свободно плавает будучи личинкой, но, повзрослев, находит подходящий камень, к которому прикрепляется навсегда. Обретя свое место, асцидия перестает нуждаться в сложной системе обработки информации и начинает поедать собственный мозг (в частности, головной ганглий). Такие же процессы происходят — что мы можем наблюдать — и с некоторыми учеными, когда они, став штатными профессорами, отдают себя в полное владение университетам.

[16] См.: [Bostrom 2012].

[17] См.: [Bostrom 2006 c].

[18] Рассмотрим проблему с прямо противоположной стороны и расследуем вероятные причины, по которым сверхмощный и сверхразумный синглтон *не стал бы* развивать свои технологические возможности: 1) синглтон прогнозирует, что не воспользуется этими возможностями; 2) затраты на разработку слишком высоки по сравнению с предполагаемой пользой (например, новейшая технология никогда не пригодится для достижения одной из целей синглтона или ставка дисконтирования столь высока, что делает практически нереальными инвестиции); 3) конечная цель синглтона требует воздерживаться от развития конкретных технологических направлений; 4) если синглтон не уверен, что сможет сохранить стабильность, то может предпочесть воздержаться от развития технологий, которые могли бы угрожать его внутреннему состоянию равновесия или привести к более тяжелым последствиям в случае его распада (например, мировое правительство не станет содействовать развитию технологий, которые могут использоваться в случае массовых акций протеста, даже если у них есть и полезные области применения, или технологий, облегчающих производство оружия массового поражения, способное вырваться из-под контроля в случае падения самого правительства); 5) у синглтона есть ранее взятые обязательства не разрабатывать определенные виды технологий; эти обязательства могут по-прежнему действовать, несмотря на их невыгодность в новых условиях. (Следует отметить, что некоторые *существующие сегодня* причины развития технологий *не будут* иметь отношения к будущему синглтону: я имею в виду причины, связанные с гонкой вооружений.)

[19] Предположим, что агент будет добывать в будущем ресурсы по экспоненциальной ставке дисконтирования, но обеспеченность ресурсами происходит в соответствии с полиномиальной моделью — из-за ограничений, которые накладывает скорость света. Будет ли это означать,

что в определенный момент агент сочтет дальнейшую экспансию бессмысленной? Нет, поскольку *так же будут вести себя и затраты на приобретение ресурсов*, хотя текущая стоимость ресурсов, которые будут получены в будущем, станет тем быстрее стремиться по асимптоте к нулю, чем дальше в будущее мы будем заглядывать. Текущая стоимость запуска еще одного зонда фон Неймана через 100 млн лет (возможно, за счет ресурсов, приобретенных ранее) будет практически обнулена за счет того же коэффициента дисконтирования, который практически обнулит текущую стоимость будущих ресурсов, добытых этим зондом.

[20] Хотя объем космического пространства, охваченный зондами, всегда будет представлять собой почти идеальную сферу, расширяющуюся пропорционально квадрату времени, прошедшему с момента запуска первого зонда ($\sim t^2$), количество содержащихся в этом объеме ресурсов будет расти с гораздо меньшей скоростью, поскольку они распределены в нем неравномерно. Поначалу, во время колонизации родной планеты, скорость будет равной $\sim t^2$; потом она достигнет пика за счет колонизации ближайших планет и звездных систем; затем, когда с объемом в форме диска, в котором содержится Млечный Путь, будет покончено, скорость роста упадет примерно до t; после снова достигнет пика во время колонизации близлежащих галактик; вслед за этим скорость роста снова упадет до $\sim t^2$, когда расширение достигнет масштабов, в которых распределение галактик можно считать однородным; далее еще один пик и плавный спад до $\sim t^2$ в процессе колонизации галактического суперкластера; и наконец, скорость роста начнет снижаться, пока не упадет до нуля в момент, когда расширение Вселенной сделает дальнейшую колонизацию невозможной.

[21] В этом контексте особенно большое значение приобретает гипотеза симуляции. Сверхразумный агент может присвоить высокую вероятность гипотезе, в соответствии с которой он существует в компьютерной имитационной модели, а его перцептивная последовательность генерируется другим сверхразумом, в результате чего могут возникнуть различные конвергентные инструментальные причины в зависимости от представлений агента о том, в какого типа имитационной модели он может скорее всего находиться; см.: [Bostrom 2003 a].

[22] Открытие фундаментальных законов физики и других фундаментальных фактов является конвергентной инструментальной целью. Поместим ее в категорию «усовершенствование когнитивных способностей», хотя также она может находиться в категории «технологическое совершенство» (поскольку новые физические явления означают создание инновационных технологий).

Глава 8. Катастрофа неизбежна?

[1] Есть другие сценарии экзистенциального риска: человечество выживет, но будет пребывать в состоянии, далеком от оптимального; человечество выживет, но безвозвратно утратит бóльшую часть своего потенциала, без которого дальнейший прогресс будет невозможен. Помимо этого, экзистенциальные риски могут быть связаны с самим процессом взрывного развития искусственного интеллекта: высока вероятность вражды между государствами, борющимися за лидерство в создании сверхразума.

[2] Особенно уязвимым будет момент, когда ИИ впервые осознает необходимость скрывать свои намерения (это явление можно назвать *рождением обмана*). В самом начале процесса осознания

ИИ еще не будет прятать свои мысли от разработчиков. Но как только поймет это окончательно, то сразу — чтобы иметь возможность продолжать работу над планом по реализации своей долгосрочной стратегии — запустит некие внутренние механизмы маскировки, скрывая в том числе и сам факт осознания (возможно, он будет использовать одни, невинно выглядящие, процессы для прикрытия других, гораздо более сложных).

[3] Даже хакеры-люди способны писать небольшие и внешне невинные программы, способные делать совершенно неожиданные вещи. (Примеры можно найти, просмотрев список победителей Международного конкурса на самый запутанный код на языке Си.)

[4] Некоторые механизмы контроля над ИИ кажутся вполне надежными в каком-то определенном контексте, но если ситуация изменится, те же механизмы могут привести к катастрофическому отказу — допустимость такого поворота событий также подчеркивал Элиезер Юдковский [Yudkowsky 2008 a].

[5] Кажется, впервые термин *самостимуляция* использовал писатель-фантаст Ларри Нивен [Niven 1973], но восходит он к реальным экспериментам по прямой электростимуляции «зон вознаграждения» мозга, которые проводили на животных Джеймс Олдс и Питер Милнер, см.: [Olds, Milner 1954; Oshima, Katayama 2010]. См. также: [Ring, Orseau 2011].

[6] См. также: [Bostrom 1997].

[7] Возможно, удастся настроить механизм обучения с подкреплением таким образом, что во время процесса самостимуляции будет происходить безопасное отключение системы, а не отказ по типу инфраструктурной избыточности. Проблема в другом: ситуация, по самым неизвестным причинам, все равно может выйти из-под контроля.

[8] Вариант, предложенный Марвином Мински; см.: [Russell, Norvig 2010, p. 1039].

[9] Для обсуждения такой темы (в отличие от многих других тем книги) большое значение имеет вопрос, какие типы цифрового разума будут наделены сознанием, в смысле будут ли у них квалиа, то есть опыт субъективных переживаний и ощущений. Нерешенным остается вопрос, как в различных ситуациях поведут себя эти антропоморфные сущности; мы в принципе не можем оценить этого, не сделав моделирования их мозга на таком уровне детализации, который мог бы привести к появлению у них сознания. Неясно также, можно ли создать пригодные к практическому использованию в процессе создания ИИ алгоритмы, например методы обучения с подкреплением, в результате работы которых у него сформируются квалиа. Пусть мы придем к выводу, что вероятность появления сознания у таких подпрограмм довольно низка, но их количество может оказаться настолько решающим, что мы не имеем права допустить даже самого незначительного риска их страданий. По нашей шкале ценностей такая опасность должна иметь серьезное значение. См. также: [Metzinger 2003, ch. 8].

[10] См.: [Bostrom 2002 a; 2003 a; Elga 2004].

Глава 9. Проблемы контроля

[1] См., например: [Laffont, Martimort 2002].

[2] Предположим, большинство избирателей мечтают, чтобы в их стране был создан сверхразум. Они голосуют за кандидата, который обещает выполнить их пожелание, но могут ли они быть

уверены, что он, придя к власти, выполнит обещания, данные в ходе предвыборной кампании, и будет реализовывать проект в соответствии с предпочтениями людей? Допустим, он сдержит слово и распорядится, чтобы правительство привлекло ученых и бизнесменов к выполнению этой задачи; но снова возникает агентская проблема: у бюрократов может быть своя точка зрения на то, что нужно делать, и проект будет реализован в соответствии с буквой, но не духом данных им инструкций. И даже если правительство честно выполнит свою часть работы, у привлеченных им исполнителей может быть собственное видение проекта. То есть проблема возникает на многих уровнях. Например, директор лаборатории, участвующей в проекте, может не спать ночами из страха, что какой-нибудь разработчик внесет несанкционированные изменения в программу, он уже представляет, как поздно ночью прокрадывается в свой кабинет профессор И. З. Менник, входит в систему и частично переписывает код, меняя конечные цели ИИ. И там, где было «служить человечеству», появляется «служить профессору И. З. Меннику».

[3] Впрочем, даже во время развития ИИ возможно использовать бихевиористский подход — правда, лишь в качестве дополнительного средства тестирования в более обширной системе мер предосторожности. Если ИИ на стадии разработки начнет вести себя ненадлежащим образом, что-то обязательно пойдет наперекосяк в его системе — однако прошу обратить особое внимание, что при обратном эффекте это правило не справедливо.

[4] Приведу классический пример, относящийся к уже далекому 1975 году, когда Стивен Домпье обнаружил, что его Altair 8800 (один из первых персональных компьютеров, примитивный и не экранированный) вызывает помехи в стоящем рядом радиоприемнике. Тогда он написал программу, управляющую электромагнитными разрядами, и с помощью приемника проиграл на своем Altair мелодию битловской Fool on the Hill («Дурак на холме»); см.: [Driscoll 2012]. Присутствовавший на демонстрации молодой Билл Гейтс был восхищен и заинтригован увиденным [Gates 1975]. Сегодня есть планы создания микропроцессоров со встроенным Wi-Fi-передатчиком [Greene 2012].

[5] Это далеко не простой вопрос: брать под сомнение то, в чем твердо уверен. Но если мы не научимся это делать, то потеряем шанс заблаговременно принять соответствующие меры — и тогда прощай заветная мечта об освоении космоса. Пожалуй, можно было бы поговорить в пользу такого принципа: если некто в прошлом каждый раз (N случаев) был убежден в абсолютной защищенности своей системы, но всякий раз оказывалось, что он ошибался, то в очередном случае рассуждая об ее безопасности, он не вправе присваивать вероятность, превышающую $1/(N + 1)$.

[6] В одном неофициальном эксперименте в роли ИИ выступил человек — естественно, обладающий незаурядным интеллектом. В роль охранника, которого каждый раз исполняли разные люди, входило ни при каком случае не выпускать «искусственного интелекта» из комнаты. «Искусственному интеллекту» дали всего два часа на то, чтобы убедить своего сторожа отпустить его на свободу, причем разрешили вступать в общение с ним лишь посредством текстовых сообщений. В трех случаях из пяти «искусственному интеллекту» удавалось сбегать; см.: [Yudkowsky 2002]. Если смог человек, то сможет и сверхразум. (Обратное, конечно, неверно. Даже если перед сверхразумом будет стоять более трудная задача — возможно, его

охранники будут сильнее мотивированы, чем люди в ходе упомянутого эксперимента, — сверхразум справится там, где потерпит поражение человек.)

[7] Однако и в этом случае не стоит переоценивать безопасность системы. Психические образы легко заменяются на визуальные с помощью графической информации. Более того, вспомним, какое воздействие на человека оказывают книги — притом что книга, насколько мы знаем, не вступает в диалог со своим читателем.

[8] См. также: [Chalmers 2010]. Однако было бы неправильным считать, что систему, за действиями которой нельзя наблюдать со стороны, *невозможно* использовать. Во-первых, происходящее внутри нее может быть определено как результирующее значение. Во-вторых, не исключено, что кто-то, либо руководствуясь собственными соображениями, либо взяв на себя такое обязательство, создает именно замкнутую систему, чтобы иметь приоритетное право на то, что происходит внутри нее. Само существование определенных классов замкнутых систем, содержащих шаблон наблюдателя, может вызвать у некоторых внешних наблюдателей чувство сомнения в антропном принципе, что, естественно, повлияет на их дальнейшие действия.

[9] Может возникнуть вопрос, почему социальная интеграция расценивается как один из методов контроля над возможностями? Наверное, правильнее было бы классифицировать ее как один из методов выбора мотивации? На том основании, что целью социальной интеграции является стимулирование нужного нам поведения системы. Довольно скоро мы начнем подробное обсуждение проблемы выбора мотивации, но сейчас, в поисках ответа на поставленный вопрос, скажем, что в этой конкретной ситуации метод контроля мы рассматриваем в свете метода выбора мотивации, имеющего прямое отношение к конечным целям системы, выбор или уточнение которых не имеет инструментального значения, поскольку они значимы сами по себе. Так как социальная интеграция не является конечной целью системы, она не относится к методам выбора мотивации. Скорее, социальная интеграция призвана ограничить огромные возможности системы, то есть не дать ей реализовать определенные сценарии, в которых она нарушает правила, не подвергаясь при этом соответствующему наказанию (возмездие и потеря выгоды от сотрудничества). Надежда лишь на одно: в результате резкого сужения потенциала, большую часть которого она не в состоянии раскрыть, система поймет, что наиболее значимой из оставшихся в ее распоряжении возможностей добиться конечных целей остается сотрудничество с человеческим обществом.

[10] Этот подход может оказаться более удачным, если считать, что у имитационных моделей будет мотивация, аналогичная человеческой.

[11] Идею я позаимствовал у Карла Шульмана.

[12] Создание шифра, способного противостоять сверхразумному дешифратору, — задача крайней сложности. Например, следы случайной последовательности могут сохраниться в мозгу какого-то наблюдателя или в микроструктуре генератора случайных чисел, откуда их сможет извлечь сверхразум. Помимо этого, если используются псевдослучайные числа, он способен разгадать или отыскать принцип их генерации. Более того, сверхразум может построить большие квантовые компьютеры или даже открыть новые физические явления, которые позволят создать совсем новые типы компьютеров.

[13] ИИ мог бы перепрограммировать себя так, чтобы *поверить* в то, что получает наградные знаки, но это не будет прямым стимулированием, если он сконструирован так, чтобы стремиться получать именно знаки (а не формулировать те или иные убеждения насчет них).

[14] См.: [Bostrom 2003 a]; см. также: [Elga 2004].

[15] См.: [Shulman 2010 a].

[16] Предположительно, доступная нам реальность базового уровня требует большего количества вычислительных ресурсов, чем имитационный мир, поскольку все вычислительные процессы, происходящие в имитационной модели, делаются на компьютерах, на которых она запущена. Реальность базового уровня также содержит множество других физических ресурсов, к которым трудно получить доступ имитационным агентам — агентам, существующим лишь милостью сверхмощных и сверхразумных имитаций-принципалов; причем у последних могут быть свои виды на эти ресурсы. (Конечно, дедуктивный вывод здесь не идеально строг — вполне допустимо, что имитационные миры обладают таким количеством ресурсов, что сверхмощные и сверхразумные имитации-принципалы в среднем имеют доступ к большему количеству ресурсов, чем сверхразумы, обитающие в реальном мире, даже если каждая неимитационная цивилизация, запускающая имитационные модели, обладает бóльшими ресурсами, чем все ее имитации, вместе взятые.)

[17] Есть и другие эзотерические соображения, имеющие отношение к этой теме и способные привести к таким последствиям, которые еще не проанализированы в должной мере. В конечном счете они могут оказаться критически важными для создания всеобъемлющего подхода к подготовке взрывного развития искусственного интеллекта. Однако маловероятно, что мы преуспеем в выяснении степени практической значимости этих эзотерических соображений, не добившись вначале определенного прогресса в анализе более приземленных вещей, которым посвящена бóльшая часть этой книги.

[18] См., например: [Quine, Ullian 1978].

[19] О них ИИ может узнать, проведя анализ характеристики производительности различных базовых вычислительных компонентов, включая размер и емкость шин данных, время доступа к разным областям памяти и т. д.

[20] Размышляя об априорной вероятности создания миров, наш сверхразум не стал бы первооткрывателем: еще в 1960-е гг. Рэй Соломонов ввел концепцию алгоритмической вероятности, которая легла в основу его, соломоновской, теории индуктивного вывода — теории предсказаний, сделанных в результате наблюдений; см.: [Li, Vitanyi 2008].

[21] Уже в следующий момент — *сразу после* рождения обмана — ИИ может решить стереть следы своих мятежных мыслей. Поэтому очень важно, чтобы «растяжки» работали постоянно. Кроме того, было бы полезно иметь «черный ящик» вроде тех, которые используются в самолетах, где могли бы храниться следы всех действий ИИ (включая точное время введения информации с клавиатуры программистами), чтобы после автоматического отключения системы их можно было проследить и проанализировать. Храниться информация может на устройстве с возможностью лишь однократной записи и многократного чтения.

[22] См.: [Asimov 1942]*. К этим трем позднее был добавлен «нулевой» закон: 0) робот не может причинить вред человечеству или своим бездействием допустить, чтобы человечеству был причинен вред [Asimov 1985]**.

[23] См.: [Gunn 1982].

[24] См.: [Russell 1986, p. 161f].

[25] У философов порою уходила вся жизнь на то, чтобы ясно сформулировать положения деонтологической этики***, тем более что и время, и мир никогда не стоят на месте, постоянно возникают новые события и обстоятельства, требующие пересмотра концепций, а значит, и обновленных изложений. Эта аналогия невольно приходит на ум в связи с нашей темой. Например, со второй половины прошлого века набирает силу новое междисциплинарное направление, названное «мысленный эксперимент», и всем, кто занимается этическими учениями или теорией познания, вновь пришлось переоценивать стандартные представления. С позиций деонтологии речь прежде всего идет, конечно, о так называемой проблеме вагонетки — этическом мысленном эксперименте, открывшем современным исследователям многие нравственные установки, присутствующие в сознании современных людей. Возьмем хотя бы наши едва уловимые представления, связанные с различиями таких понятий, как последствия действия и бездействия, последствия намеренные и непреднамеренные; см.: [Kamm 2007].

[26] См.: [Armstrong 2010].

[27] Здравый смысл подсказывает, что если планируешь использовать многочисленные предохранительные устройства для изоляции ИИ, а также другие методы контроля и мотивации, то самым разумным было бы действовать так, *будто* каждое устройство и каждый метод является тем единственным, который *следовало бы* применять *именно* в данном случае. Говоря языком программистов, когда ставишь одно дырявое ведро в другое дырявое ведро — вода все равно вытекает наружу.

[28] Вариант той же идеи: создать ИИ, мотивированный действовать в соответствии со своим представлением, каким мог бы быть неявно заданный стандарт. В данной ситуации конечная цель ИИ — всегда действовать в соответствии с неявно заданным стандартом, а задача определения того, каким он может быть, выполняется в рамках достижения инструментальной цели.

Глава 10. Оракулы, джинны, монархи и инструменты

[1] Я сознательно остановился на антропоморфных названиях, но им не следует придавать привычный нам смысл. Это всего лишь ярлыки для обозначения внешних отличий различных интеллектуальных систем, которые можно создать.

---

* Рассказ Айзека Азимова Runaround известен русскоязычному читателю с 1963 года как «Хоровод»; в сборнике «Я, робот» (М.: Центрполиграф, 2003) опубликован под названием «Вокруг да около».

** *Айзек Азимов*. Роботы и Империя. М.: Эксмо, 2003.

*** *Деонтологическая этика*, или деонтология (от др.-греч. δέον — «должное») — учение о проблемах морали и нравственности; раздел этики.

[2] При вопросе о результатах следующих выборов вряд ли было бы удобно пользоваться ответом, содержащим полный перечень всех прогнозных вариантов, касающихся как людей, так и ситуаций.

[3] Привязанных к определенному набору команд и компьютеру.

[4] См.: [Kuhn 1962; De Blanc 2011].

[5] Применить метод консенсуса к ИИ-джиннам и ИИ-монархам было бы сложнее, поскольку для достижения одной и той же цели можно использовать различные и при этом почти одинаково эффективные последовательности базовых действий (например, направление определенных наборов электрических сигналов к исполнительным механизмам системы); разные агенты могут выбирать несколько отличные действия, и тогда консенсус окажется невозможным. В отличие от этого, у вопросов, сформулированных нужным образом, не так много вариантов подходящих ответов (например, «да» и «нет»). (О концепции фокусной, или фокальной, точки, или точки Шеллинга, см.: [Schelling 1980].)

[6] В некотором отношении поведение мировой экономики в чем-то напоминает поведение джинна, только джинна какого-то вялого и тщедушного, хотя и чрезвычайно алчного. Мировая экономика, которая может возникнуть в будущем, будет напоминать джинна, обладающего коллективным сверхразумом.

Правда, есть один момент, когда современная экономика *не напоминает* джинна: я могу приказать ей (конечно, за оплату ее услуг) обеспечить мне пиццу с доставкой на дом, но не в моих силах дать ей команду обеспечить мир во всем мире. Причем причина не в том, что экономика не обладает достаточной властью, просто она недостаточно скоординированна. В этом смысле мировая экономика скорее похожа не на одного джинна или иного единого агента, а на *партию* джиннов, обслуживающую разных и конкурирующих между собой владельцев.

Вряд ли окажется перспективным путь, если мы начнем наделять каждого джинна бóльшим могуществом — это не сделает нашу экономику более сильной и не обеспечит мир на планете. Чтобы экономика могла функционировать как сверхразумный джинн, нужно не только повышать ее производительность, то есть добиваться при минимальных затратах максимального выпуска товаров и услуг (в том числе тех, которым требуются принципиально новые технологии), но и приспосабливать ее к тому, чтобы она была заинтересована лучше решать проблемы общемировой координации.

[7] Если джинн почему-то будет неспособен отказаться от выполнения команды — и перепрограммировать себя так, чтобы избавиться от этого слабого места, — он может действовать так, чтобы исключить возможность появления новых команд.

[8] Для джинна или монарха, находящихся в поисках нужного варианта ответа, может оказаться полезным даже оракул, отвечающий в духе «да» и «нет». Помощь оракула удобно использовать для написания исходного кода для джинна или монарха — нужно лишь задать ему ряд вопросов, приблизительно такого плана: «В бинарной версии кода первого ИИ, который, по твоему мнению, мог бы стать джинном, *n*-й символ — это ноль?»

[9] Можно разработать более сложного оракула (или джинна), принимающего вопросы (или команды) только от уполномоченных на это лиц. Но где гарантия, что эти люди не окажутся коррумпированными или не станут жертвами шантажа со стороны ИИ?

[10] Концепция «вуаль неведения», или «вуаль невежества», принадлежит одному из самых известных политических философов XX века Джону Ролзу; согласно ей при формировании принципов справедливого распределения нужно абстрагироваться от возможных последствий для своего личного благосостояния. Другими словами, когда мы размышляем над устройством такого понятия, как общественный договор, нам следует представить, будто мы находимся под «вуалью неведения», не позволяющей нам узнать, ни каким человеком мы станем, ни какую социальную роль будем выполнять. Идея в том, чтобы мы думали построить такое общество, которое могло бы быть наиболее справедливым и желательным безотносительно наших эгоистических интересов и когнитивных искажений. В противном случае мы стали бы действовать в пользу собственной выгоды и предпочли бы такой социальный порядок, при котором имели бы незаслуженные привилегии [Rawls 1971].

[11] См.: [Karnofsky 2012].

[12] Возможным исключением может быть ПО, непосредственно замкнутое на достаточно мощные исполнительные механизмы, скажем, системы раннего предупреждения о ракетном нападении, напрямую соединенные с ядерными боеголовками или передающие информацию офицерам, уполномоченным на нанесение ядерного удара. Ошибки в его работе способны привести к абсолютно рискованным ситуациям. В истории человечества это происходило минимум дважды. Первый случай: 9 ноября 1979 года в результате компьютерного сбоя Объединенное командование воздушно-космической обороны Североамериканского континента получило ложный сигнал о начале полномасштабного нападения СССР на США. Немедленно началась подготовка ответного удара, но данные с радарных систем раннего предупреждения показали, что ни одной ракеты со стороны СССР запущено не было [McLean, Stewart 1979]. Второй случай: 26 сентября 1983 года ошибочно сработала «Око» — советская спутниковая система обнаружения стартов межконтинентальных баллистических ракет с континентальной части США, — сообщив о ракетном ударе со стороны Соединенных Штатов. Оперативный дежурный командного пункта подполковник Станислав Петров правильно определил, что эта тревога ложная, — практически он один предотвратил ядерную войну [Lebedev 2004]. Вряд ли она привела бы к исчезновению человечества, даже если был бы задействован весь ядерный потенциал, имевшийся у всех стран на пике холодной войны, но, безусловно, вызвала бы неисчислимые смерти и страдания и крах современной цивилизации [Gaddis 1982; Parrington 1997]. Что угрожает нам в будущем? Может быть накоплен еще больший ядерный потенциал, изобретено более мощное смертоносное оружие, наши модели ядерного Армагеддона (в частности, оценки суровости ядерной зимы) могут оказаться несостоятельными.

[13] Этот подход можно отнести к категории метода точной спецификации, основанного на системе четко прописанных правил (см. главу 9).

[14] Ничего не изменится и в том случае, если критерий успеха будет определять лишь *меру* успешности решения, а не его точное определение.

[15] Апологеты ИИ-оракула заявили бы, что у его пользователя по крайней мере есть возможность заметить изъян в предлагаемом решении — что он не соответствует намерениям пользователя, хотя и отвечает формально заданному критерию успеха. Вероятность обнаружения ошибки на этом этапе зависит от множества факторов, включая то, насколько понятны для человека результаты работы оракула и насколько доброжелательно он подходит к отбору тех черт потенциального сценария, которые представляет вниманию пользователя.

Можно не полагаться на ответы оракула, а попытаться создать отдельный инструмент, который мог бы инспектировать предложения ИИ и сообщать нам, что произойдет, если мы с ними согласимся. Но чтобы обеспечить это в полной мере, потребуется еще один сверхразум, чьему мнению мы должны будем доверять, то есть проблема надежности по-прежнему не будет решена. Можно также попробовать повысить безопасность за счет использования множества оракулов, перепроверяющих друг друга, но это не защитит нас в том случае, если все оракулы совершат одну и ту же ошибку — что может произойти, например, в ситуации, когда все они пользуются одним и тем же формальным определением того, что считать удовлетворительным решением.

[16] См.: [Bird, Layzell 2002; Thompson 1997]; см. также: [Yaeger 1994; p. 13–14].

[17] См.: [Williams 1966].

[18] См.: [Leigh 2010].

[19] Пример взят из работы Элиезера Юдковского, см.: [Yudkowsky 2011].

[20] См.: [Wade 1976]. Проводились также компьютерные эксперименты по симулированию некоторых аспектов биологической эволюции, и иногда тоже с очень неожиданными результатами (см., например: [Yaeger 1994]).

[21] При наличии довольно большой — ограниченной, но физически невозможной — вычислительной мощности *было бы* возможно получить универсальный сверхразум даже на базе имеющихся сейчас алгоритмов. (Например, AIX — UNIX-подобная операционная система IBM, см.: [Hutter 2001].) Но для достижения нужного уровня вычислительной мощности будет недостаточно соблюдения закона Мура даже в течение еще ста лет.

Глава 11. Сценарии многополярного мира

[1] Совсем не обязательно, что это наиболее вероятный или желательный сценарий, поскольку он — один из наиболее простых с точки зрения анализа средствами стандартной экономической теории и потому представляет собой удобную стартовую точку для нашего обсуждения.

[2] См.: [American Horse Council, 2005]; см. также: [Salem, Rowan 2001].

[3] См.: [Acemoglu 2003; Mankiw 2009; Zuleta 2008].

[4] См.: [Fredriksen 2012, p. 8; Salverda et al. 2009, p. 133].

[5] Важно, чтобы часть капитала была инвестирована в активы, которые растут вместе с рынком. Повысить шансы, что волна не пройдет мимо, может диверсификация портфеля активов, например приобретение паев индексного фонда.

[6] Пенсионные системы многих европейских стран являются *распределительными*, то есть пенсии выплачиваются скорее за счет текущих отчислений и налогов работающих граждан, чем накопленных сбережений. Такая система не будет автоматически удовлетворять нашим требованиям, поскольку в случае внезапного всплеска безработицы доходы, которые распределяются в виде пенсий, могут оскудеть. Однако у правительства может быть возможность восполнить дефицит из каких-то других источников.

[7] См.: [American Horse Council, 2005].

[8] Чтобы 7 млрд человек получали годовую пенсию в размере 90 000 долларов, потребуется 630 трлн долларов в год, что в десять раз превышает величину современного мирового ВВП. По имеющимся данным, за последние 100 лет он вырос в 19 раз — с 2 млрд долларов в 1900 году до 37 млрд долларов в 2000-м (в международных долларах 1900 года), см.: [Maddison 2007]. Поэтому если темпы роста, которые мы видели в прошлом веке, сохранятся на протяжении следующих двухсот лет, а население планеты не вырастет, обеспечение каждого годовой пенсией в 90 000 долларов будет стоить всего 3% мирового ВВП. Благодаря взрывному развитию искусственного интеллекта такой рост может случиться гораздо раньше. См. также: [Hanson 1998 a; 1998 b; 2008].

[9] За последние 70 тысяч лет могло вырасти в миллион раз; см.: [Kremer 1993; Huff et al. 2010].

[10] См.: [Cochran, Harpending 2009]; см. также: [Clark 2007] и [Allen 2008] — с критикой.

[11] См.: [Kremer 1993].

[12] См.: [Basten et al. 2013]. Также возможны сценарии и продолжения роста; но надо сказать, что неопределенность резко вырастает при прогнозе более чем на пару поколений.

[13] В среднем в мире минимальный коэффициент фертильности для замещения в 2013 году равнялся 2,33 ребенка на одну женщину. Это число учитывает факт, что два ребенка нужно для того, чтобы заменить родителей, плюс еще «треть» для компенсации: 1) более высокая вероятность рождения мальчиков; 2) ранняя смерть до достижения фертильного возраста; для развитых стран этот показатель ниже, около 2,1 ребенка, из-за более низкой смертности; см.: [Espenshade et al. 2003, Introduction, table 1, p. 580]. Не будь иммиграции, численность населения большинства развитых стран снижалась бы. Приведу несколько примеров стран с фертильностью, недотягивающей до уровня замещения: Сингапур — 0,79 (самая низкая в мире), Япония — 1,39, Китай — 1,55, ЕС — 1,58, Россия — 1,61, Бразилия — 1,81, Иран — 1,86, Великобритания — 1,90. Даже население США медленно уменьшалось бы с их коэффициентом фертильности 2,05; см.: [CIA, 2013].

[14] Этот час может пробить лишь через миллиарды лет.

[15] Карл Шульман отмечает, что если люди рассчитывают сохранить свой образ жизни в условиях цифровой экономики, им нужно добиться, чтобы политическое устройство цифрового мира обеспечивало бы защиту их интересов на протяжении очень долгого времени [Shulman 2012]. Например, если события в цифровом мире разворачиваются в тысячу раз быстрее, чем вне его, то человеку придется полагаться на политическую стабильность в нем на протяжении 50 000 лет постоянных внутренних изменений. Но если цифровой политический мир будет хоть в чем-то похож на наш, за эти тысячи лет в нем произойдет огромное количество

революций, войн и бунтов, способных сделать жизнь обычных людей «снаружи» невыносимой. Даже при минимальных шансах (не превышающих 0,01% за год) начала мировой ядерной войны или подобного ей катаклизма возможна практически неизбежная гибель людей, живущих в гораздо более медленном времени, чем их цифровые имитации. Чтобы устранить эту проблему, может потребоваться более стабильный порядок в цифровой реальности, например правление синглтона, постепенно приобретающего всю бóльшую стабильность.

[16] Даже при условии, что машины будут намного эффективнее людей, все-таки может сохраниться *некоторый* уровень зарплат, при котором окажется выгоднее нанимать работников-людей, скажем, за один цент в час. Но если для самих людей это обернется единственным источником дохода, наш род просто вымрет, потому что не сможет прокормить себя на такие суммы. Конечно, у людей останется еще доход на капитал. Если теперь предположить, что население будет расти до тех пор, пока доход на душу населения не дойдет до уровня прожиточного минимума, может показаться, что после этого люди станут работать усерднее. Допустим, прожиточный минимум равен одному доллару в день на человека. Тогда получается, что население будет расти до тех пор, пока он не снизится до девяноста центов, и людям придется работать на десять часов больше, чтобы получить эти недостающие десять центов. Однако это не обязательно так, поскольку прожиточный минимум зависит от количества выполненной работы — люди, занимающиеся трудом, сжигают больше калорий. Представим себе, что каждый час работы увеличивает издержки на питание на два цента. И теперь можно получить модель, в которой человечество находится в равновесии.

[17] Может показаться, что такая ситуация невозможна, поскольку «организм, находящийся в подобном овощном состоянии» не сможет голосовать и защищать свои права. Но он может оставить доверенность ИИ на управление делами и защиту его политических интересов. (Все наши предположения исходят из того, что права собственности будут уважаться.)

[18] Не очень понятно, какой термин выбрать: слово *убивать* звучит более жестоко, чем это может оказаться на самом деле; слово *отключать* звучит, может быть, более приемлемо, но для нашего случая оно слишком упрощает ситуацию. А сложность в том, что фактически речь идет о двух отдельных процедурах: остановке активной деятельности и удалении информационных шаблонов. Смерть человека обычно предполагает одновременно и то и другое, но в случае с эмуляторами эти процедуры могут быть разделены во времени. Мы могли бы сравнить *временную* остановку программы с обычным человеческим сном, а *постоянную* остановку — с комой. Еще сильнее усложняет ситуацию возможность копирования эмуляторов и запуск их с различными скоростями — это вообще не имеет прямых аналогий с человеческой жизнью. См.: [Bostrom 2006 b; Bostrom, Yudkowsky 2014].

[19] Придется искать компромисс между общей мощностью параллельных вычислений и их скоростью, поскольку высочайшая скорость вычислений достижима только за счет снижения их эффективности. Это будет особенно верно с началом эры обратимых вычислений.

[20] Эмуляторы можно испытывать, подвергая их искушению. Путем многократных проверок, как эмулятор, запущенный из состояния готовности, реагирует на различные последовательности стимулов, можно убедиться в его надежности. Но чем дальше в процессе развития он

продвигается от своего исходного состояния, тем сильнее снижается уверенность в нем работодателя. (Поскольку проницательный эмулятор может догадываться, что работает в режиме имитации, следует быть осторожными, экстраполируя его поведение на ситуации, в которых гипотеза о симуляции будет меньше влиять на принимаемые им решения.)

[21] Некоторые эмуляторы скорее будут идентифицировать себя со своим кланом — то есть со всеми копиями и вариациями, полученными из одного шаблона, — чем думать о себе как отдельном субъекте. Такие эмуляторы, скорее всего, не станут считать свое отключение смертью, зная, что другие члены клана выживут. Они могут понимать, что рано или поздно их настройки сбросят к некоторому исходному состоянию и они потеряют память, но будут так же мало беспокоиться об этом, как и любители вечеринок, которые часто не могут вспомнить утром события предыдущей ночи, — считайте это ретроградной амнезией, а не смертью.

[22] С этической оценкой можно подходить и ко многим другим факторам. Даже если все работники испытывают «чувство глубокого удовлетворения» от условий своего труда, результат все-таки может быть сомнительным с нравственной точки зрения по другим основаниям, хотя эти основания и являются предметом споров между сторонниками различных этических учений. Однако важнейшим факторов все равно следует считать субъективное ощущение благополучия. См. также: [Bostrom, Yudkowsky 2014].

[23] См.: [World Values Survey, 2008].

[24] См.: [Helliwell et al. 2012].

[25] См.: [Bostrom 2004]; см. также: [Chislenko 1996; Moravec 1988].

[26] Будут ли системы для обработки информации, возникшие в рамках этого сценария, обладать сознанием (в смысле наличия квалиа и возможности субъективных переживаний)? Трудно сказать. Отчасти из-за эмпирической неопределенности: какого рода когнитивными субъектами они будут. Отчасти из-за философской неопределенности: какие типы систем обладают сознанием. Можно попытаться переформулировать наш вопрос и спросить: будут ли они иметь моральный статус или будут ли они такими, что у нас появятся предпочтения относительно их «благополучия». Но на эти вопросы ответить не легче, чем на вопрос о сознании. На самом деле может оказаться, что ответы на вопросы и о сознании, и о моральном статусе, и о наших предпочтениях будут зависеть от того, смогут ли обсуждаемые системы субъективно переживать условия, в которых находятся.

[27] Свидетельства того, что и в геологической, и человеческой истории прослеживается тенденция в пользу большей сложности, можно найти в книге Роберта Райта «Не ноль. Логика человеческой судьбы» [Wright 2001]. Противоположные аргументы (которые Райт критикует в главе 9 своей книги) см.: [Gould 1990]. См. также: [Pinker 2011], где приведены аргументы в пользу того, что мы являемся свидетелями устойчивой и длительной тенденции снижения уровня насилия и жестокости.

[28] Об эффекте выбора наблюдателя см.: [Bostrom 2002 a].

[29] См.: [Bostrom 2008 a]. Чтобы исключить влияние эффекта наблюдателя, потребуется гораздо более тщательное изучение деталей нашей эволюционной истории; см., например: [Carter 1983; 1993; Hanson 1998 d; Ćirković et al. 2010].

[30] См.: [Kansa 2003].

[31] См., например: [Zahavi, Zahavi 1997].

[32] См.: [Miller 2000].

[33] См.: [Kansa 2003]; см. также: [Frank 1999] — очень интересная книга Роберта Франка «„Роскошная“ лихорадка», дающая серьезную пищу для размышлений.

[34] В принципе, не вполне ясно, как лучше всего измерять степень глобальной политической интеграции. Можно сравнить эпоху охотничье-собирательской культуры, когда в руководящие структуры в общем входило максимум несколько сот человек, принимающих все решения, и наше время, когда крупнейшие политические структуры в общем насчитывают более миллиарда человек. Это означает разницу в семь порядков, всего на порядок меньше ситуации, в которой единой политической структурой стал бы весь мир. Однако в те времена, когда максимальным уровнем интеграции являлось племя, население планеты было несравнимо меньше. Тогда племя могло включать в себя до тысячной доли всех жителей Земли. Оценка доли населения планеты, вовлеченной в процесс политической интеграции, а не абсолютных цифр, кажется более правильной в современных условиях (особенно учитывая, что переход к искусственному интеллекту может вызвать взрывной рост количества жителей планеты за счет появления имитационных моделей или иных форм цифрового разума). Развиваются принципиально отличные от государственных и других официальных институций глобальные организации и сети, которые тоже следует принимать во внимание.

[35] Одна из возможных причин, почему первая революция ИИ должна закончиться довольно быстро, — избыточное количество аппаратных средств. Но это не исключает и прорыва в ПО, связанного с переходом от имитационных моделей к композиционным моделям ИИ.

[36] См.: [Shulman 2010 b].

[37] Как получится уравновесить все *pro* и *contra*, будет зависеть от работы, которую пытается выполнить суперорганизм, и насколько большими возможностями будет обладать наиболее развитый единый шаблон эмуляторов. Человеческая потребность в крупных организационных структурах, в которые постоянно объединяются множества различных типов людей, отчасти вызвана тем, что людей, талантливых сразу во всем, очень мало.

[38] Естественно, изготовить множество копий интеллектуального программного агента очень легко. Но для уверенности, что скопированные имитационные модели будут обладать одинаковыми конечными целями, этого недостаточно. Чтобы два агента имели одинаковые конечные цели (*одинаковые* в прямом смысле этого слова — то есть «одни и те же»), эти цели должны совпадать в своих *дейктических* (то есть указательных) элементах. Если Боб эгоист, то имитация Боба тоже будет эгоистом. Тем не менее цели их не одинаковые, поскольку Боба заботит благополучие Боба, а имитацию Боба — благополучие имитации Боба.

[39] [Shulman 2010 b, p. 6].

[40] Это скорее относится к людям и имитационным моделям, а не к системам ИИ, которые могут быть разработаны таким образом, что будут способны включать скрытые модули или функциональные процессы, которые практически невозможно обнаружить. Однако *ИИ, специально созданные для обеспечения прозрачности*, позволят проводить более тщательные контрольные

проверки, чем системы, по архитектуре схожие с человеческим мозгом. Давление общества может заставить ИИ раскрыть свой исходный код или изменить себя в сторону большей прозрачности — особенно если прозрачность является предварительным условием для доверия и возможности заключать выгодные сделки. См.: [Hall 2007].

[41] Есть издержки, которые кажутся относительно незначительными, особенно на фоне общих переговоров, где ставки слишком высоки (например, возможность провала согласования на глобальном уровне ключевых вопросов); имеются в виду издержки при поиске взаимовыгодных политических решений или когда некоторые агенты могут принципиально отстаивать позицию «независимости», степень которой явно понизится в результате присоединения к всеобъемлющим международным соглашениям, предусматривающим механизмы контрольных проверок и надзора.

[42] Возможно, ИИ будут этого добиваться, меняя соответствующим образом свой исходный код, а всем своим наблюдателям предоставляя доступ к себе лишь в режиме «только для чтения». Машинный интеллект с менее прозрачной архитектурой (имитационные модели мозга) может поступить иначе: публично применить к себе один из методов выбора мотивации. Еще один вариант выбора посредника — услуги независимого агента, обладающего принудительной функцией. Например, это может быть полицейский суперорганизм, в чье предназначение входит не только принуждение к исполнению соглашений, достигнутых между несколькими сторонами, но и обеспечение контроля за выполнением обязательств.

[43] Возможно, сам естественный отбор благоволил тем, кто не обращал внимания на угрозы и имел настолько сильный характер, что явно предпочел бы скорее умереть в бою, чем поступиться своими интересами хотя бы в малом. Эта непреклонность могла приносить большую пользу, недвусмысленно сигнализируя окружающим, с кем им придется иметь дело. (Естественно, чисто инструментальное вознаграждение не обязательно играло сколь-нибудь значимую роль в сознательной мотивации агента — он вполне мог ценить справедливость и честь как таковые.)

[44] Для вынесения окончательного вердикта по этому вопросу пришлось бы проводить более глубокий анализ. Есть сложности разной степени, но мы сейчас не будем на них задерживаться.

Глава 12. Выработка ценностей

[1] У этой базовой идеи есть множество вариаций. Об одной из них мы рассказывали в главе 8: агент не обязан каждую секунду стремиться все доводить до максимума, вполне возможно существование такого агента, которого бы все «устраивало», то есть агента, отвечающего критерию разумной достаточности. В следующей главе мы коротко затронем проблему альтернативных подходов к принятию решений. Но поскольку сейчас эти вопросы несущественны, мы не будем сбиваться на другие темы, а сосредоточим внимание на агенте, максимизирующем ожидаемую полезность.

[2] При условии, что функция полезности этого ИИ не совсем примитивна. Например, очень легко создать агента, который всегда выбирает действие, максимизирующее ожидаемую полезность

в случае, если функция полезности, например, константа: $U(w) = 0$. При такой функции полезности каждое действие одинаково хорошо максимизировало бы ожидаемую полезность.

[3] Вероятно, мы забыли ту цветовую мешанину, которую наблюдали в раннем младенчестве, когда мозг еще не научился интерпретировать поступающую в него визуальную информацию.

[4] См. также: [Yudkowsky 2011] и [Muehlhauser, Helm 2012] — см. обзор в части пятой.

[5] Вполне возможно, что прогресс в области программирования в конечном счете поможет преодолеть и эти сложности. Используя современные инструменты, один-единственный программист может создавать такие продукты, которые не снились целой команде, вынужденной писать сразу в машинном коде. Сегодня разработчики ИИ могут пользоваться такими выразительными возможностями, как высококачественные библиотеки для машинного обучения и научных вычислений, позволяющие легко собрать, например, приложение для подсчета людей с помощью веб-камеры из библиотек, написать которое с чистого листа мало кому по силам. Благодаря накоплению целого пласта «многоразового» программного обеспечения, созданного специалистами, но доступного неспециалистам, у будущих программистов будет огромный выбор выразительных средств. Например, разработчики роботов смогут воспользоваться стандартными библиотеками изображений лиц, коллекциями типичных офисных объектов, специальными библиотеками траекторий движения и многими другими инструментами, еще недоступными в настоящее время.

[6] См.: [Dawkins 1995, p. 132] — хотя речь не о том, что страданий в мире *больше*, чем радости.

[7] Однако размер популяций всегда был эффективным, поэтому — несмотря на все страдания, войны и смерти — среднее число особей в нашей популяции стабильно обеспечивало передачу свойственных ей генов от поколения к поколению. Вопреки всему наши предки не выродились и не погибли; см.: [Shulman, Bostrom 2012].

[8] Безусловно, с моральной точки зрения было бы намного справедливее, если мы смогли бы легко добиваться подобных результатов, не заставляя страдать множество невинных существ. Но если имитационным моделям все-таки придется претерпевать бессмысленные страдания, то эту несправедливость мы попробуем возместить, сохранив их файлы, а много позже, при более благоприятных условиях — когда человечество обеспечит себе полную безопасность — запустить их снова. В каком-то смысле это возрождение будет напоминать религиозную идею загробной жизни с последующим воскрешением — вполне в духе теологической концепции, пытающейся примирить нашу бренную жизнь с существованием зла.

[9] Один из ведущих специалистов в области обучения с подкреплением Ричард Саттон определяет этот вид обучения не с методологической точки зрения, а в категориях проблематики самого подхода: по его мнению, любой способ, пригодный для решения этой проблемы, является методом обучения с подкреплением [Sutton, Barto 1998, p. 4]. Напротив, наше обсуждение напрямую касается методов, в которых конечной целью агента является стремление получать максимальное совокупное вознаграждение (в том смысле, что «совокупное вознаграждение» представляет собой восприятие общей ценности всех видов поощрения). Например, решить проблему обучения с подкреплением возможно и таким образом: обучить агента с совершенно иными конечными целями имитировать в самых разных ситуациях поведение агента,

стремящегося к максимизации вознаграждения, — в соответствии с мнением Саттона и такой прием допустимо считать «методом обучения с подкреплением», но только в этом случае он не приведет к возникновению эффекта самостимуляции. Однако замечание Саттона верно по отношению к большинству приемов, которые используют в своей практике специалисты в области обучения с подкреплением.

[10] Даже если удастся каким-то образом создать машинный интеллект «человеческого типа», совсем не обязательно, что его конечные цели начнут напоминать конечные цели человека. Разве только условия воспитания цифрового дитя будут близки к условиям воспитания обычного ребенка. Не представляю, как это можно обеспечить, но предположим, кому-то удалось. И все равно результат не будет гарантирован, поскольку даже небольшая разница во врожденных способностях приведет к совершенно иным реакциям на события. Однако вполне допускаю, что в будущем для цифрового разума человеческого типа разработают более надежный механизм ценностного приращения (с использованием новых лекарственных препаратов, имплантатов или их цифровых эквивалентов).

[11] Невольно возникает вопрос: почему мы, *люди*, похоже, никогда не пытаемся «отключить механизм», иногда вынуждающий нас изменять своей прежней системе ценностей? Видимо, роль играют многие факторы. Во-первых, человеческая система мотивации пока плохо описана в качестве алгоритма, отстраненно вычисляющего максимум функции полезности. Во-вторых, у нас может не быть подходящих средств видоизменять пути, которыми мы приобретаем ценности. В-третьих, у нас могут быть инструментальные причины (связанные, в частности, с социальными сигналами, о которых мы говорили в главе 7) иногда приобретать новые конечные цели, поскольку окружающие способны догадываться о наших намерениях, и тогда нам приходится в собственных интересах пересматривать свои цели. В-четвертых, встречаются моменты, когда мы действительно *активно* сопротивляемся чьему-то тлетворному влиянию, заставляющему нас пересмотреть свою систему ценностей. В-пятых, есть вероятный и довольно любопытный вариант: мы наделяем некоторыми конечными ценностями своего рода агента, способного приобретать новые конечные ценности обычным человеческим способом.

[12] Или попытаться создать такую систему мотивации, чтобы ИИ был индифферентен к замене целей; см.: [Armstrong 2010].

[13] Мы опираемся на объяснения, данные Дэниелом Дьюи [Dewey 2011]. Использованы также идеи из работ: [Hutter 2005; Legg 2008; Yudkowsky 2001; Hay 2005].

[14] Чтобы избежать ненужного усложнения, мы остановимся на агентах с детерминированным поведением, которые не дисконтируют будущее вознаграждение.

[15] С математической точки зрения поведение агента можно формализовать при помощи *агентской функции*, ставящей в соответствие каждой возможной истории взаимодействий свое действие. Явно задать агентскую функцию в табличном виде невозможно за исключением случаев самых простых агентов. Вместо этого агенту дается возможность вычислить, какое действие лучше выполнять. Поскольку способов вычисления одной и той же агентской функции может быть много, это ведет к индивидуализации агента в виде *агентской программы*. Агентская программа — это такая программа или алгоритм, которая вычисляет действие,

соответствующее каждой истории взаимодействий. Хотя часто удобнее и полезнее — с математической точки зрения — считать, что агент взаимодействует с другими в некоторой формально определенной среде, важно помнить, что это является идеализацией. На реальных агентов действуют реальные физические стимулы. Это означает не только, что агент взаимодействует со средой посредством датчиков и исполнительных механизмов, но также, что «мозг» или контроллер агента *сам является частью физической реальности*. Поэтому на его поведение, в принципе, могут воздействовать физические помехи извне (а не только объекты восприятия, или перцепты, полученные с датчиков). То есть с какого-то момента становится необходимым считать агента *реализацией агента*. Реализация агента — это физическая структура, которая в отсутствие влияния среды выполняет агентскую функцию. (Определения даны в соответствии с работой Дэниела Дьюи [Dewey 2011].)

[16] Дьюи предлагает следующее определение оптимальности для агента, обучающегося ценностям:

$$y_k = \arg\max_{y_k} \sum_{x_k yx_{k+\min}} P_1\left(yx_{\le m} \mid yx_x y_k\right) \sum_U U\left(yx_{\le m}\right) P_2\left(U \mid yx_{\le m}\right)$$

Здесь P_1 и P_2 — две вероятностные функции. Вторая сумма располагает в определенном порядке некоторый подходящий класс функций полезности по всем возможным историям взаимодействия. В версии, представленной в тексте, мы явно выделили некоторые зависимости, а также упростили обозначение возможных миров.

[17] Нужно заметить, что набор функций полезности U должен быть таким, чтобы полезность можно было сравнивать и усреднять. В принципе, это непросто, кроме того, не всегда очевидно, как представлять различные этические теории в терминах количественно выраженной функции полезности. См., например: [MacAskill 2010].

[18] В более общем случае нужно обеспечить ИИ адекватным представлением условного распределения вероятностей $P(v(U) \mid w)$, поскольку v не всегда может напрямую дать ответ, истинно ли утверждение $v(U)$ в мире w для любой пары «возможный мир — функция полезности» (w, U).

[19] Рассмотрим вначале Y — класс действий, возможных для агента. Одна из сложностей связана с тем, что именно следует считать действием: только базовую моторную команду (вроде «отправить электрический импульс по каналу вывода #00101100») или команду более высокого уровня (вроде «удерживать фокус камеры на лице»)? Поскольку мы скорее пытаемся дать определение оптимальности, а не разработать план практического применения метода, можно ограничить область только базовыми моторными командами (а поскольку набор таких команд может со временем меняться, нам следует проиндексировать Y по времени). Однако чтобы двигаться в сторону практической реализации, очевидно, будет необходимо создать некий процесс иерархического планирования, в рамках которого придется решить, как применять формулу к классу действий более высокого уровня. Еще одна сложность связана с тем, как анализировать внутренние действия системы (вроде записи данных в рабочую память). Поскольку внутренние действия могут иметь важные последствия, в идеале хотелось бы, чтобы в Y были включены и базовые внутренние действия, и моторные команды. Но есть определенные пределы, как далеко можно зайти в этом направлении — вычисление ожидаемой полезности любого действия

из *Y* требует выполнения многочисленных вычислительных действий, и если каждое из них также считается действием из *Y*, которое должно быть оценено в соответствии с моделью ИИ-ОЦ, мы имеем дело с бесконечной регрессией, которая вообще не позволит тронуться с места. Чтобы исключить эту ситуацию, нужно сузить количество явных попыток оценить ожидаемую функцию полезности ограниченным количеством наиболее важных возможностей для совершения действий. После этого систему нужно наделить некоторым эвристическим процессом, который определит список наиболее важных возможностей совершения действий для дальнейшего рассмотрения. (В конечном счете система могла бы сама принимать решения относительно некоторых возможных действий и вносить изменения в этот эвристический процесс, чтобы постепенно приближаться к идеалу, описанному в модели ИИ-ОЦ.)

Теперь рассмотрим *W* — класс возможных миров. Одна из сложностей связана с описанием *W* так, чтобы он оказался достаточно представительным. Отсутствие каких-то важных *w* в *W* приведет к тому, что ИИ не сможет составить представление о некой реальной ситуации и примет неверное решение. Предположим, что для определения вида *W* мы используем какую-то онтологическую теорию. Например, включаем в *W* все возможные миры, составляющие некий пространственно-временной континуум, населенный элементарными частицами, описанными в стандартных физических моделях. Если эта стандартная модель окажется неполной или неправильной, эпистемологическая основа ИИ будет нарушена. Можно попробовать использовать более широкий класс *W*, чтобы покрыть больше возможностей, но даже будучи уверенными, что учтены все возможные физические вселенные, мы не можем исключать, что за скобками остались еще какие-то. Может быть, дуалистические возможные миры, в которых осознаваемые факты не вытекают из физических? Или дейктических фактов? А может быть, нормативных? Математических? Возможно, каких-то иных видов фактов, которые мы, смертные, просмотрели, но которые могут быть важными с точки зрения устройства мира? Есть люди, убежденные в правильности той или иной онтологической теории. (Те, кто создает будущее ИИ, часто принимают как должное веру в материалистическую онтологию, которая предполагает первичность физического и вторичность психического.) Хотя даже недолгое размышление об истории идей поможет понять, что есть высокая вероятность ложности нашей любимой онтологии. Если ученые XIX века попытались бы дать основанное на физических законах описание *W*, они, вероятно, не включили бы в него возможность неэвклидова пространства–времени, квантовой («многомировой») теории Эверетта, космологического мультиверса или иных подобных гипотез — то есть возможностей, вероятность которых сегодня представляется довольно высокой. Вполне может быть, что и в наши дни существуют возможности, о которых не подозревает нынешнее поколение людей. (В то же время, если *W* будет слишком большим, могут возникнуть технические трудности, связанные с операциями над трансфинитными множествами.) Идеальным решением мог бы стать подход, в границах которого ИИ наделяется какой-то открытой онтологией с возможностью ее самостоятельного расширения на базе тех же принципов, которыми пользуемся мы сами, принимая решение, признавать или нет новый тип метафизических возможностей.

Теперь рассмотрим $P(w \mid Ey)$. Определение этой условной вероятности, строго говоря, не является частью проблемы загрузки ценностей. Чтобы считаться разумным, ИИ уже должен

уметь каким-то образом оценивать вероятность возникающих в реальном мире возможностей. Неспособная на это система не будет представлять опасности, о которой мы говорим. Однако существует риск, что эпистемология ИИ окажется достаточно хорошей, чтобы сделать его инструментально эффективным, и при этом недостаточно хорошей, чтобы правильно оценивать возможности, имеющие важное нормативное значение. (В этом смысле проблема определения $P(w|Ey)$ связана с проблемой определения W). Определение $P(w|Ey)$ также требует преодоления и других трудностей, в частности: как представлять неопределенность, связанную с логически невозможными событиями.

Упомянутые выше вопросы — как определить класс возможных действий, класс возможных миров и распределение вероятности, связывающее событие с классами возможных миров, — имеют довольно общий характер, поскольку те же самые вопросы возникают в случае широкого диапазона формально определяемых агентов. Остается рассмотреть вопросы, более специфические для метода обучения ценностям, а именно как определить U, $V(U)$ и $P(V(U)|w)$.

U — это класс функций полезности. U и W связаны, поскольку каждая функция полезности $U(w)$ в U должна в идеале присваивать полезность каждого возможного мира w из W. Но U тоже должна быть довольно широкой в том смысле, что должна содержать много разных функций полезности — это повысит нашу уверенность, что хотя бы одна из них справится с задачей адекватного представления требуемых ценностей.

Причина написания $P(V(U)|w)$, а не просто $P(U|w)$, в том, чтобы подчеркнуть факт присвоения вероятностей утверждениям. Сама функция полезности утверждением не является, но ее можно трансформировать в утверждение. Например, можно сказать о некоторой функции полезности $U(.)$, что она описывает предпочтения некоторого субъекта, или представляет утверждения некоторой этической теории, или что эту функцию полезности хотел бы использовать в системе ИИ принципал, если бы долго и глубоко размышлял на эту тему. Тогда «критерий ценности» $V(.)$ может выглядеть как функция, которая в качестве аргумента использует функцию полезности U, а в качестве значения выдает утверждение, что U удовлетворяет критерию V. Определив утверждение $V(U)$, мы, скорее всего, получим условную вероятность $P(V(U)|w)$ из того же источника, который используем для получения и других распределений вероятности нашего ИИ. (Если мы уверены, что все существенные с нормативной точки зрения факты приняты во внимание при задании возможных миров W, тогда в каждом из возможных миров $P(V(U)|w)$ будет равняться нулю или единице.) Остается вопрос, как определить V, — это обсудим далее в основном тексте.

[20] Здесь приведены не единственные сложности метода обучения ценностям. Неясно, например, как наделить ИИ набором достаточно разумных исходных убеждений до того момента, когда он окрепнет настолько, что сможет воспротивиться попыткам программистов их скорректировать.

[21] См.: [Yudkowsky 2001].

[22] «Аве Мария» — термин из американского футбола. Так называется очень длинный пас вперед, сделанный в отчаянной ситуации — обычно когда время на исходе, — в надежде, что кто-то из игроков поймает мяч у зачетного поля противника и выполнит тачдаун.

[23] Подход «Аве Мария» основан на идее, что сверхразум может формулировать свои предпочтения точнее, чем мы, люди, излагаем свои. Например, ИИ может сделать это при помощи кода. Поэтому если наш ИИ представляет другие сверхразумные системы в виде вычислительных процессов, воспринимающих окружающую их среду, то он сможет предположить, как эти системы могли бы реагировать на разные гипотетические стимулы, например «окна», выскакивающие в их поле зрения, с исходным кодом нашего ИИ и предложением сформулировать свои инструкции для нас в каком-то заранее выбранном и удобном для понимания формате. После этого наш ИИ мог бы изучить эти воображаемые инструкции (фактически из своей собственной модели, работающей по принципу «от обратного», в которой и существуют эти «другие» системы сверхразума) и выполнить их, поскольку изначально был мотивирован нами на это.

[24] Альтернативный вариант — создать детектор, который в рамках модели мира нашего ИИ ищет представления физических структур, созданных сверхразумными цивилизациями. Затем мы могли бы исключить шаг определения функций предпочтения этих гипотетических сверхразумных систем и наделить наш ИИ конечными ценностями, предполагающими попытку скопировать те физические структуры, которые, как ему кажется, скорее всего создали бы эти гипотетические системы.

Однако и для этого варианта характерны технические трудности. Например, поскольку наш ИИ, даже достигнув уровня сверхразума, скорее всего, не будет знать с достаточной точностью, какие именно физические структуры создают другие сверхразумные системы, он может попытаться аппроксимировать их. Для этого ему потребуется метрика, с помощью которой он мог бы оценивать сходство двух физических артефактов. Но метрики, основанные исключительно на физических показателях, могут быть неадекватными; например, вывод, что мозг больше похож на камамбер, чем на компьютер, работающий в режиме имитационной модели, был бы в корне неправильным.

Более правильным мог бы быть подход, основанный на поиске «радиомаячков» — сообщений относительно функций полезности, закодированных в каком-то подходящем простом формате. Тогда наш ИИ мог бы искать признаки этих гипотетических сообщений о функциях полезности во Вселенной, а нам оставалось бы надеяться, что дружественные инопланетные системы ИИ создали множество таких «радиомаячков», предвидя (благодаря своему сверхразуму), что более примитивные цивилизации (вроде нашей, человеческой) построят ИИ, чтобы их искать.

[25] Если *все* цивилизации попытаются решить проблему загрузки ценностей при помощи подхода «Аве Мария», дорога окажется тупиковой. Кому-то придется выбрать более трудный путь.

[26] См.: [Christiano 2012].

[27] Искусственному интеллекту, который мы создаем, может быть, вообще не потребуется искать эту модель. Как и мы, он мог бы просто размышлять над тем, какие следствия могли бы быть у столь сложного косвенно заданного определения (возможно, изучая свою среду и следуя тому же ходу рассуждений, которым воспользовались бы и мы).

[28] См. главы 9 и 11.

[29] Например, экстази способен временно повышать эмпатию, а окситоцин — доверие; см.: [Vollenweider et al. 1998; Bartz et al. 2011]. Однако этот эффект меняется в широком диапазоне и сильно зависит от контекста.

[30] Улучшенных субагентов можно было бы убивать, ставить на паузу, сбрасывать до более раннего состояния или лишать полномочий и не подвергать дальнейшему улучшению до тех пор, пока вся система не станет настолько зрелой и безопасной, что эти субагенты перестанут представлять для нее угрозу.

[31] Ответ на этот вопрос может не быть очевидным и по отношению к человеческому обществу, оснащенному великолепным арсеналом новейших средств слежения, биомедицинских методов психологического манипулирования; кроме того, достаточно богатому, чтобы позволить себе огромный штат сотрудников спецслужб, следящих за обычными гражданами (и друг за другом).

[32] См.: [Armstrong 2007; Shulman 2010 b].

[33] Остается открытым вопрос, до какой степени контролер уровня n должен контролировать не только агентов уровня $(n - 1)$, но и агентов уровня $(n - 2)$, чтобы убедиться, так ли хорошо агенты уровня $(n - 1)$ выполняют свою работу. Чтобы узнать, насколько правильно агенты уровня $(n - 1)$ управляют агентами уровня $(n - 1)$, агенту уровня n придется брать под контроль и агентов уровня $(n - 3)$?

[34] Этот метод занимает промежуточное место между методами выбора мотивации и контроля над возможностями. С технической точки зрения та часть системы, которая состоит из людей, контролирующих набор агентов-программ первого уровня, управляет методами контроля над возможностями, а та, что состоит из множества уровней контролирующих друг друга агентов-программ, управляет методами выбора мотивации (постольку, поскольку эта схема определяет мотивацию системы).

[35] На самом деле заслуживают внимания и многие другие издержки, но описывать их здесь не представляется возможным. Например, связанные с тем, что агенты, находящиеся на вершине этой иерархии, могут оказаться коррумпированными или начнут злоупотреблять своей властью.

[36] Чтобы эта гарантия была эффективной, к ее разработке нужно подойти добросовестно. Это поможет избежать манипулирования эмоциональным состоянием эмуляторов и влиять на их принятие решений, в результате чего (например) можно вселить в эмулятора вечный страх, что его отключат или не дадут возможности рационально оценивать имеющиеся у него варианты действий.

[37] См., например: [Brinton 1965; Goldstone 1980; 2001]. (Прогресс социальных наук в этом направлении станет отличным подарком для мировых деспотий: в их распоряжении окажутся более точные предсказательные модели социальных беспорядков, которые помогут им оптимизировать свои стратегии контроля над населением и мягко подавлять мятежи в зародыше с меньшими потерями для всех.)

[38] См.: [Bostrom 2011 a; 2009 b].

[39] В случае полностью искусственной системы можно обеспечить некоторые преимущества институциональной структуры без необходимости создавать субагентов. Например, в процессе

принятия решений можно было бы использовать несколько различных точек зрения, не выделяя их в отдельные сущности с полным набором черт, характерных для независимого агента. Однако в случае, когда система не состоит из субагентов, будет сложнее обеспечить полноценное наблюдение за последствиями поведения, вызванного предлагаемыми изменениями, и возврат к предыдущей версии, если эти последствия окажутся нежелательными.

Глава 13. Выбор критериев выбора

[1] В ходе недавних опросов профессиональных философов была выявлена доля респондентов, которые «поддерживают или склоняются к поддержке» тех или иных теорий. В области нормативной этики: *деонтология* — 25,9%; *консеквенциализм* — 23,6%; *этика добродетели* —18,2%. В области метаэтики: *моральный реализм* — 56,4%; *моральный антиреализм* — 27,7%. В области моральных суждений: *когнитивизм* — 65,7%; *нонкогнитивизм* — 17,0%; см.: [Bourget, Chalmers 2009].

[2] См.: [Pinker 2011].

[3] Обсуждение этого вопроса см. в работе: [Shulman et al. 2009].

[4] См.: [Moore 2011].

[5] См.: [Bostrom 2006 b].

[6] См.: [Bostrom 2009 b].

[7] См.: [Bostrom 2011 a].

[8] Если быть совсем точным, то нам следует полагаться на его мнение за исключением тех случаев, когда у нас есть весомые основания считать, что наши суждения более точные. Например, мы лучше сверхразума знаем, о чем мы думаем в тот или иной момент, — если, конечно, он не научился сканировать наш мозг. Но если у сверхразума есть доступ к нашим взглядам, то этим замечанием руководствоваться не стоит. Тогда мы вполне можем положиться на его мнение, в каких случаях нам доверять собственным суждениям, а в каких нет. (В отдельных случаях, когда речь идет, например, о дейктических знаниях, сверхразум может «встать на наше место» и объяснить, во что нам рациональнее верить.) О философских дискуссиях на темы моральных суждений и эпистемологического авторитета см. в статье Адама Эльги: [Elga 2007].

[9] [Yudkowsky 2004]. См. также: [Mijic 2010].

[10] Например, Дэвид Льюис предложил *диспозиционную теорию ценности*, которая предполагает, что некая вещь *X* значима для *A* тогда и только тогда, когда *A* хотел бы ею обладать, будучи идеально рациональным и идеально информированным об *X*; см.: [Smith et al. 1989]. Родственные идеи были озвучены и ранее, см., например: [Sen, Williams 1982; Railton 1986; Sidgwick, Jones 2010]. Отчасти напоминает их и другой общий философский подход к выработке суждений: *метод рефлексивного равновесия* — процесс итерационной взаимной корректировки наших интуитивных представлений о ситуации, общих правил, которыми мы обычно руководствуемся в аналогичных случаях, и принципами, в соответствии с которыми, как нам кажется, эти элементы могут быть пересмотрены для того, чтобы система стала более согласованной; см., например: [Rawls 1971; Goodman 1954].

[11] Предполагается, что, работая над предотвращением подобных катастрофических исходов, ИИ должен действовать *максимально легкими касаниями*, то есть так, чтобы можно было избежать несчастья и при этом не радикально вмешиваться в судьбу человечества в других смыслах.

[12] [Yudkowsky 2004].

[13] Замечание Ребекки Роуч (личное собщение).

[14] Приведу три этих ценностных принципа, которые могли бы стать конечными целями сверхразума: 1) «защищать людей, будущее человечества и нашу человечную природу» (именно «человечную», а не просто «человеческую», то есть такую природную среду, какую нам *хотелось бы иметь*); 2) «человечество не должно провести остаток вечности, отчаянно сожалея о том, что сделали программисты»; 3) «помогать людям».

[15] Некоторые религиозные сообщества делают особый акцент на вере, противопоставляя ее разумному доводу, которого, по их мнению, недостаточно для обретения духовного опыта — даже в наиболее идеализированной форме разума, даже после упорного и беспристрастного изучения всех священных текстов, откровений и толкований. Те, кто придерживается этих взглядов, вряд ли увидят в КЭВ оптимальный метод принятия решений (однако все-таки решат предпочесть именно его, а не еще более несовершенные методы, которые могут быть реализованы в случае отказа от КЭВ).

[16] Подобный силам природы, действующий незаметно и регулирующий жизнь людей, — такой ИИ скорее можно считать «системным оператором» пространства, занятого человеческой цивилизацией, см.: [Yudkowsky 2001].

[17] «Был бы нанесен» — потому что так *может* случиться, *если* когерентное экстраполированное волеизъявление человечества не согласится рассматривать моральные соображения в отношении этих субъектов, видимо, из-за сомнений в наличии у них морального статуса (хотя сейчас нам кажется более вероятным, что он у них есть). Но даже в случае блокировки решения КЭВ о прямой защите их интересов допустимо думать о возможности, что в рамках существующих правил те индивидуумы, которые желают эти интересы защитить и обеспечить благополучие субъектов за пределами базы экстраполяции, все-таки смогут добиться компромисса по этому вопросу (за счет отказа от части своей доли на ресурсы). Реально это или нет, будет зависеть от того, станет ли результат КЭВ выглядеть как набор базовых принципов, которые позволят в подобных случаях приходить к компромиссу (что, в свою очередь, предполагает решение проблемы стратегического торга).

[18] Индивидуумы, внесшие вклад в создание безопасного и полезного человечеству сверхразума, могут получить за свой труд *некоторую* особую награду, которая в первую очередь не должна быть эксклюзивным правом определять характер деятельности человечества по овладению космическим пространством. Однако идея, что все, входящие в базу экстраполяции, получают *равную* долю, может быть настолько хорошей отправной точкой, что ее не стоит отбрасывать в сторону. В любом случае следует найти способ косвенным образом вознаградить тех, кто этого заслуживает, — альтруистов, работающих на благо всего человечества. Это можно сделать, не присваивая таким людям специального значения в базе экстраполяции, если КЭВ одобрит сам принцип (в том смысле, что присвоит ему какой-то минимальный ненулевой вес).

[19] См.: [Bostrom et al. 2013].

[20] Если у морального суждения, которое мы делаем, есть некий (довольно четко выраженный) смысл, понятный другим, сверхразум сможет определить его значение. Если суждение обладает свойствами высказывания (то есть имеет пропозициональный, или истинностный, характер, позволяющий ему быть истинным или ложным), сверхразум все равно сможет отыскать истинное значение, выраженное в виде: «агент *X* должен сейчас сделать *O*». Так или иначе, но с подобной задачей он справится лучше нас. Даже не имеющий изначально способностей к оценке моральных суждений ИИ может ими овладеть, если обладает сверхмощью в области совершенствования интеллекта. Один из способов сделать это — разобраться, как функционирует этот механизм в человеческом мозгу, и разработать аналогичный, но более быстрый, действующий с более точной фактической информацией.

[21] В силу неопределенности вопроса, связанного с метаэтикой, нужно решить, что должен делать ИИ, если не выполняются предварительные условия МП. Один из вариантов — постановить, что ИИ должен отключиться, если присвоит довольно высокую вероятность тому, что когнитивизм не работает или что подходящих абсолютных моральных истин не существует. Или предложить ему воспользоваться каким-то альтернативным методом вроде КЭВ.

Можно также уточнить метод МП, чтобы было понятнее, как поступать в различных неоднозначных или вырожденных случаях. Например, если теория ошибок верна (и, как следствие, любые утвердительные моральные суждения вида «я должен сейчас делать *T*» ложны), тогда должна реализовываться запасная стратегия (например, отключение). Нам также нужно указать, как поступать, если существует несколько возможных действий, каждое из которых будет отвечать критерию моральной правоты. Например, можно сказать, что в таких случаях ИИ следует выполнить одно из возможных действий, которые предпочла бы коллективная экстраполяция человечества. Можно также оговорить, что произойдет, если в базовом словаре истинной этической теории не окажется терминов вроде «моральная правота». Например, в рамках консеквенциалистской теории одни действия могут считаться лучше других, но отсутствовать такое понятие, как «морально правильное действие». Тогда следует сказать, что в случае истинности такой теории ИИ следует выполнить одно из действий, наиболее приемлемых с моральной точки зрения, если таковое имеется; если приемлемых действий бесконечно много и для каждого из них есть лучшее с моральной точки зрения, тогда ИИ следует выполнить любое из тех действий, которое лучше лучшего действия, выбранного в такой ситуации человеком; если такового нет, то действие, которое как минимум не хуже того лучшего действия, выбранного в такой ситуации человеком.

Размышляя, как можно уточнить метод МП, нужно помнить о нескольких важных моментах. Во-первых, следует придерживаться консервативного подхода и использовать запасной вариант практически во всех случаях, оставив вариант «моральной правоты» только для тех ситуаций, которые мы полностью понимаем. Во-вторых, определение МП следует дополнить общим модулятором, что оно «должно трактоваться доброжелательно и пересматриваться так же, как его пересмотрели бы люди, если могли бы долго и глубоко размышлять над ним, прежде чем его сформулировать, и т. д.».

[22] Из этих терминов лишь «знания» кажется наиболее подходящим для формального анализа (в информационно-теоретическом смысле). Однако чтобы представить, что означает для человека «знать что-то», сверхразуму может потребоваться изощренный набор представлений, связанных со сложными психологическими свойствами. Человек не «знает» всю информацию, которая хранится в его мозгу.

[23] Одним из косвенных показателей меньшей непрозрачности (в очень незначительной степени) терминологии КЭВ является факт, что для анализа модели МП не существует терминологической базы даже такого уровня. Если МП когда-нибудь обрастет «своей» терминологией, то с философской точки зрения это будет огромным прорывом. На самом деле теория идеального наблюдателя — одно из основных направлений в метаэтике — как раз и выполняет функцию описания. См., например: [Smith et al. 1989].

[24] Это требует решения проблемы фундаментальной нормативной неопределенности. Можно показать, что не всегда приемлемо действовать в соответствии с той этической теорией, которая имеет максимальную вероятность оказаться истинной. Можно также показать, что не всегда приемлемо выполнять действие, которое имеет максимальную вероятность быть правильным. Похоже, необходимо как-то корректировать вероятности на «степень неприемлемости» результата и важность того, что поставлено на кон. Некоторые идеи см.: [Bostrom 2009 a].

[25] Можно даже заявить, что любое определение моральной правоты должно отвечать некоторому условию адекватности, например чтобы даже средний гражданин мог разобраться, что хорошо, а что плохо.

[26] Первая опасность состоит в том, что, создавая ИИ, в котором используется модель МП, у нас нет уверенности в *своей* абсолютной моральной правоте, даже если мы исходим из того, что *сам ИИ* всегда будет действовать правильно с этической точки зрения. Возможно, с нашей стороны это вообще будет проявлением высокомерия (если учитывать, сколько людей могут не одобрять проект разработки ИИ). Отчасти проблему можно решить за счет тонкой настройки модели МП. Предположим, мы оговорим, как ИИ следует действовать, то есть делать то, что было бы морально правильно, только если разработчики были морально правы, создав его, — в противном случае он должен отключиться. Непохоже, что, создавая *такой ИИ*, мы были бы морально неправы, поскольку даже если разрабатывать его и было неправильно, то единственным последствием нашего труда стало бы его немедленное отключение, если, конечно, до того момента он не успеет совершить никакого мысленного преступления. (И все-таки мы могли бы быть неправы, например, потому что упустили возможность создать другой ИИ.) Вторая опасность — слишком большая «плотность добра». Представьте, что есть множество морально правильных действий, которые мог бы совершить ИИ, — в смысле *моральной допустимости,* притом что некоторые из них с моральной точки зрения лучше других. Один вариант — чтобы ИИ выбирал в таких случаях лучшее с моральной точки зрения действие (или одно из лучших, если таких несколько). Другой — чтобы он выбирал из одинаково морально допустимых действий такое, которое удовлетворяет какому-то иному (не моральному) критерию. Например, ИИ мог бы выбирать из морально допустимых действий такое, которое

было бы предпочтительнее с точки зрения КЭВ. Такой ИИ никогда не совершит ничего морально недопустимого и поэтому лучше защитит наши интересы, чем тот, который ориентируется лишь на лучшие с моральной точки зрения действия.

[27] Оценивая моральную допустимость нашего действия по созданию ИИ, нужно интерпретировать допустимость объективно. Например, в формальном смысле доктор действует морально допустимо, прописывая пациенту лекарство, которое по своему назначению (и по мнению врача) должно помочь заболевшему человеку — но у пациента на этот препарат аллергия, и в результате он может умереть. Если ИИ сосредоточится на объективной моральной допустимости, он обеспечит себе более выигрышную эпистемологическую позицию.

[28] Еще более прямолинейно: это зависит от *убеждения* ИИ в том, какая этическая теория является истинной (или точнее, от распределения вероятностей, присвоенных им этическим теориям).

[29] Трудно даже представить, какой невероятно чудесной была бы их жизнь. В работе «Письмо из Утопии» я даже поэтически представил эту будущую жизнь [Bostrom 2008 c], а в работе «Почему я хочу жить в постчеловеческом мире» мною приведены аргументы в пользу того, что некоторые из этих возможностей были бы хороши и *для нас*, людей, живущих сейчас [Bostrom 2008 b].

[30] Тактика продвижения одной рекомендации, если при этом считаешь более правильной другую, может показаться обманом или манипуляцией. Но так можно делать, не боясь показаться неискренним. Почему бы открыто не признавать превосходство идеала, но тем не менее выступать за неидеальный вариант в качестве лучшего из возможных компромиссов?

[31] Или какое-то иное слово, предполагающее положительный смысл: *хороший*, *отличный*, *чудесный*.

[32] В программировании принцип «делай то, что я имею в виду» называется ненужной добавкой, или бантиком (Do What I Mean, DWIM); см.: [Teitelman 1966].

[33] Мы выбрали для обсуждения три проектных решения: описание цели, принятие решений и познание мира, но не собираемся поднимать вопрос об их возможной декомпозиции.

[34] С этической точки зрения было бы правильным, чтобы проект предполагал распределение небольшой доли благ, которые будут получены в результате создания сверхразума, в качестве особого вознаграждения тем, кто внес морально допустимый вклад в его успех. Направить на формирование стимулирующего пакета значительную долю благ было бы недостойно. Практически это было бы равно ситуации, когда благотворительный фонд тратит 90% собранных средств на премии сотрудникам и на рекламные кампании по привлечению новых пожертвований.

[35] Как можно вознаградить мертвых? В голову приходят несколько вариантов. Во-первых, организация мемориальных торжеств и создание монументов, которые могли бы считаться наградой в том случае, если люди желали посмертной славы. Во-вторых, имена умерших могли бы быть увековечены в таких сферах, как культура, искусство, архитектура и природа. В-третьих, большинство людей беспокоит судьба их потомков, поэтому специальные привилегии можно даровать их детям и внукам. В-четвертых, сверхразум был бы готов создать сравнительно правдоподобные имитационные модели умерших — модели, обладающие разумом и напоминающие

свои оригиналы настолько, что ушедшие могли бы считаться воскресшими. Скорее всего, последний вариант достижим лишь для тех, кто после смерти был подвергнут криогенной заморозке, но, может быть, сверхразуму не составит труда сделать что-то подобное и на основании других сохранившихся материальных следов умершего человека: личная переписка, публикации, аудиовизуальные материалы, цифровые записи, а также воспоминания живущих. Кроме того, сверхразум найдет и другие возможности, для нас не столь очевидные.

[36] «Ограбление Паскаля» — см.: [Bostrom 2009 b]; неограниченные выигрыши — см.: [Bostrom 2011 a]: фундаментальная нормативная неопределенность — см.: [Bostrom 2009 a].

[37] См., например: [Price 1991; Joyce 1999; Drescher 2006; Yudkowsky 2010; Dai 2009].

[38] См., например: [Bostrom 2009 a].

[39] Уменьшить проблемы, возникающие из-за неверно заданного подхода к принятию решений, можно при помощи косвенной нормативности. Например, при помощи модели КЭВ. При правильном применении она могла бы компенсировать некоторые ошибки, связанные с определением подхода к принятию решений ИИ. Ее использование позволило бы поставить цели, которые наше КЭВ хотело бы внедрить в ИИ, в зависимости от его подхода к принятию решений. Если бы наши идеализированные двойники знали, что они определяют цели для ИИ, который использует определенный подход к принятию решений, то могли бы скорректировать их так, чтобы он вел себя правильно, несмотря на ошибки в этом подходе, — немного напоминает абсурдную ситуацию, когда вслед за линзой, искажающей изображение, ставят другую линзу, исправляющую эту ошибку.

[40] Некоторые эпистемологические системы могут быть цельными и не предполагать явно сформулированных принципов. В этом случае преемственность будет обеспечиваться не выполнением этих принципов, а скорее некоторой эпистемологической точкой отсчета — теми или иными предпочтениями, которыми и определяется реакция субъекта на поток событий.

[41] Обсуждение проблемы искажения — см.: [Bostrom 2011 a].

[42] Например, одним из камней преткновения в рассуждениях о роли наблюдателя является правомерность следующего допущения: из факта вашего существования следует гипотеза, что существованию большего количества наблюдателей N следует присвоить вероятность, пропорциональную N. Аргументы против этого допущения можно найти в незавершенном эксперименте о самонадеянном философе, описанном в книге: [Bostrom 2002 a]. Аргументы в его защиту см.: [Olum 2002]; критику этих аргументов — см.: [Bostrom, Ćirković 2003]. Убежденность в правомерности данного допущения может влиять на различные эмпирические гипотезы, имеющие потенциально высокую стратегическую важность, например на теорему о конце света Картера–Лесли («Аргумент Судного дня»), гипотезу о симуляции и гипотезу о «великом фильтре». См.: [Bostrom 2002 a; 2003 a; 2008 a; Carter 1983; Ćirković et al. 2010; Hanson 1998 d; Leslie 1996; Tegmark, Bostrom 2005]. То же можно сказать и о других аспектах теории выбора наблюдателя.

[43] См., например: [Howson, Urbach 1993]. Есть и другие интересные результаты, сужающие диапазон ситуаций, в которых два байесовских агента могут рационально не согласиться друг с другом, притом что их мнение основано на общем знании; см.: [Aumann 1976; Hanson 2006].

[44] См. концепцию «верховного судии»: [Yudkowsky 2004].

[45] В эпистемологии остаются нерешенными много важных вопросов, некоторые из них были нами упомянуты. Важно то, что нам, возможно, не обязательно иметь все ответы, чтобы обеспечить результат, практически неотличимый от наилучшего. Вполне может сработать смешанная модель, основанная на широком диапазоне априорных распределений.

Глава 14. Стратегический ландшафт

[1] Этот принцип впервые сформулирован в моей статье «Грабеж с насилием Паскаля» [Bostrom 2009 b, p. 190], там же отмечается, что это не тавтология. Визуальным аналогом может быть коробка большого, но ограниченного объема, представляющая собой пространство базовых возможностей, которые могут быть получены при помощи некоторой технологии. Представьте, что вы насыпаете в коробку песок, символизирующий усилия исследователей. От того, как вы его распределяете, зависит, в каком месте коробки образуется горка. Но если продолжать сыпать песок, рано или поздно заполнится весь ее объем.

[2] См.: [Bostrom 2002 b].

[3] Это не согласуется с традиционным взглядом на научно-технологическую политику. Харви Аверч считает, что между 1945 и 1984 гг. в США она была сфокусирована на поиске оптимального уровня правительственного финансирования научных и технологических предприятий, а также на обсуждении проблемы, насколько активно правительство должно поощрять лучших, чтобы обеспечить наибольшее процветание национальной экономики и скорейший рост ее военной мощи. В этих расчетах технологический прогресс всегда благо, но Аверч отмечает рост критических замечаний в адрес этого посыла [Averch 1985]. См. также: [Graham 1997].

[4] См.: [Bostrom 2002 b].

[5] Конечно, здесь явно просматривается тавтология. Легко можно построить аргументацию в пользу обратного порядка разработки подобных технологий. Скажем, для человечества будет лучше вначале решить менее трудную проблему — создать нанотехнологии, — потому что в результате мы создадим лучшие институты, усилим международную координацию и станем более зрелыми в вопросах глобальной стратегии. Может быть, правильнее потренироваться на какой-то проблеме, с которой связана не столь сложная угроза, как в случае машинного интеллекта? Нанотехнологии (или синтетическая биология, или какое-то менее опасное направление, с которым мы столкнемся впервые) могли бы стать своеобразной стремянкой, встав на которую мы дотянулись бы до уровня возможностей, которых будет достаточно, чтобы справиться с более высоким уровнем угрозы со стороны сверхразума. Этот аргумент следует оценивать для каждой ситуации отдельно. Например, в случае нанотехнологий следует помнить о самых разных последствиях их появления: скачок производительности аппаратного обеспечения за счет создания наноподложек микросхем; влияние на экономических рост дешевизны физического капитала; широкое распространение совершенных технологий слежения; возможность формирования синглтона как прямого или косвенного следствия прорыва в области нанотехнологий; огромные возможности развития таких направлений, как разработка машинного интеллекта, нейроморфного ИИ и полная эмуляция головного мозга.

Рассмотрение всех этих вопросов (и других, возникающих в результате появления новых технологий, с которыми связан экзистенциальный риск) выходит за рамки нашей книги. Здесь мы лишь отмечаем причину, которая на первый взгляд обосновывает желание начать с разработки сверхразума, подчеркивая, впрочем, что в некоторых случаях возможно появление аргументов, способных изменить эту предварительную оценку.

[6] См.: [Pinker 2011; Wright 2001].

[7] Может возникнуть соблазн выдвинуть гипотезу, что ускорять одновременно все бессмысленно, поскольку тогда (на первый взгляд) это ускорение не будет иметь видимых последствий, но см., например: [Shoemaker 1969].

[8] Уровень подготовленности зависит не от количества усилий, потраченных на подготовку, а от того, насколько тщательно сконфигурированы условия и насколько хорошо готовы к нужным действиям те, кто уполномочен на принятие ключевых решений.

[9] Еще одним фактором может быть уровень международного доверия в период, предшествующий взрывному развитию интеллекта. Мы рассмотрим его в разделе «Сотрудничество».

[10] Забавно другое: кажется, среди тех, кто сейчас всерьез интересуется проблемой контроля, непропорционально много представителей полярно иного хвоста распределения интеллектуальных способностей, хотя у этого впечатления могут быть различные альтернативные объяснения. Но если область станет модной, в нее, несомненно, потянутся бездари и психи.

[11] За этот термин я благодарен Карлу Шульману.

[12] Насколько близко должен воспроизводить мозг машинный интеллект, чтобы считаться полноценной имитационной моделью, а не нейроморфным ИИ? Подходящим критерием могло бы быть воспроизведение моделью или системы ценностей, или полного набора когнитивных и оценочных особенностей какого-то конкретного или просто среднестатистического человека, поскольку, вероятно, именно это важно с точки зрения проблемы контроля. Копирование этих качеств требует от модели более высокой точности.

[13] Величина этого ускорения, конечно же, зависит от того, насколько велики приложенные усилия, и от источника использованных при этом ресурсов. Ускорения развития нейробиологии не произойдет, если все дополнительные ресурсы, распределенные на исследования в области компьютерного моделирования мозга, будут сняты с направления общих нейробиологических исследований — если только фокус на моделировании не окажется более эффективным способом развития нейробиологии в целом, чем стандартный портфель исследований в этой области.

[14] См.: [Drexler 1986, p. 242]. Эрик Дрекслер (в личном общении) подтвердил, что эта реконструкция соответствует ходу рассуждений, который он хотел представить в своих работах. Наверное, если кто-то захочет превратить эту схему в логически завершенную цепочку умозаключений, придется добавить множество явных допущений.

[15] Возможно, нам не стоит приветствовать *мелкие* катастрофы по той причине, что они повышают нашу бдительность и не дают произойти *средним* катастрофам, которые были бы необходимы для принятия сильных мер предосторожности, способных предотвратить экзистенциальные катастрофы? (И, конечно, как и в случае биологических иммунных систем, нужно

помнить о возможной гиперреакции, аналога аллергических реакций и аутоиммунных нарушений.)

[16] См.: [Lenman 2000; Burch-Brown 2014].

[17] См.: [Bostrom 2007].

[18] Обратите внимание, что этот аргумент касается только порядка, но не времени, когда произойдут указанные события. Более раннее появление сверхразума поможет снизить другие экзистенциальные риски только в том случае, если его вмешательство изменит последовательность ключевых событий: например, если сверхразум появится раньше, чем будут пройдены определенные вехи в развитии нанотехнологий или синтетической биологии.

[19] Если решить проблему контроля *несоизмеримо* сложнее, чем проблему эффективности машинного интеллекта, и если возможности проекта очень слабо коррелируют с размером проекта, тогда, возможно, лучше вначале реализовать мелкие проекты, особенно если различия в возможностях среди мелких проектов выше. Даже если мелкие проекты отличаются в среднем более низким уровнем компетентности, чем крупные, вероятность того, что у какого-то из них уровень компетентности окажется достаточно высоким, чтобы решить проблему контроля, будет не ниже, чем в случае крупного проекта.

[20] Никто не отрицает, что можно представить появление новых инструментов, которые способны еще больше облегчить общение в глобальных масштабах — например, для высококачественного перевода, точного поиска, повсеместной работы смартфонов, создания привлекательной виртуальной среды для общения и так далее — и для которых будет желательным (или даже обязательным) прогресс в области аппаратного обеспечения.

[21] Инвестиции в развитие технологии полной эмуляции головного мозга ускорят движение в сторону создания полноценной имитационной модели не только прямо (за счет появления новых технических методов), но и косвенно, за счет создания условий, которые обеспечат увеличение финансирования, а также повысят заметность этого направления и доверие к нему.

[22] Сколько может потерять человечество, если контуры будущего будут определяться желаниями одного случайного человека, а не суперпозицией желаний всего народонаселения? Ответ на этот вопрос очень сильно зависит от стандартов оценки, которые мы используем, а также от того, идет ли речь об идеализированных или «необработанных» желаниях.

[23] Например, в то время как два человеческих мозга «обмениваются информацией» очень медленно, посредством языка, ИИ может быть разработан таким образом, что две копии одной и той же программы смогут легко и быстро передавать друг другу и информацию, и навыки. Изначально предназначенный для киберпространства машинный интеллект не нуждается в громоздкой системе приспособления к природной среде обитания, унаследованной людьми от своих предков. Устройство цифрового мозга позволяет ему воспользоваться всеми преимуществами быстрых последовательных вычислений, недоступных мозгу биологическому, а также быстро инсталлировать новые высокооптимизированные функциональные модули (например, для обработки символьной информации, распознавания образов, симуляции, глубинной обработки данных и планирования). У ИИ могут быть заметные преимущества нетехнического

характера, например его легче запатентовать и с ним не связаны типичные для использования КИМГМ сложности этического плана.

[24] Если p_1 и p_2 — вероятности риска неудачи на каждом шаге, то суммарная вероятность неудачи равна $p_1 + (1 - p_1)\, p_2$, поскольку фатальная неудача может случится лишь один раз.

[25] Конечно, возможен и такой вариант, что у лидирующего проекта не будет такой большой форы и он не сможет сформировать синглтон. Также может быть, что синглтон сформируется прежде, чем появится ИИ, даже без вмешательства КИМГМ, в этом случае причины предпочитать сценарий «сначала КММ» отпадают полностью.

[26] Есть ли возможность у сторонников КИМГМ скорректировать усилия по поддержке этого направления так, чтобы при этом минимизировать ущерб для направления ИИ? В этом смысле, вероятно, лучше развивать технологии сканирования, чем нейрокомпьютерного моделирования. (Поддержка разработки аппаратного обеспечения вряд ли сыграет какую-то роль, поскольку благодаря большому коммерческому интересу прогресс в этой области и так довольно заметен.) Развитие технологии сканирования может повысить вероятность многополярного исхода, поскольку устранит потенциальное узкое место метода и позволит создать первую популяцию эмуляторов на базе множества различных человеческих шаблонов, а не воспроизводить сто тысяч миллиардов одинаковых копий нескольких исходных сканов. Прогресс в области технологии сканирования также повысит вероятность того, что узким местом окажется аппаратное обеспечение, а это может замедлить взлет.

[27] У нейроморфного ИИ может не быть и других свойств компьютерной модели мозга, повышающих безопасность проекта, в частности, аналогичного человеческому профиля сильных и слабых когнитивных черт (благодаря которому мы могли бы, опираясь на свои знания о людях, формировать ожидания об уровне развития системы на разных стадиях ее развития).

[28] Если мотивом поддержки направления КИМГМ является желание, чтобы технология полной эмуляции головного мозга была создана раньше разработок ИИ, то следует помнить, что ускорение развития этого направления изменит порядок появления двух форм машинного интеллекта только в том случае, если они и так должны были бы появиться почти сразу друг за другом с небольшим преимуществом в пользу ИИ. В противном случае инвестиции в технологию эмуляции или просто ускорят создание КИМГМ (сократив аппаратный навес и время, оставшееся на подготовку к этому), не изменив порядок появления КИМГМ и ИИ, или вообще не будут иметь никакого успеха (возможен даже обратный эффект, если ускорение разработок приведет к созданию нейроморфного ИИ).

[29] См.: [Hanson 2009].

[30] Конечно, и в этом случае имеется *некоторый* экзистенциальный риск, из-за которого отложить прогресс было бы правильно даже с субъективной точки зрения — это или позволит ныне живущим людям прожить чуть дольше, или обеспечит дополнительное время для попыток снизить опасность.

[31] Предположим, мы можем предпринять некоторое действие, которое приблизит взрывное развитие интеллекта на один год. Предположим также, что люди, в настоящее время населяющие

Землю, умирают со скоростью 1% в год и что по умолчанию риск гибели человечества в результате взрывного развития равен 20% (число произвольное, взято просто для иллюстрации). Тогда ускорение появления взрывного развития интеллекта на один год означает (с субъективной точки зрения) увеличение риска с 20% до 21%, то есть на 5%. Однако большинство живущих за год до начала взрывного развития интеллекта людей заинтересованы отложить его, если тем самым можно снизить риск на один процентный пункт (поскольку большинство людей знают, что их риск смерти в следующий год гораздо меньше 1%, учитывая, что основная смертность приходится на сравнительно узкий сегмент старых и больных людей). Так можно получить модель, в которой население ежегодно голосует за то, чтобы отложить взрывное развитие искусственного интеллекта еще на год, притом что все живущие согласны с тем, что было бы хорошо, если рано или поздно он произошел бы. В реальной жизни, конечно, такого бесконечного откладывания, скорее всего, не будет — из-за плохой координации, ограниченных возможностей прогнозирования или предпочтений иных вещей, нежели личное выживание. Если вместо стандартов, затрагивающих личные интересы каждого, использовать экономический коэффициент дисконтирования, позитивный эффект от ускорения резко снизится, так как упадет приведенная стоимость того, что ныне живущие люди обеспечат себе астрономически длинные жизни. Этот эффект проявится особенно заметно, если коэффициент дисконтирования применять к субъективному времени каждого индивидуума, а не к звездному. Если будущие блага дисконтируются по ставке x% годовых, а фоновый уровень экзистенциального риска от других источников равен y% в год, то оптимальный момент для взрывного развития искусственного интеллекта настанет тогда, когда его откладывание еще на год приведет к снижению связанного с ним экзистенциального риска меньше, чем на $x + y$ процентных пунктов.

[32] Я в долгу перед Карлом Шульманом и Стюартом Армстронгом за помощь в создании этой модели. См. также: «Дэвид Чалмерс сообщает о полном согласии среди курсантов и преподавателей Военной академии США в Вест-Пойнте, считающих, что правительство не должно сдерживать исследования в области ИИ даже перед лицом потенциальной катастрофы из-за опасности, что решающее преимущество могут получить конкурирующие державы [Chalmers 2010]» [Shulman 2010 a, p. 3].

[33] Слова о том, что информация — это всегда плохо, являются вплне ожидаемым утверждением. Конечно, все зависит от того, что это за информация, поскольку иногда она может оказаться положительной, например если выяснится, что разрыв между лидером и преследователями гораздо больше, чем предполагалось.

[34] С таким столкновением может быть связан экзистенциальный риск, особенно если ему предшествовало создание новых военных технологий или беспрецедентный рост арсеналов.

[35] Занятые в проекте люди могут жить в совершенно разных местах и связываться друг с другом по защищенным каналам. Но эта тактика является компромиссной с точки зрения безопасности: географическая разбросанность помогает защититься от военного нападения, но повышает операционную уязвимость, поскольку становится труднее обеспечить сохранность информации, предотвратить ее утечки и похищение сотрудников враждебными структурами.

[36] Обратите внимание, что высокий коэффициент дисконтирования по времени может привести к возникновению ощущения гонки даже в тех случаях, когда известно об отсутствии реальных конкурентов. Высокий коэффициент дисконтирования означает, что о далеком будущем можно почти не беспокоиться. В зависимости от ситуации это может препятствовать проведению оторванных от действительности исследований и разработок, что, скорее всего, приведет к задержке революции машинного интеллекта (хотя, вероятно, сделает ее более резкой, когда она наконец случится, — из-за аппаратного навеса). Но высокий коэффициент дисконтирования — или нежелание думать о будущих поколениях — может также снизить значимость экзистенциального риска. Это потворствует авантюрам, в которых имеется возможность получить быструю выгоду за счет резкого роста риска экзистенциальной катастрофы, устраняет стимулы для инвестирования в безопасность и создает стимулы для более раннего появления машинного интеллекта, фактически имитируя ощущение гонки. Но, в отличие от реальной гонки, высокий коэффициент дисконтирования (или игнорирование будущих поколений) не будет провоцировать конфликт. Устранение ощущения гонки — главное преимущество сотрудничества. В ситуации сотрудничества облегчается обмен идеями о возможных путях решения проблемы контроля, что также является преимуществом, хотя в некотором смысле оно уравновешивается тем фактом, что облегчается также обмен идеями и о возможных путях решения проблемы компетентности. Чистым результатом облегчения обмена идеями может быть некоторое повышение уровня коллективного интеллекта соответствующего исследовательского сообщества.

[37] С одной стороны, контроль над проектом со стороны правительственных структур единственной страны повышает риск, что эта страна монополизирует блага от его реализации. Этот результат хуже того, в котором никому не подотчетные альтруисты стремятся распределить блага между всеми. Более того, такой контроль вовсе не обязательно приведет к распределению благ среди всех граждан этой страны: в зависимости от конкретной страны есть более или менее высокий риск, что блага будут присвоены политической элитой или несколькими лицами, действующими в своих интересах.

[38] Уточним только, что использование стимулирующего пакета (о котором шла речь в главе 12) в некоторых обстоятельствах может побудить людей активно участвовать в работе над проектом, а не оставаться пассивными зрителями.

[39] Похоже, закон убывающей предельной полезности работает и в гораздо меньших масштабах. Большинство людей предпочли бы гарантированный доступ к одной звезде, чем один шанс из миллиарда владеть галактикой из миллиарда звезд. На самом деле большинство людей скорее согласились бы на одну миллиардную ресурсов Земли, чем на один шанс из миллиарда владеть ею целиком.

[40] См.: [Shulman 2010 a].

[41] Использование агрегированных этических теорий вызывает определенные проблемы, когда серьезно относишься к идее, что космос может быть бесконечным; см.: [Bostrom 2011 b]. Проблемы возникают и в том случае, если воспринимать серьезно идею о возможности невероятно огромной, хотя и ограниченной, полезности; см.: [Bostrom 2009 b].

[42] Если увеличивать размер компьютера, в конечном счете столкнешься с релятивистскими ограничениями из-за задержек в передаче информации между различными его частями — сигналы не могут передаваться быстрее скорости света. Если сжимать компьютер, то столкнешься с пределами миниатюризации на квантовом уровне. Если увеличивать плотность компьютера, упрешься в границы черной дыры. Однако нужно признать, что есть вероятность открытия новых физических свойств, которые когда-нибудь позволят обойти эти ограничения.

[43] Линейно с ростом ресурсов бесконечно могло бы расти количество копий человека. Хотя неясно, насколько высоко среднестатистический земной житель оценил бы возможность наличия множества своих копий. Даже у тех, кто предпочел бы прожить множество жизней, функция полезности может не быть линейной относительно количества копий. Вполне возможно, что количество копий, как и продолжительность жизни, для среднестатистического человека отвечает закону убывающей предельной полезности.

[44] Высокая степень сотрудничества характерна для синглтона на самом верхнем уровне принятия решений. При этом синглтон *мог бы* отличаться отсутствием сотрудничества и конфликтами на нижних уровнях, если этого захотели бы составляющие его агенты верхнего уровня.

[45] Если каждая из соперничающих команд, работающих над созданием ИИ, убеждена, что ее конкуренты идут в неверном направлении и не имеют никаких шансов приблизить взрывное развитие интеллекта, то исчезает одна из причин для сотрудничества — избежание ощущения гонки — и проектные группы могут независимо друг от друга принять решение замедлить скорость разработки, будучи уверенными, что серьезных конкурентов вокруг них нет.

[46] Работающего над докторской диссертацией.

[47] Предполагалось, что эта формулировка должна пониматься так, как если бы включала оговорку о необходимости принимать во внимание благо животных и других разумных существ (включая цифровые), которые живут или будут жить в будущем. Это не означает выдачу разработчикам ИИ лицензии на замену их этических представлений представлениями более широкого сообщества. Принцип сочетается с методом когерентного экстраполированного волеизъявления, который мы обсуждали в главе 12, когда база экстраполяции состоит из всех людей.

Необходимое пояснение: формулировка не обязательно предполагает отсутствие в постпереходный период прав собственности на искусственный сверхразум или алгоритмы и данные, из которых он состоит. Она нейтральна к тому, какие правовые или политические системы будут лучше с точки зрения организации транзакций в гипотетическом постчеловеческом обществе будущего. Что формулировка *точно* предполагает, так это то, что выбор такой системы в той мере, в какой он определяется типом создаваемого сверхразума, следует делать на базе указанного критерия; то есть система устройства постпереходного общества должна работать на благо всего человечества и служить общим этическим идеалам, а не интересам тех, кто оказался первым создателем сверхразума.

[48] Очевидно, что принцип распределения сверхдоходов можно уточнить. Например, порог распределения может быть сформулирован в суммах на душу населения, или доля сверхдохода, которую может оставить у себя победитель, должна быть выше средней, чтобы сильнее

мотивировать будущее производство (например, использовать принцип теории справедливости Ролза). Можно, в принципе, уйти от сумм в денежном выражении и описать задачу следующим образом: «влияние на будущее человечества», «степень того, насколько интересы каждой из сторон будут учитываться в функции полезности будущего синглтона» или что-то подобное.

Глава 15. Цейтнот

[1] Некоторые исследования важны не потому, что приводят к открытиям, но по каким-то иным причинам, например, потому что развлекают, развивают, возвеличивают или вдохновляют тех, кто ими занят.

[2] Я не имею в виду, что *никто* не должен заниматься чистой математикой или философией. Или что эти направления означают пустую трату времени по сравнению со всеми остальными областями науки или жизни в целом. Вероятно, очень хорошо, что некоторые люди могут посвятить себя интеллектуальной деятельности и следовать за своим научным любопытством, куда бы оно ни завело, независимо от любых соображений относительно полезности или возможных последствий этих занятий. Я имею в виду, что как минимум некоторые из этих лучших умов могли бы, поняв, что в обозримом будущем их когнитивная эффективность может совершенно обесцениться, пожелать переключить свое внимание на теоретические проблемы, скорость решения которых имеет для человечества некоторое значение.

[3] Однако следует быть внимательными в случаях, когда неопределенность может быть полезна, — вспомните, например, модель риска, связанного с гонкой вооружений (врезка 13), где мы выяснили, что дополнительная стратегическая информация может быть вредна. Говоря в общем, нас должна беспокоить потенциальная опасность информации, см.: [Bostrom 2011 b]. Возникает соблазн сказать, что нам нужно больше анализировать потенциально опасную информацию. Вероятно, это правда, хотя в таком случае нас не может не беспокоить, что в результате такого анализа, в свою очередь, появится потенциально опасная информация.

[4] См.: [Bostrom 2007].

[5] Я благодарен Карлу Шульману за то, что он подчеркнул этот момент.

Библиография

Acemoglu 2003 — *Acemoglu Daron*. Labor- and Capital-Augmenting Technical Change // Journal of the European Economic Association, 2003, 1 (1), p. 1–37.

Albertson, Thomson 1976 — *Albertson D. G., Thomson J. N*. The Pharynx of *Caenorhabditis Elegans* // Philosophical Transactions of the Royal Society B: Biological Sciences, 1976, 275 (938), p. 299–325.

Allen 2008 — *Allen Robert C*. A Review of Gregory Clark's «A Farewell to Alms: A Brief Economic History of the World» // Journal of Economic Literature, 2008, 46 (4), p. 946–973.

American Horse Council, 2005 — National Economic Impact of the US Horse Industry // American Horse Council, 2005 (Retrieved 2013, July 30; http://www.horsecouncil.org/national-economic-impact-us-horse-industry).

Andres et al. 2012 — *Andres B., Koethe U., Kroeger T., Helmstaedter M., Briggman K. L., Denk W., Hamprecht F. A*. 3D Segmentation of SBFSEM Images of Neuropil by a Graphical Model over Supervoxel Boundaries // Medical Image Analysis, 2012, 16 (4), p. 796–805.

Armstrong 2012 — *Armstrong Alex*. Computer Competes in Crossword Tournament // I Programmer, 2012, March 19.

Armstrong 2007 — *Armstrong Stuart*. Chaining God: A Qualitative Approach to AI, Trust and Moral Systems // Unpublished manuscript, 2007, October 20 (Retrieved 2012, December 31; http://www.neweuropeancentury.org/GodAI.pdf).

Armstrong 2010 — *Armstrong Stuart*. Utility Indifference. Technical Report 2010–1. Oxford: Future of Humanity Institute, University of Oxford, 2010.

Armstrong 2013 — *Armstrong Stuart*. General Purpose Intelligence: Arguing the Orthogonality Thesis // Analysis and Metaphysics, 2013, 12, p. 68–84.

Armstrong, Sandberg 2013 — *Armstrong Stuart*, *Sandberg Anders*. Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox // Acta Astronautica, 2013, 89, p. 1–13.

Armstrong, Sotala 2012 — *Armstrong Stuart*, *Sotala Kaj*. How We're Predicting AI — or Failing To // Beyond AI: Artificial Dreams / Eds. Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, Radek Schuster. Pilsen: University of West Bohemia, 2012, p. 52–75 (Retrieved 2013, February 2).

Asimov 1942 — *Asimov Isaac*. Runaround // Astounding Science-Fiction, 1942, March, p. 94–103.

Asimov 1985 — *Asimov Isaac*. Robots and Empire. New York: Doubleday, 1985.

Aumann 1976 — *Aumann Robert J.* Agreeing to Disagree // Annals of Statistics, 1976, 4 (6), p. 1236–1239.

Averch 1985 — *Averch Harvey Allen*. A Strategic Analysis of Science and Technology Policy. Baltimore: Johns Hopkins University Press, 1985.

Azevedo et al. 2009 — *Azevedo F. A. C., Carvalho L. R. B., Grinberg L. T., Farfel J. M., Ferretti R. E. L., Leite R. E. P., Jacob W., Lent R., Herculano-Houzel S*. Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain // Journal of Comparative Neurology, 2009, 513 (5), p. 532–541.

Baars 1997 — *Baars Bernard J*. In the Theater of Consciousness: The Workspace of the Mind. New York: Oxford University Press, 1997.

Baratta 2004 — *Baratta Joseph Preston*. The Politics of World Federation: United Nations, UN Reform, Atomic Control. Westport, CT: Praeger, 2004.

Barber 1991 — *Barber E. J. W.* Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean. Princeton, NJ: Princeton University Press, 1991.

Bartels et al. 2008 — *Bartels J., Andreasen D., Ehirim P., Mao H., Seibert S., Wright E. J., Kennedy P*. Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex // Journal of Neuroscience Methods, 2008, 174 (2), p. 168–176.

Bartz et al. 2011 — *Bartz Jennifer A., Zaki Jamil, Bolger Niall, Ochsner Kevin N*. Social Effects of Oxytocin in Humans: Context and Person Matter // Trends in Cognitive Science, 2011, 15 (7), p. 301–309.

Basten et al. 2013 — *Basten Stuart, Lutz Wolfgang, Scherbov Sergei*. Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility // Demographic Research, 2013, 28, p. 1145–11466.

Baum 2004 — *Baum Eric B*. What Is Thought? Bradford Books. Cambridge, MA: MIT Press, 2004.

Baum et al. 2011 — *Baum Seth D., Goertzel Ben, Goertzel Ted G*. How Long Until Human Level AI? Results from an Expert Assessment // Technological Forecasting and Social Change, 2011, 78 (1), p. 185–195.

Beal, Winston 2009 — *Beal J., Winston P*. Guest Editors' Introduction: The New Frontier of Human-Level Artificial Intelligence // IEEE Intelligent Systems, 2009, 24 (4), p. 21–23.

Bell, Gemmel 2009 — *Bell C. Gordon, Gemmell Jim*. Total Recall: How the E-Memory Revolution Will Change Everything. New York: Dutton, 2009.

Benyamin et al. 2013 — *Benyamin B., Pourcain B. St., Davis O. S., Davies G., Hansell M. K., Brion M.-J. A., Kirkpatrick R. M*. Childhood Intelligence is Heritable, Highly Polygenic and Associated With FNBP1L // Molecular Psychiatry, 2013, January 23.

Berg, Rietz 2003 — *Berg Joyce E., Rietz Thomas A*. Prediction Markets as Decision Support Systems // Information Systems Frontiers, 2003, 5 (1), p. 79–93.

Berger et al. 2008 — *Berger Theodore W., Chapin J. K., Gerhardt G. A., Soussou W. V., Taylor D. M., Tresco P. A*. Brain-Computer Interfaces: An International Assessment of Research and Development Trends. Springer, 2008.

Berger et al. 2012 — *Berger T. W., Song D., Chan R. H., Marmarelis V. Z., LaCoss J., Wills J., Hampson R. E., Deadwyler S. A., Granacki J. J.* A Hippocampal Cognitive Prosthesis: Multi-Input, Multi-Output Nonlinear Modeling and VLSI Implementation // IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2012, 20 (2), p. 198–211.

Berliner 1980 a — *Berliner Hans J*. Backgammon Computer-Program Beats World Champion // Artificial Intelligence, 1980, 14 (2), p. 205–220.

Berliner 1980 b — *Berliner Hans J*. Backgammon Program Beats World Champ // SIGART Newsletter, 1980, 69, p. 6–9.

Bernardo, Smith 1994 — *Bernardo José M., Smith Adrian F. M*. Bayesian Theory. 1st ed. Wiley Series in Probability & Statistics. New York: Wiley, 1994.

Birbaumer et al. 2008 — *Birbaumer N., Murguialday A. R., Cohen L*. Brain-Computer Interface in Paralysis // Current Opinion in Neurology, 2008, 21 (6), p. 634–638.

Bird, Layzell 2002 — *Bird Jon, Layzell, Paul*. The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors // Proceedings of the 2002 Congress on Evolutionary Computation, 2002, 2, p. 1836–1841.

Blair 1957 — *Blair Clay, Jr*. Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country // Life, 1957, February 25, p. 89–104.

Bobrow 1968 — *Bobrow Daniel G*. Natural Language Input for a Computer Problem Solving System // Semantic Information Processing / Ed. Marvin Minsky. Cambridge, MA: MIT Press, 1968, p. 146–227.

Bostrom 1997 — *Bostrom Nick*. “Predictions from Philosophy? How Philosophers Could Make Themselves Useful, 1997 (Unpublished manuscript. Last revised 1998, September 19).

Bostrom 2002 a — *Bostrom Nick*. Anthropic Bias: Observation Selection Effects in Science and Philosophy. New York: Routledge, 2002.

Bostrom 2002 b — *Bostrom Nick*. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards // Journal of Evolution and Technology, 2002, 9.

Bostrom 2003 a — *Bostrom Nick*. Are We Living in a Computer Simulation? // Philosophical Quarterly, 2003, 53 (211), p. 243–255.

Bostrom 2003 b — *Bostrom Nick*. Astronomical Waste: The Opportunity Cost of Delayed Technological Development // Utilitas, 2003, 15 (3), p. 308–314.

Bostrom 2003 c — *Bostrom Nick*. Ethical Issues in Advanced Artificial Intelligence — Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence / Eds. Iva Smit and George E. Lasker. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics, 2003, 2, p. 12–17.

Bostrom 2004 — *Bostrom Nick*. The Future of Human Evolution // Death and Anti-Death. Vol. 2: Two Hundred Years After Kant, Fifty Years After Turing / Ed. Charles Tandy. Palo Alto, CA: Ria University Press, 2004, 2, p. 339–371.

Bostrom 2006 a — *Bostrom Nick*. How Long Before Superintelligence? // Linguistic and Philosophical Investigations, 2006, 5 (1), p. 11–30.

Bostrom 2006 b — *Bostrom Nick*. Quantity of Experience: Brain-Duplication and Degrees of Consciousness // Minds and Machines, 2006, 16 (2), p. 185–200.

Bostrom 2006 c — *Bostrom Nick*. What is a Singleton? // Linguistic and Philosophical Investigations, 2006, 5 (2), p. 48–54.

Bostrom 2007 — *Bostrom Nick*. Technological Revolutions: Ethics and Policy in the Dark // Nanoscale: Issues and Perspectives for the Nano Century / Eds. Nigel M. de S. Cameron, M. Ellen Mitchell. Hoboken, NJ: Wiley, 2007, p. 129–152.

Bostrom 2008 a — *Bostrom Nick*. Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing // MIT Technology Review, 2008, May/June, p. 72–77.

Bostrom 2008 b — *Bostrom Nick*. Why I Want to Be a Posthuman When I Grow Up // Medical Enhancement and Posthumanity / Eds. Bert Gordijn, Ruth Chadwick. New York: Springer, 2008, p. 107–137.

Bostrom 2008 c — *Bostrom Nick*. Letter from Utopia // Studies in Ethics, Law, and Technology, 2008, 2 (1), p. 1–7.

Bostrom 2009 a — *Bostrom Nick*. Moral Uncertainty — Towards a Solution? // Overcoming Bias (blog), 2009, January 1.

Bostrom 2009 b — *Bostrom Nick*. Pascal's Mugging // Analysis, 2009, 69 (3), p. 443–445.

Bostrom 2009 c — *Bostrom Nick*. The Future of Humanity // New Waves in Philosophy of Technology / Eds. Jan Kyrre Berg Olsen, Evan Selinger, Søren Riis. New York: Palgrave Macmillan, 2009, p. 186–215.

Bostrom 2011 a — *Bostrom Nick*. Information Hazards: A Typology of Potential Harms from Knowledge // Review of Contemporary Philosophy, 2011, 10, p. 44–79.

Bostrom 2011 b — *Bostrom Nick*. Infinite Ethics // Analysis and Metaphysics, 2011, 10, p. 9–59.

Bostrom 2012 — *Bostrom Nick*. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents // Minds and Machines / Ed. Vincent C. Müller. Journal for Artificial Intelligence, Philosophy and Cognitive Science, 2012, 22 (2), p. 71–85.

Bostrom, Ćirković 2003 — *Bostrom Nick, Ćirković Milan M*. The Doomsday Argument and the Self-Indication Assumption: Reply to Olum // Philosophical Quarterly, 2003, 53 (210), p. 83–91.

Bostrom, Ord 2006 — *Bostrom Nick, Ord Toby*. The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics // Ethics, 2006, 116 (4), p. 656–679.

Bostrom, Roache 2011 — *Bostrom Nick, Roache Rebecca*. Smart Policy: Cognitive Enhancement and the Public Interest // Enhancing Human Capacities / Eds. Julian Savulescu, Ruud ter Meulen, Guy Kahane. Malden, MA: Wiley—Blackwell, 2011, p. 138–149.

Bostrom, Sandberg 2009 a — *Bostrom Nick, Sandberg Anders*. Cognitive Enhancement: Methods, Ethics, Regulatory Challenges // Science and Engineering Ethics, 2009, 15 (3), p. 311–341.

Bostrom, Sandberg 2009 b — *Bostrom Nick, Sandberg Anders*. The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement // Human Enhancement. 1st ed. / Eds. Julian Savulescu, Nick Bostrom. New York: Oxford University Press, 2009, p. 375–416.

Bostrom et al. 2013 — *Bostrom Nick, Sandberg Anders, Douglas Tom*. The Unilateralist's Curse: The Case for a Principle of Conformity. Working Paper, 2013 (Retrieved 2013, February 28; http://www.nickbostrom.com/papers/unilateralist.pdf).

Bostrom, Yudkowsky 2014 — *Bostrom Nick, Yudkowsky Eliezer*. The Ethics of Artificial Intelligence // Cambridge Handbook of Artificial Intelligence / Eds. Keith Frankish, William Ramsey. New York: Cambridge University Press, 2014.

Boswell 1917 — *Boswell James*. Boswell's Life of Johnson. New York: Oxford University Press, 1917.

Bouchard 2004 — *Bouchard T. J*. Genetic Influence on Human Psychological Traits: A Survey // Current Directions in Psychological Science, 2004, 13 (4), p. 148–151.

Bourget, Chalmers 2009 — *Bourget David, Chalmers David* // The PhilPapers Surveys, 2009, November (http://philpapers.org/surveys/).

Bradbury 1999 — *Bradbury Robert J*. Matrioshka Brains, 1999 (Archived version. As revised 2004, August 16; http://web.archive.org/web/20090615040912/http://www.aeiveos.com/~bradbury/).

Brinton 1965 — *Brinton Crane.* The Anatomy of Revolution. Revised ed. New York: Vintage Books, 1965.

Bryson, Ho 1969. — *Bryson Arthur E., Jr., Ho Yu-Chi*. Applied Optimal Control: Optimization, Estimation, and Control. Waltham, MA: Blaisdell, 1969.

Buehler et al. 2009 — *Buehler Martin, Iagnemma Karl, Singh Sanjiv* (Eds.). The DARPA Urban Challenge: Autonomous Vehicles in City Traffic. Springer Tracts in Advanced Robotics 56. Berlin: Springer, 2009.

Burch-Brown 2014 — *Burch-Brown J*. Clues for Consequentialists // Utilitas, 2014, 26 (1), p. 105–119.

Burke 2001 — *Burke Colin*. Agnes Meyer Driscoll vs. the Enigma and the Bombe, 2001 (Unpublished manuscript. Retrieved 2013, February 22; http://userpages.umbc.edu/~burke/driscoll1–2011.pdf).

Canbäck et al. 2006 — *Canbäck S., Samouel P., Price D*. Do Diseconomies of Scale Impact Firm Size and Performance? A Theoretical and Empirical Overview // Journal of Managerial Economics, 2006, 4 (1), p. 27–70.

Carmena et al. 2003 — *Carmena J. M., Lebedev M. A., Crist R. E., O'Doherty J. E., Santucci D. M., Dimitrov D. F., Patil P. G., Henriquez C. S., Nicolelis M. A*. Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates // Public Library of Science Biology, 2003, 1 (2), p. 193–208.

Carroll, Ostlie 2007 — *Carroll Bradley W., Ostlie Dale A*. An Introduction to Modern Astrophysics. 2nd ed. San Francisco: Pearson Addison Wesley, 2007.

Carroll 1993 — *Carroll John B*. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. New York: Cambridge University Press, 1993.

Carter 1983 — *Carter Brandon*. The Anthropic Principle and its Implications for Biological Evolution // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 1983, 310 (1512), p. 347–363.

Carter 1993 — *Carter Brandon*. The Anthropic Selection Principle and the Ultra-Darwinian Synthesis // The Anthropic Principle: Proceedings of the Second Venice Conference on Cosmology and Philosophy / Eds F. Bertola, U. Curi. Cambridge: Cambridge University Press, 1993, p. 33–66.

CFTC/SEC Report on May 6, 2010 — Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. Washington, DC, 2010.

Chalmers 2010 — *Chalmers David John*. The Singularity: A Philosophical Analysis // Journal of Consciousness Studies, 2010, 17 (9–10), p. 7–65.

Chason et al. 2011 — *Chason R. J., Csokmay J., Segars J. H., DeCherney A. H., Armant D. R*. Environmental and Epigenetic Effects Upon Preimplantation Embryo Metabolism and Development // Trends in Endocrinology and Metabolism, 2011, 22 (10), p. 412–420.

Chen, Ravallion 2010 — *Chen S., Ravallion M*. The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty // Quarterly Journal of Economics, 2010, 125 (4), p. 1577–1625.

Chislenko 1996 — *Chislenko Alexander*. Networking in the Mind Age: Some Thoughts on Evolution of Robotics and Distributed Systems, 1996 (Unpublished manuscript).

Chislenko 1997 — *Chislenko Alexander*. Technology as Extension of Human Functional Architecture, 1997 (Extropy Online).

Chorost 2005 — *Chorost Michael*. Rebuilt: How Becoming Part Computer Made Me More Human. Boston: Houghton Mifflin, 2005.

Christiano 2012 — *Christiano Paul F*. «Indirect Normativity» Write-up // Ordinary Ideas (blog), 2012, April 21.

CIA, 2013 — The World Factbook. Central Intelligence Agency, 2013 (https://www.cia.gov/library/publications/the-world- factbook/rankorder/2127rank.html? countryname=United%20States&countrycode=us&regionCode=noa&rank=121#us).

Cicero. On Divination — *Cicero Marcus Tullius*. On Divination // Cicero: On Old Age, On Friendship, On Divination (Loeb Classical Library, No. 154) / Transl. W. A. Falconer. Cambridge, MA: Harvard University Press, 1923.

Cirasella, Kopec 2006 — *Cirasella Jill, Kopec Danny*. The History of Computer Games // Exhibit at Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50). Dartmouth College, 2006, July 13–15.

Ćirković 2004 — *Ćirković Milan M*. Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings // Foundations of Physics, 2004, 34 (2), p. 239–261.

Ćirković et al. 2010 — *Ćirković Milan M*., *Sandberg Anders*, *Bostrom Nick*. Anthropic Shadow: Observation Selection Effects and Human Extinction Risks // Risk Analysis, 2010, 30 (10), p. 1495–1506.

Clark, Chalmers 1998 — *Clark Andy, Chalmers David J*. The Extended Mind // Analysis, 1998, 58 (1), p. 7–19.

Clark 2007 — *Clark Gregory*. A Farewell to Alms: A Brief Economic History of the World. 1st ed. Princeton, NJ: Princeton University Press, 2007.

Clavin 2012 — *Clavin Whitney*. Study Shows Our Galaxy Has at Least 100 Billion Planets // Jet Propulsion Laboratory, 2012, January 11.

CME Group 2010 — What Happened on May 6th? Chicago: CME Group, 2010, May 10.

Coase 1937 — *Coase R. H*. The Nature of the Firm // Economica, 1937, 4 (16), p. 386–405.

Cochran, Harpending 2009 — *Cochran Gregory, Harpending Henry*. The 10,000 Year Explosion: How Civilization Accelerated Human Evolution. New York: Basic Books, 2009.

Cochran et al. 2006 — *Cochran G., Hardy J., Harpending H*. Natural History of Ashkenazi Intelligence // Journal of Biosocial Science, 2006, 38 (5), p. 659–693.

Cook 1984 — *Cook James Gordon*. Handbook of Textile Fibres: Natural Fibres. Cambridge: Woodhead, 1984.

Cope 1996 — *Cope David*. Experiments in Musical Intelligence. Computer Music and Digital Audio Series. Madison, WI: A-R Editions, 1996.

Cotman, Berchtold 2002 — *Cotman Carl W., Berchtold Nicole C*. Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity // Trends in Neurosciences, 2002, 25 (6), p. 295–301.

Cowan 2001 — *Cowan Nelson*. The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity // Behavioral and Brain Sciences, 2001, 24 (1), p. 87–114.

Crabtree 1999 — *Crabtree Steve*. New Poll Gauges Americans' General Knowledge Levels // Gallup News, 1999, July 6.

Cross, Walker 1994 — *Cross Stephen E., Walker Edward*. Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning // Intelligent Scheduling / Eds. Monte Zweben, Mark Fox. San Francisco, CA: Morgan Kaufmann, 1994, p. 711–729.

Crow 2000 — *Crow James F*. The Origins, Patterns and Implications of Human Spontaneous Mutation // Nature Reviews Genetics, 2000, 1 (1), p. 40–47.

Cyranoski 2013 — *Cyranoski David*. Stem Cells: Egg Engineers // Nature, 2013, 500 (7463), p. 392–394.

Dagnelie 2012 — *Dagnelie Gislin*. Retinal Implants: Emergence of a Multidisciplinary Field // Current Opinion in Neurology, 2012, 25 (1), p. 67–75.

Dai 2009 — *Dai Wei*. Towards a New Decision Theory // Less Wrong (blog), 2009, August 13.

Dalrymple 2011 — *Dalrymple David*. Comment on Kaufman J. «Whole Brain Emulation: Looking at Progress on C. Elegans» // Less Wrong (blog), 2011, October 29.

Davies et al. 2011 — *Davies G., Tenesa A., Payton A., Yang J., Harris S. E., Liewald D., Ke X*. Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic // Molecular Psychiatry, 2011, 16 (10), p. 996–1005.

Davis et al. 2010 — *Davis Oliver S. P., Butcher Lee M., Docherty Sophia J., Meaburn Emma L., Curtis Charles J. C., Simpson Michael A., Schalkwyk Leonard C., Plomin Robert*. A Three-Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects // Behavior Genetics, 2010, 40 (6), p. 759–767.

Dawkins 1995 — *Dawkins Richard*. River Out of Eden: A Darwinian View of Life. Science Masters Series. New York: Basic Books, 1995.

De Blanc 2011 — *De Blanc Peter*. Ontological Crises in Artificial Agents' Value Systems. Machine Intelligence Research Institute, San Francisco, CA, 2011, May 19.

De Long 1998 — *De Long J. Bradford*. Estimates of World GDP, One Million B.C. — Present, 1998 (Unpublished manuscript).

Dean 2005 — *Dean Cornelia*. Scientific Savvy? In U.S., Not Much // New York Times, 2005, August 30.

Deary 2001 — *Deary Ian J*. Human Intelligence Differences: A Recent History // Trends in Cognitive Sciences, 2001, 5 (3), p. 127–130.

Deary 2012 — *Deary Ian J*. Intelligence // Annual Review of Psychology, 2012, 63, p. 453–482.

Degnan et al. 2002 — *Degnan G. G., Wind T. C., Jones E. V., Edlich R. F*. Functional Electrical Stimulation in Tetraplegic Patients to Restore Hand Function // Journal of Long-Term Effects of Medical Implants, 2002, 12 (3), p. 175–188.

Devlin et al. 1997 — *Devlin B., Daniels M., Roeder K*. The Heritability of IQ // Nature, 1997, 388 (6641), p. 468–471.

Dewey 2011 — *Dewey Daniel*. Learning What to Value // Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings / Eds. Jürgen Schmidhuber, Kristinn R. Thórisson, Moshe Looks. Lecture Notes in Computer Science 6830. Berlin: Springer, 2011, p. 309–314.

Dowe and Hernandez-Orallo 2012 — *Dowe D. L., Hernandez-Orallo J*. IQ Tests Are Not for Machines, Yet // Intelligence, 2012, 40 (2), p. 77–81.

Drescher 2006 — *Drescher Gary L*. Good and Real: Demystifying Paradoxes from Physics to Ethics. Bradford Books. Cambridge, MA: MIT Press, 2006.

Drexler 1986 — *Drexler K. Eric*. Engines of Creation. Garden City, NY: Anchor, 1986.

Drexler 1992 — *Drexler K. Eric*. Nanosystems: Molecular Machinery, Manufacturing, and Computation. New York: Wiley, 1992.

Drexler 2013 — *Drexler K. Eric*. Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization. New York: PublicAffairs, 2013.

Driscoll 2012 — *Driscoll Kevin*. Code Critique: «Altair Music of a Sort» // Paper presented at Critical Code Studies Working Group Online Conference, 2012, February 6.

Dyson 1960 — *Dyson Freeman J*. Search for Artificial Stellar Sources of Infrared Radiation // Science, 1960, 131 (3414), p. 1667–1668.

Dyson 1979 — *Dyson Freeman J*. Disturbing the Universe. 1st ed. Sloan Foundation Science Series. New York: Harper & Row, 1979.

Elga 2004 — *Elga Adam*. Defeating Dr. Evil with Self-Locating Belief // Philosophy and Phenomenological Research, 2004, 69 (2), p. 383–396.

Elga 2007 — *Elga Adam*. Reflection and Disagreement // Nous, 2007, 41 (3), p. 478–502.

Eliasmith et al. 2012 — *Eliasmith Chris, Stewart Terrence C., Choo Xuan, Bekolay Trevor, DeWolf Travis, Tang Yichuan, Rasmussen Daniel*. A Large-Scale Model of the Functioning Brain // Science, 2012, 338 (6111), p. 1202–1205.

Ellis 1999 — *Ellis J. H*. The History of Non-Secret Encryption // Cryptologia, 1999, 23 (3), p. 267–273.

Elyasaf et al. 2011 — *Elyasaf Achiya, Hauptmann Ami, Sipper Moche*. Ga-Freecell: Evolving Solvers for the Game of Freecell // Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference, 1931–1938. GECCO' 11. New York: ACM, 2011.

Eppig et al. 2010 — *Eppig C., Fincher C. L., Thornhill R*. Parasite Prevalence and the Worldwide Distribution of Cognitive Ability // Proceedings of the Royal Society B: Biological Sciences, 2010, 277 (1701), p. 3801–3808.

Espenshade et al. 2003 — *Espenshade T. J., Guzman J. C., Westoff C. F*. The Surprising Global Variation in Replacement Fertility // Population Research and Policy Review, 2003, 22 (5–6), p. 575–583.

Evans 1964 — *Evans Thomas G*. A Heuristic Program to Solve Geometric-Analogy Problems // Proceedings of the April 21–23, 1964, Spring Joint Computer Conference. AFIPS '64. New York: ACM, 1964, p. 327–338.

Evans 1968 — *Evans Thomas G*. A Program for the Solution of a Class of Geometric-Analogy Intelligence-Test Questions // In Semantic Information Processing / Ed. Marvin Minsky. Cambridge, MA: MIT Press, 1968, p. 271–353.

Faisal et al. 2008 — *Faisal A. A., Selen L. P., Wolpert D. M*. Noise in the Nervous System // Nature Reviews Neuroscience, 2008, 9 (4), p. 292–303.

Faisal et al. 2005 — *Faisal A. A., White J. A., Laughlin S. B*. Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring // Current Biology, 2005, 15 (12), p. 1143–1149.

Feldman 2000 — *Feldman Jacob*. Minimization of Boolean Complexity in Human Concept Learning // Nature, 2000, 407 (6804), p. 630–633.

Feldman, Ballard 1982 — *Feldman J. A., Ballard Dana H*. Connectionist Models and Their Properties // Cognitive Science, 1982, 6 (3), p. 205–254.

Foley et al. 2007 — *Foley J. A., Monfreda C., Ramankutty N., Zaks D*. Our Share of the Planetary Pie // Proceedings of the National Academy of Sciences of the United States of America, 2007, 104 (31), p. 12585–12586.

Forgas et al. 2010 — The Psychology of Attitudes and Attitude Change (Sydney Symposium of Social Psychology) // Eds. Joseph P. Forgas, Joel Cooper, William D. Crano. New York: Psychology Press, 2010.

Frank 1999 — *Frank Robert H*. Luxury Fever: Why Money Fails to Satisfy in an Era of Excess. New York: Free Press, 1999.

Fredriksen 2012 — *Fredriksen Kaja Bonesmo*. Less Income Inequality and More Growth — Are They Compatible? Part 6: The Distribution of Wealth. Technical report, OECD Economics Department Working Papers 929. OECD Publishing, 2012.

Freitas 1980 — *Freitas Robert A., Jr*. A Self-Replicating Interstellar Probe // Journal of the British Interplanetary Society, 1980, 33, p. 251–264.

Freitas 2000 — *Freitas Robert A., Jr*. Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations // Foresight Institute, 2000, April (Retrieved 2013, July 28; http://www.foresight.org/nano/Ecophagy.html).

Freitas, Merkle 2004 — *Freitas Robert A., Jr., Merkle Ralph C*. Kinematic Self-Replicating Machines. Georgetown, TX: Landes Bioscience, 2004.

Gaddis 1982 — *Gaddis John Lewis*. Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy. New York: Oxford University Press, 1982.

Gammoned.net, 2012 — Snowie // Gammoned.net. (Archived version. Retrieved 2012, June 30; http://web.archive.org/web/20070920191840/http://www.gammoned.com/snowie.h).

Gates 1975 — *Gates Bill*. Software Contest Winners Announced // Computer Notes, 1975, 1 (2), p. 1.

Georgieff 2007 — *Georgieff Michael K*. Nutrition and the Developing Brain: Nutrient Priorities and Measurement // American Journal of Clinical Nutrition, 2007, 85 (2), p. 614S–620S.

Gianaroli 2000 — *Gianaroli Luca*. Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy // Supplement, Human Reproduction, 2000, 15 (4), p. 69–75.

Gilovich et al. 2002 — Heuristics and Biases: The Psychology of Intuitive Judgment / Eds. Thomas Gilovich, Dale Griffin, Daniel Kahneman. New York: Cambridge University Press, 2002.

Gilster 2012 — *Gilster Paul*. ESO: Habitable Red Dwarf Planets Abundant // Centauri Dreams (blog), 2012, March 29.

Goldstone 1980 — *Goldstone Jack A*. Theories of Revolution: The Third Generation // World Politics, 1980, 32 (3), p. 425–453.

Goldstone 2001 — *Goldstone Jack A*. Towards a Fourth Generation of Revolutionary Theory // Annual Review of Political Science, 2001, 4, p. 139–187.

Good 1965 — *Good Irving John*. Speculations Concerning the First Ultraintelligent Machine // Advances in Computers / Eds. Franz L. Alt, Morris Rubinoff. New York: Academic Press, 1965, 6, p. 31–88.

Good 1970 — *Good Irving John*. Some Future Social Repercussions of Computers // International Journal of Environmental Studies, 1970, 1 (1–4), p. 67–79.

Good 1976 — *Good Irving John*. Book review of «The Thinking Computer: Mind Inside Matter» // International Journal of Man-Machine Studies, 1976, 8, p. 617–620.

Good 1982 — *Good Irving John*. Ethical Machines // Intelligent Systems: Practice and Perspective / Eds. J. E. Hayes, Donald Michie, Y.-H. Pao. Machine Intelligence 10. Chichester: Ellis Horwood, 1982, p. 555–560.

Goodman 1954 — *Goodman Nelson*. Fact, Fiction, and Forecast. 1st ed. London: Athlone Press, 1954.

Gott et al. 2005 — *Gott J. R., Juric M., Schlegel D., Hoyle F., Vogeley M., Tegmark M., Bahcall N., Brinkmann J*. A Map of the Universe // Astrophysical Journal, 2005, 624 (2), p. 463–483.

Gottfredson 2002 — *Gottfredson Linda S*. G: Highly General and Highly Practical // The General Factor of Intelligence: How General Is It? / Eds. Robert J. Sternberg, Elena L. Grigorenko. Mahwah, NJ: Lawrence Erlbaum, 2002, p. 331–380.

Gould 1990 — *Gould S. J*. Wonderful Life: The Burgess Shale and the Nature of History. New York: Norton, 1990.

Graham 1997 — *Graham Gordon*. The Shape of the Past: A Philosophical Approach to History. New York: Oxford University Press, 1997.

Gray, McCormick 1996 — *Gray C. M., McCormick D. A*. Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex // Science, 1996, 274 (5284), p. 109–113.

Greene 2012 — *Greene Kate*. Intel's Tiny Wi-Fi Chip Could Have a Big Impact // MIT Technology Review, 2012, September 21.

Guizzo 010 — *Guizzo Erico*. World Robot Population Reaches 8.6 Million // IEEE Spectrum, 2010, April 14.

Gunn 1982 — *Gunn James E*. Isaac Asimov: The Foundations of Science Fiction. Science-Fiction Writers. New York: Oxford University Press, 1982.

Haberl et al. 2007 — *Haberl Helmut, Erb Karl-Heinz, Krausmann Fridolin*. Global Human Appropriation of Net Primary Production (HANPP) // Encyclopedia of Earth, 2013, September 3.

Haberl et al. 2007 — *Haberl H., Erb K. H., Krausmann F., Gaube V., Bondeau A., Plutzar C., Gingrich S., Lucht W., Fischer-Kowalski M*. Quantifying and Mapping the Human Appropriation of Net Primary Production in Earth's Terrestrial Ecosystems // Proceedings of the National Academy of Sciences of the United States of America, 2007, 104 (31), p. 12942–12947.

Hájek 2009 — *Hájek Alan*. Dutch Book Arguments // The Oxford Handbook of Rational and Social Choice / Eds. Anand Paul, Pattanaik Prasanta, Puppe Clemens. New York: Oxford University Press, 2009, p. 173–195.

Hall 2007 — *Hall John Storrs*. Beyond AI: Creating the Conscience of the Machine. Amherst, NY: Prometheus Books, 2007.

Hampson et al. 2012 — *Hampson R. E., Song D., Chan R. H., Sweatt A. J., Riley M. R., Gerhardt G. A., Shin D. C., Marmarelis V. Z., Berger T. W., Deadwyler S. A*. A Nonlinear Model for Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal Ensemble Stimulation // IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2012, 20 (2), p. 184–197.

Hanson 1994 — *Hanson Robin*. If Uploads Come First: The Crack of a Future Dawn // Extropy, 1994, 6 (2).

Hanson 1995 — *Hanson Robin*. Could Gambling Save Science? Encouraging an Honest Consensus // Social Epistemology, 1995, 9 (1), p. 3–33.

Hanson 1998 a — *Hanson Robin*. Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization // Unpublished manuscript, 1998, July 1 (Retrieved 2012, April 26; http://hanson.gmu.edu/filluniv.pdf).

Hanson 1998 b — *Hanson Robin*. Economic Growth Given Machine Intelligence, 1998 (Unpublished manuscript. Retrieved 2013, May 15; http://hanson.gmu.edu/aigrow.pdf).

Hanson 1998 c — *Hanson Robin*. Long-Term Growth as a Sequence of Exponential Modes, 1998 (Unpublished manuscript. Last revised 2000, December; http://hanson.gmu.edu/longgrow.pdf).

Hanson 1998 d — *Hanson Robin*. Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions, 1998, September 23 (Unpublished manuscript. Retrieved 2012, August 12; http://hanson.gmu.edu/hardstep.pdf).

Hanson 2000 — *Hanson Robin*. Shall We Vote on Values, But Bet on Beliefs? 2000, September (Unpublished manuscript. Last revised 2007, October; http://hanson.gmu.edu/futarchy.pdf).

Hanson 2006 — *Hanson Robin*. Uncommon Priors Require Origin Disputes // Theory and Decision, 2006, 61 (4), p. 319–328.

Hanson 2008 — *Hanson Robin*. Economics of the Singularity // IEEE Spectrum, 2008, 45 (6), p. 45–50.

Hanson 2009 — *Hanson Robin*. Tiptoe or Dash to Future? // Overcoming Bias (blog), 2009, December 23.

Hanson 2012 — *Hanson Robin*. Envisioning the Economy, and Society, of Whole Brain Emulations. Paper presented at the AGI Impacts conference, 2012.

Hart 2008 — *Hart Oliver*. Economica Coase Lecture Reference Points and the Theory of the Firm // Economica, 2008, 75 (299), p. 404–411.

Hay 2005 — *Hay Nicholas James*. Optimal Agents // B.Sc. thesis, University of Auckland, 2005.

Hedberg 2002 — *Hedberg Sara Reese*. Dart: Revolutionizing Logistics Planning // IEEE Intelligent Systems, 2002, 17 (3), p. 81–83.

Helliwell et al. 2012 — *Helliwell John, Layard Richard, Sachs Jeffrey*. World Happiness Report. The Earth Institute. 2012.

Helmstaedter et al. 2011 — *Helmstaedter M., Briggman K. L., Denk W*. High-Accuracy Neurite Reconstruction for High-Throughput Neuroanatomy // Nature Neuroscience, 2011, 14 (8), p. 1081–1088.

Heyl 2005 — *Heyl Jeremy S*. The Long-Term Future of Space Travel // Physical Review D, 2005, 72 (10), p. 1–4.

Hibbard 2011 — *Hibbard Bill*. Measuring Agent Intelligence via Hierarchies of Environments // Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings / Eds. Jürgen Schmidhuber, Kristinn R. Thórisson, Moshe Looks. Lecture Notes in Computer Science 6830. Berlin: Springer, 2011, p. 303–308.

Hinke et al. 1993 — *Hinke R. M., Hu X., Stillman A. E., Herkle H., Salmi R., Ugurbil K*. Functional Magnetic Resonance Imaging of Broca's Area During Internal Speech // Neuroreport, 1993, 4 (6), p. 675–678.

Hinxton Group, 2008 — Consensus Statement: Science, Ethics and Policy Challenges of Pluripotent Stem Cell-Derived Gametes. Hinxton, Cambridgeshire, UK, 2008, April 11 (http://www.hinxtongroup.org/Consensus_HG08_FINAL.pdf).

Hoffman 2009 — *Hoffman David E*. The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy. New York: Doubleday, 2009.

Hofstadter 1999 — *Hofstadter Douglas R*. Gödel, Escher, Bach: An Eternal Golden Braid. New York: Basic Books, 1999 [1979].

Holley 2009 — *Holley Rose*. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs // D-Lib Magazine, 2009, 15 (3–4).

Horton et al. 2008 — *Horton Sue, Alderman Harold, Rivera Juan A*. Copenhagen Consensus 2008 Challenge Paper: Hunger and Malnutrition. Technical report. Copenhagen Consensus Center, 2008, May 11.

Howson, Urbach 1993 — *Howson Colin, Urbach Peter*. Scientific Reasoning: The Bayesian Approach. 2nd ed. Chicago: Open Court, 1993.

Hsu 2012 — *Hsu Stephen*. Investigating the Genetic Basis for Intelligence and Other Quantitative Traits. Lecture given at UC Davis Department of Physics Colloquium, Davis, CA, 2012, February 13.

Huebner 2008 — *Huebner Bryce*. Do You See What We See? An Investigation of an Argument Against Collective Representation // Philosophical Psychology, 2008, 21 (1), p. 91–112.

Huff et al. 2010 — *Huff C. D., Xing J., Rogers A. R., Witherspoon D., Jorde L. B*. Mobile Elements Reveal Small Population Size in the Ancient Ancestors of *Homo Sapiens* // Proceedings of the National Academy of Sciences of the United States of America, 2010, 107 (5), p. 2147–2152.

Huffman, Pless 2003 — *Huffman W. Cary, Pless Vera*. Fundamentals of Error-Correcting Codes. New York: Cambridge University Press, 2003.

Hunt 2011 — *Hunt Patrick*. Late Roman Silk: Smuggling and Espionage in the 6th Century CE // Philolog, Stanford University (blog), 2011, August 2.

Hutter 2001 — *Hutter Marcus*. Towards a Universal Theory of Artificial Intelligence Based on Algorithmic Probability and Sequential Decisions // Machine Learning: ECML 2001: 12th European Conference on Machine Learning, Freiburg, Germany, September 5–7, 2001. Proceedings / Eds. Luc De Raedt, Peter Flach. Lecture Notes in Computer Science 2167. New York: Springer, 2001, p. 226–238.

Hutter 2005 — *Hutter Marcus*. Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability. Texts in Theoretical Computer Science. Berlin: Springer, 2005.

Iliadou et al. 2011 — *Iliadou A. N., Janson P. C., Cnattingius S*. Epigenetics and Assisted Reproductive Technology // Journal of Internal Medicine, 2011, 270 (5), p. 414–420.

Isaksson 2007 — *Isaksson Anders*. Productivity and Aggregate Growth: A Global Picture. Technical report 05/2007. Vienna, Austria: UNIDO (United Nations Industrial Development Organization) Research and Statistics Branch, 2007.

Jones 2009 — *Jones Garret*. Artificial Intelligence and Economic Growth: A Few Finger-Exercises, 2009, January (Unpublished manuscript. Retrieved 2012, November 5; http://mason.gmu.edu/~gjonesb/AIandGrowth).

Jones 1985 — *Jones Vincent C*. Manhattan: The Army and the Atomic Bomb. United States Army in World War II. Washington, DC: Center of Military History, 1985.

Joyce 1999 — *Joyce James M*. The Foundations of Causal Decision Theory. Cambridge Studies in Probability, Induction and Decision Theory. New York: Cambridge University Press, 1999.

Judd et al. 2012 — *Judd K. L., Schmedders K., Yeltekin S*. Optimal Rules for Patent Races // International Economic Review, 2012, 53 (1), p. 23–52.

Kalfoglou et al. 2004 — *Kalfoglou A., Suthers K., Scott J., Hudson K*. Reproductive Genetic Testing: What America Thinks. Genetics and Public Policy Center, 2004.

Kamm 2007 — *Kamm Frances M*. Intricate Ethics: Rights, Responsibilities, and Permissible Harm. Oxford Ethics Series. New York: Oxford University Press, 2007.

Kandel et al. 2000 — Principles of Neural Science / Eds. Eric Kandel, James Schwartz, Thomas Jessell. 4th ed. New York: McGraw-Hill, 2000.

Kansa 2003 — *Kansa Eric*. Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox, 2003 (Unpublished manuscript, archived version).

Karnofsky 2012 — *Karnofsky Holden*. Comment on «Reply to Holden on Tool AI» // Less Wrong (blog), 2012, August 1.

Kasparov 1996 — *Kasparov Garry*. The Day That I Sensed a New Kind of Intelligence // Time, 1996, March 25, no. 13.

Kaufman 2011 — *Kaufman Jeff*. Whole Brain Emulation and Nematodes // Jeff Kaufman's Blog (blog), 2011, November 2.

Keim et al. 1999 — *Keim G. A., Shazeer N. M., Littman M. L., Agarwal S., Cheves C. M., Fitzgerald J., Grosland J., Jiang F., Pollard S., Weinmeister K*. Proverb: The Probabilistic Cruciverbalist // Proceedings of the Sixteenth National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1999, p. 710–717.

Kell et al. 2013 — Kell Harrison J., Lubinski David, Benbow Camilla P. Who Rises to the Top? Early Indicators // Psychological Science, 2013, 24 (5), p. 648–659.

Keller 2004 — *Keller Wolfgang*. International Technology Diffusion // Journal of Economic Literature, 2004, 42 (3), p. 752–782.

KGS, 2012 — KGS Game Archives: Games of KGS player zen19 // KGS Go Server, 2012 (Retrieved 2013, July 22; http://www.gokgs.com/gameArchives.jsp? user=zen19d&oldAccounts=t&year=2012 &month=3).

Knill et al. 2000 — *Knill Emanuel, Laflamme Raymond, Viola Lorenza*. Theory of Quantum Error Correction for General Noise // Physical Review Letters, 2000, 84 (11), p. 2525–2528.

Koch et al. 2006 — *Koch K., McLean J., Segev R., Freed M. A., Berry M. J., Balasubramanian V., Sterling P*. How Much the Eye Tells the Brain // Current Biology, 2006, 16 (14), p. 1428–1434.

Kong et al. 2012 — *Kong A., Frigge M. L., Masson G., Besenbacher S., Sulem P., Magnusson G., Gudjonsson S. A., Sigurdsson A*. Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk // Nature, 2012, 488, p. 471–475.

Koomey 2011 — *Koomey Jonathan G*. Growth in Data Center Electricity Use 2005 to 2010. Technical report, 08/01/2011. Oakland, CA: Analytics Press, 2011.

Koubi 1999 — *Koubi Vally*. Military Technology Races // International Organization, 1999, 53 (3), p. 537–565.

Koubi, Lalman 2007 — *Koubi Vally*, *Lalman David*. Distribution of Power and Military R&D // Journal of Theoretical Politics, 2007, 19 (2), p. 133–152.

Koza et al. 2003 — *Koza J. R., Keane M. A., Streeter M. J., Mydlowec W., Yu J., Lanza G.* Genetic Programming IV: Routine Human-Competitive Machine Intelligence. 2nd ed. Genetic Programming. Norwell, MA: Kluwer Academic, 2003.

Kremer 1993 — *Kremer Michael*. Population Growth and Technological Change: One Million B.C. to 1990 // Quarterly Journal of Economics, 1993, 108 (3), p. 681–716.

Kruel 2011 — *Kruel Alexander*. Interview Series on Risks from AI // Less Wrong Wiki (blog), 2011 (Retrieved 2013, Oct. 26; http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI).

Kruel 2012 — *Kruel Alexander*. Q&A with Experts on Risks From AI #2 // Less Wrong (blog), 2012, January 9.

Krusienski, Shih 2011 — *Krusienski D. J., Shih J. J.* Control of a Visual Keyboard Using an Electrocorticographic Brain-Computer Interface // Neurorehabilitation and Neural Repair, 2011, 25 (4), p. 323–331.

Kuhn 1962 — *Kuhn Thomas S*. The Structure of Scientific Revolutions. 1st ed. Chicago: University of Chicago Press, 1962.

Kuipers 2012 — *Kuipers Benjamin*. An Existing, Ecologically-Successful Genus of Collectively Intelligent Artificial Creatures. Paper presented at the 4th International Conference, ICCCI 2012, Ho Chi Minh City, Vietnam, 2012, November 28–30.

Kurzweil 2001 — *Kurzweil Ray*. Response to Stephen Hawking // Kurzweil Accelerating Intelligence, 2001, September 5 (Retrieved 2012, December 31; http://www.kurzweilai.net/response-to-stephen-hawking).

Kurzweil 2005 — *Kurzweil Ray*. The Singularity Is Near: When Humans Transcend Biology. New York: Viking, 2005.

Laffont, Martimort 2002 — *Laffont Jean-Jacques, Martimort David*. The Theory of Incentives: The Principal-Agent Model. Princeton, NJ: Princeton University Press, 2002.

Lancet, 2008 — Iodine Deficiency — Way to Go Yet // The Lancet, 2008, 372 (9633), p. 88.

Landauer 1986 — *Landauer Thomas K*. How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-Term Memory // Cognitive Science, 1986, 10 (4), p. 477–493.

Lebedev 2004 — *Lebedev Anastasiya*. The Man Who Saved the World Finally Recognized // MosNews, 2004, May 21.

Lebedev, Nicolelis 2006 — *Lebedev M. A., Nicolelis M. A*. Brain-Machine Interfaces: Past, Present and Future // Trends in Neuroscience, 2006, 29 (9), p. 536–546.

Legg 2008 — *Legg Shane*. Machine Super Intelligence. PhD diss., University of Lugano, 2008.

Leigh 2010 — *Leigh E. G., Jr*. The Group Selection Controversy // Journal of Evolutionary Biology, 2010, 23 (1), p. 6–19.

Lenat 1982 — *Lenat Douglas B*. Learning Program Helps Win National Fleet Wargame Tournament // SIGART Newsletter, 1982, 79, p. 16–17.

Lenat 1983 — *Lenat Douglas B*. EURISKO: A Program that Learns New Heuristics and Domain Concepts // Artificial Intelligence, 1983, 21 (1–2), p. 61–98.

Lenman 2000 — *Lenman James*. Consequentialism and Cluelessness // Philosophy & Public Affairs, 2000, 29 (4), p. 342–370.

Lerner 1997 — *Lerner Josh*. An Empirical Exploration of a Technology Race // RAND Journal of Economics, 1997, 28 (2), p. 228–247.

Leslie 1996 — *Leslie John*. The End of the World: The Science and Ethics of Human Extinction. London: Routledge, 1996.

Lewis 1988 — *Lewis David*. Desire as Belief // Mind: A Quarterly Review of Philosophy, 1988, 97 (387), p. 323–332.

Li, Vitanyi 2008 — *Li Ming, Vitaanyi Paul M. B*. An Introduction to Kolmogorov Complexity and Its Applications. Texts in Computer Science. New York: Springer, 2008.

Lin et al. 2012 — *Lin Thomas, Mausam, Etzioni Oren*. Entity Linking at Web Scale // Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX '12) / Eds. James Fan, Raphael Hoffman, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek, Partha Pratim Talukdar. Madison, WI: Omnipress, 2012, p. 84–88.

Lloyd 2000 — *Lloyd Seth*. Ultimate Physical Limits to Computation // Nature, 2000, 406 (6799), p. 1047–1054.

Louis Harris & Associates, 1969 — Louis Harris & Associates. Science, Sex, and Morality Survey, study no. 1927 // Life Magazine. 1969, (New York) 4.

Lynch 2010 — *Lynch Michael*. Rate, Molecular Spectrum, and Consequences of Human Mutation // Proceedings of the National Academy of Sciences of the United States of America, 2010, 107 (3), p. 961–968.

Lyons 2011 — *Lyons Mark K*. Deep Brain Stimulation: Current and Future Clinical Applications // Mayo Clinic Proceedings, 2011, 86 (7), p. 662–672.

MacAskill 2010 — *MacAskill William*. Moral Uncertainty and Intertheoretic Comparisons of Value // BPhil thesis, University of Oxford, 2010.

McCarthy 2007 — *McCarthy John*. From Here to Human-Level AI // Artificial Intelligence, 2007, 171 (18), p. 1174–1182.

McCorduck 1979 — *McCorduck Pamela*. Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence. San Francisco: W. H. Freeman, 1979.

Mack 2011 — *Mack C. A*. Fifty Years of Moore's Law // IEEE Transactions on Semiconductor Manufacturing, 2011, 24 (2), p. 202–207.

MacKay 2003 — *MacKay David J. C*. Information Theory, Inference, and Learning Algorithms. New York: Cambridge University Press, 2003.

McLean, Stewart 1979 — *McLean George, Stewart Brian*. Norad False Alarm Causes Uproar // The National, 1979, Aired November 10. Ottawa, ON: CBC, 2012. News Broadcast.

Maddison 1999 — *Maddison Angus*. Economic Progress: The Last Half Century in Historical Perspective // Facts and Fancies of Human Development: Annual Symposium and Cunningham Lecture, 1999 / Ed. Ian Castles. Occasional Paper Series, 1/2000. Academy of the Social Sciences in Australia, 1999.

Maddison 2001 — *Maddison Angus*. The World Economy: A Millennial Perspective. Development Centre Studies. Paris: Development Centre of the Organisation for Economic Co-operation / Development, 2001.

Maddison 2005 — *Maddison Angus*. Growth and Interaction in the World Economy: The Roots of Modernity. Washington, DC: AEI Press, 2005.

Maddison 2007 — *Maddison Angus*. Contours of the World Economy, l-2030 AD: Essays in Macro-Economic History. New York: Oxford University Press, 2007.

Maddison 2010 — *Maddison Angus*. Statistics of World Population, GDP and Per Capita GDP 1–2008 AD, 2010 (Retrieved 2013 October 26; http://www.ggdc.net/maddison/Historical_Statistics/vertical-file_02–2010.xls).

Mai et al. 2007 — *Mai Q., Yu Y., Li T., Wang L., Chen M. J., Huang S. Z., Zhou C., Zhou Q*. Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts // Cell Research, 2007, 17 (12), p. 1008–1019.

Mak, Wolpaw 2009 — *Mak J. N., Wolpaw J. R*. Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects // IEEE Reviews in Biomedical Engineering, 2009, 2, p. 187–199.

Mankiw 2009 — *Mankiw N. Gregory*. Macroeconomics. 7th ed. New York, NY: Worth, 2009.

Mardis 2011 — *Mardis Elaine R*. A Decade's Perspective on DNA Sequencing Technology // Nature, 2011, 470 (7333), p. 198–203.

Markoff 2011 — *Markoff John*. Computer Wins on «Jeopardy!»: Trivial, It's Not // New York Times, 2011, February 16.

Markram 2006 — *Markram Henry*. The Blue Brain Project // Nature Reviews Neuroscience, 2006, 7 (2), p. 153–160.

Mason 2003 — *Mason Heather*. Gallup Brain: The Birth of In Vitro Fertilization // Gallup, 2003, August 5.

Menzel, Giurfa 2001 — *Menzel Randolf, Giurfa Martin*. Cognitive Architecture of a Mini-Brain: The Honeybee // Trends in Cognitive Sciences, 2001, 5 (2), p. 62–71.

Metzinger 2003 — *Metzinger Thomas*. Being No One: The Self-Model Theory of Subjectivity. Cambridge, MA: MIT Press, 2003.

Mijic 2010 — *Mijic Roko*. Bootstrapping Safe AGI Goal Systems. Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, 2010, March 8.

Mike 2013 — *Mike Mike*. Face of Tomorrow, 2013 (http://faceoftomorrow.org).

Milgrom, Roberts 1990 — *Milgrom Paul, Roberts John*. Bargaining Costs, Influence Costs, and the Organization of Economic Activity // Perspectives on Positive Political Economy / Eds. James E. Alt, Kenneth A. Shepsle. New York: Cambridge University Press, 1990, p. 57–89.

Miller 1956 — *Miller George A*. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information // Psychological Review, 1956, 63 (2), p. 81–97.

Miller 2000 — *Miller Geoffrey*. The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature. New York: Doubleday, 2000.

Miller 2012 — *Miller James D*. Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World. Dallas, TX: BenBella Books, 2012.

Minsky 1967 — *Minsky Marvin*. Computation: Finite and Infinite Machines. Englewood Cliffs, NJ: Prentice-Hall, 1967.

Minsky 1968 — Semantic Information Processing / Ed. Marvin Minsky. Cambridge, MA: MIT Press, 1968.

Minsky 1984 — *Minsky Marvin*. Afterword to Vernor Vinge's novel, «True Names», 1984 (Unpublished manuscript. Retrieved 2012, December 31; http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html).

Minsky 2006 — *Minsky Marvin*. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. New York: Simon & Schuster, 2006.

Minsky, Papert 1969 — *Minsky Marvin*, *Papert Seymour*. Perceptrons: An Introduction to Computational Geometry. 1st ed. Cambridge, MA: MIT Press, 1969.

Moore 2011 — *Moore Andrew*. Hedonism // The Stanford Encyclopedia of Philosophy / Ed. Edward N. Zalta. Stanford, CA: Stanford University, 2011.

Moravec 1976 — *Moravec Hans P*. The Role of Raw Power in Intelligence, 1976 (Unpublished manuscript. Retrieved 2012, August 12; http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Po).

Moravec 1980 — *Moravec Hans P*. Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover. PhD diss., Stanford University, 1980.

Moravec 1988 — *Moravec Hans P*. Mind Children: The Future of Robot and Human Intelligence. Cambridge, MA: Harvard University Press, 1988.

Moravec 1998 — *Moravec Hans P*. When Will Computer Hardware Match the Human Brain? // Journal of Evolution and Technology, 1998, 1.

Moravec 1999 — *Moravec Hans P*. Rise of the Robots // Scientific American, 1999, December, p. 124–135.

Muehlhauser, Helm 2012 — *Muehlhauser Luke, Helm Louie*. The Singularity and Machine Ethics // Singularity Hypotheses: A Scientific and Philosophical Assessment / Eds. Amnon Eden, Johnny Søraker, James H. Moor, Eric Steinhart. The Frontiers Collection. Berlin: Springer, 2012.

Muehlhauser, Salamon 2012 — *Muehlhauser Luke*, *Salamon Anna*. Intelligence Explosion: Evidence and Import // Singularity Hypotheses: A Scientific and Philosophical Assessment / Eds. Amnon Eden, Johnny Søraker, James H. Moor, Eric Steinhart. The Frontiers Collection. Berlin: Springer, 2012.

Müller, Bostrom <В печати> — *Müller Vincent C., Bostrom Nick*. Future Progress in Artificial Intelligence: A Poll Among Experts // Impacts and Risks of Artificial General Intelligence / Ed. Vincent C. Müller. Special issue, Journal of Experimental and Theoretical Artificial Intelligence (Forthcoming).

Murphy 2012 — *Murphy Kevin P*. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2012.

Nachman, Crowell 2000 — *Nachman Michael W., Crowell Susan L*. Estimate of the Mutation Rate per Nucleotide in Humans // Genetics, 2000, 156 (1), p. 297–304.

Nagy, Chang 2007 — *Nagy Z. P., Chang C. C*. Artificial Gametes // Theriogenology, 2007, 67 (1), p. 99–104.

Nagy et al. 2008 — *Nagy Z. P., Kerkis I., Chang C. C*. Development of Artificial Gametes // Reproductive BioMedicine Online, 2008, 16 (4), p. 539–544.

NASA, 2013 — International Space Station: Facts and Figures // NASA, 2013 (http://www.nasa.gov/worldbook/intspacestation_worldbook.html).

Newborn 2011 — *Newborn Monty*. Beyond Deep Blue: Chess in the Stratosphere. New York: Springer, 2011.

Newell et al. 1958 — *Newell Allen, Shaw J. C., Simon Herbert A*. Chess-Playing Programs and the Problem of Complexity // IBM Journal of Research and Development, 1958, 2 (4), p. 320–335.

Newell et al. 1959 — *Newell Allen, Shaw J. C., Simon, Herbert A*. Report on a General Problem-Solving Program: Proceedings of the International Conference on Information Processing // Information Processing. Paris: UNESCO, 1959, p. 256–264.

Nicolelis, Lebedev 2009 — *Nicolelis Miguel A. L., Lebedev Mikhail A*. Principles of Neural Ensemble Physiology Underlying the Operation of Brain-Machine Interfaces // Nature Reviews Neuroscience, 2009, 10 (7), p. 530–540.

Nilsson 1984 — *Nilsson Nils J*. Shakey the Robot, Technical Note 323. Menlo Park, CA: AI Center, SRI International, 1984, April.

Nilsson 2009 — *Nilsson Nils J*. The Quest for Artificial Intelligence: A History of Ideas and Achievements. New York: Cambridge University Press, 2009.

Nisbett et al. 2012 — *Nisbett R. E., Aronson J., Blair C., Dickens W., Flynn J., Halpern D. F., Turkheimer E*. Intelligence: New Findings and Theoretical Developments // American Psychologist, 2012, 67 (2), p. 130–159.

Niven 1973 — *Niven Larry*. The Defenseless Dead // Ten Tomorrows / Ed. Roger Elwood. New York: Fawcett, 1973, p. 91–142.

Nordhaus 2007 — *Nordhaus William D*. Two Centuries of Productivity Growth in Computing // Journal of Economic History, 2007, 67 (1), p. 128–159.

Norton 2011 — *Norton John D*. Waiting for Landauer // Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 2011, 42 (3), p. 184–198.

Olds, Milner 1954 — *Olds James, Milner Peter*. Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain // Journal of Comparative and Physiological Psychology, 1954, 47 (6), p. 419–427.

Olum 2002 — *Olum Ken D*. The Doomsday Argument and the Number of Possible Observers // Philosophical Quarterly, 2002, 52 (207), p. 164–184.

Omohundro 2007 — *Omohundro Stephen M*. The Nature of Self-Improving Artificial Intelligence. Paper presented at Singularity Summit 2007, San Francisco, CA, 2007, September 8–9.

Omohundro 2008 — *Omohundro Stephen M*. The Basic AI Drives // Artificial General Intelligence 2008: Proceedings of the First AGI Conference / Eds. Pei Wang, Ben Goertzel, Stan Franklin. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS, 2008, p. 483–492.

Omohundro 2012 — *Omohundro Stephen M*. Rational Artificial Intelligence for the Greater Good // Singularity Hypotheses: A Scientific and Philosophical Assessment / Eds. Amnon Eden, Johnny Søraker, James H. Moor, Eric Steinhart. The Frontiers Collection. Berlin: Springer, 2012.

O'Neill 1974 — *O'Neill Gerard K*. The Colonization of Space // Physics Today, 1974, 27 (9), p. 32–40.

Oshima, Katayama 2010 — *Oshima Hideki, Katayama Yoichi*. Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans // Neurologia medico-chirurgica, 2010, 50 (9), p. 845–852.

Parfit 1986 — *Parfit Derek*. Reasons and Persons. New York: Oxford University Press, 1986.

Parfit 2011 — *Parfit Derek*. On What Matters. 2 vols. The Berkeley Tanner Lectures. New York: Oxford University Press, 2011.

Parrington 1997 — Parrington, Alan J. 1997. "Mutually Assured Destruction Revisited." Airpower Journal 11 (4).

Pasqualotto et al. 2012 — *Pasqualotto Emanuele, Federici Stefano, Belardinelli Marta Olivetti*. Toward Functioning and Usable Brain-Computer Interfaces (BCIs): A Literature Review // Disability and Rehabilitation: Assistive Technology, 2012, 7 (2), p. 89–103.

Pearl 2009 — *Pearl Judea*. Causality: Models, Reasoning, and Inference. 2nd ed. New York: Cambridge University Press, 2009.

Perlmutter, Mink 2006 — *Perlmutter J. S., Mink J. W*. Deep Brain Stimulation // Annual Review of Neuroscience, 2006, 29, p. 229–257.

Pinker 2011 — *Pinker Steven*. The Better Angels of Our Nature: Why Violence Has Declined. New York: Viking, 2011.

Plomin et al. 2013 — *Plomin R., Haworth C. M., Meaburn E. L., Price T. S.*, Wellcome Trust Case Control Consortium 2, *Davis O. S*. Common DNA Markers Can Account for More than Half of the Genetic Influence on Cognitive Abilities // Psychological Science, 2013, 24 (2), p. 562–568.

Popper 2012 — *Popper Nathaniel*. Flood of Errant Trades Is a Black Eye for Wall Street // New York Times, 2012, August 1.

Pourret et al. 2008 — Bayesian Networks: A Practical Guide to Applications / Eds. Olivier Pourret, Patrick Naim, Bruce Marcot. Chichester, West Sussex, UK: Wiley, 2008.

Powell et al. 2009 — *Powell A., Shennan S., Thomas M. G*. Late Pleistocene Demography and the Appearance of Modern Human Behavior // Science, 2009, 324 (5932), p. 1298–1301.

Price 1991 — *Price Huw*. Agency and Probabilistic Causality // British Journal for the Philosophy of Science, 1991, 42 (2), p. 157–176.

Qian et al. 2005 — *Qian M., Wang D., Watkins W. E., Gebski V., Yan Y. Q., Li M., Chen Z. P*. The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China // Asia Pacific Journal of Clinical Nutrition, 2005, 14 (1), p. 32–42.

Quine, Ullian 1978 — *Quine Willard Van Orman, Ullian Joseph Silbert*. The Web of Belief / Ed. Richard Malin Ohmann. Vol. 2. New York: Random House, 1978.

Railton 1986 — *Railton Peter*. Facts and Values // Philosophical Topics, 1986, 14 (2), p. 5–31.

Rajab et al. 2006 — *Rajab Moheeb Abu, Zarfoss Jay, Monrose Fabian, Terzis Andreas*. A Multifaceted Approach to Understanding the Botnet Phenomenon // Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement. New York: ACM, 2006, p. 41–52.

Rawls 1971 — *Rawls John*. A Theory of Justice. Cambridge, MA: Belknap, 1971.

Read, Trentham 2005 — *Read J. I., Trentham Neil*. The Baryonic Mass Function of Galaxies // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2005, 363 (1837), p. 2693–2710.

Repantis et al. 2010 — *Repantis D., Schlattmann P., Laisney O., Heuser I*. Modafinil and Methylphenidate for Neuroenhancement in Healthy Individuals: A Systematic Review // Pharmacological Research, 2010, 62 (3), p. 187–206.

Rhodes 1986 — *Rhodes Richard*. The Making of the Atomic Bomb. New York: Simon & Schuster, 1986.

Rhodes 2008 — *Rhodes Richard*. Arsenals of Folly: The Making of the Nuclear Arms Race. New York: Vintage, 2008.

Rietveld et al. 2013 — *Rietveld Cornelius A., Medland Sarah E., Derringer Jaime, Yang Jian, Esko Tonu, Martin Nicolas W., Westra Harm-Jan, Shakhbazov Konstantin, Abdellaoui Abdel*. GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment // Science, 2013, 340 (6139), p. 1467–1471.

Ring, Orseau 2011 — *Ring Mark, Orseau Laurent*. Delusion, Survival, and Intelligent Agents // Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings / Eds. Jürgen Schmidhuber, Kristinn R. Thórisson, Moshe Looks. Lecture Notes in Computer Science 6830. Berlin: Springer, 2011, p. 11–20.

Ritchie et al. 2007 — *Ritchie Graeme, Manurung Ruli, Waller Annalu*. A Practical Application of Computational Humour // Proceedings of the 4th International Joint Workshop on Computational Creativity / Eds. Amilcar Cardoso, Geraint A. Wiggins. London: Goldsmiths, University of London, 2007, p. 91–98.

Roache 2008 — *Roache Rebecca*. Ethics, Speculation, and Values // NanoEthics, 2008, 2 (3), p. 317–327.

Robles et al. 2008 — *Robles J. A., Lineweaver C. H., Grether D., Flynn C., Egan C. A., Pracy M. B., Holmberg J., Gardner E*. A Comprehensive Comparison of the Sun to Other Stars: Searching for Self-Selection Effects // Astrophysical Journal, 2008, 684 (1), p. 691–706.

Roe 1953 — *Roe Anne*. The Making of a Scientist. New York: Dodd, Mead, 1953.

Roy 2012 — *Roy Deb*. About, 2012 (Retrieved October 14; http://web.media.mit.edu/~dkroy/).

Rubin, Watson 2011 — *Rubin Jonathan, Watson Ian*. Computer Poker: A Review // Artificial Intelligence, 2011, 175 (5–6), p. 958–987.

Rumelhart et al. 1986 — *Rumelhart D. E., Hinton G. E., Williams R. J.* Learning Representations by Back-Propagating Errors // Nature, 1986, 323 (6088), p. 533–536.

Russell 1986 — *Russell Bertrand*. The Philosophy of Logical Atomism // The Philosophy of Logical Atomism and Other Essays 1914–1919 / Ed. John G. Slater. The Collected Papers of Bertrand Russell. Boston: Allen & Unwin, 1986, 8, p. 157–244.

Russell, Griffin 2001 — *Russell Bertrand, Griffin Nicholas*. The Selected Letters of Bertrand Russell: The Public Years, 1914–1970. New York: Routledge, 2001.

Russell, Norvig 2010 — *Russell Stuart J., Norvig Peter*. Artificial Intelligence: A Modern Approach. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

Sabrosky 1952 — *Sabrosky Curtis W*. How Many Insects Are There? // Insects / Ed. United States Department of Agriculture. Yearbook of Agriculture. Washington, DC: United States Government Printing Office, 1952, p. 1–7.

Salamon 2009 — *Salamon Anna*. When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth // Working Paper, 2009, December 27.

Salem, Rowan 2001 — *Salem D. J., Rowan A. N*. The State of the Animals: 2001. Public Policy Series. Washington, DC: Humane Society Press, 2001.

Salverda et al. 2009 — *Salverda W., Nolan B., Smeeding T. M*. The Oxford Handbook of Economic Inequality. Oxford: Oxford University Press, 2009.

Samuel 1959 — *Samuel A. L*. Some Studies in Machine Learning Using the Game of Checkers // IBM Journal of Research and Development, 1959, 3 (3), p. 210–219.

Sandberg 1999 — *Sandberg Anders*. The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains // Journal of Evolution and Technology, 1999, 5.

Sandberg 2010 — *Sandberg Anders*. An Overview of Models of Technological Singularity. Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, 2010, March 8.

Sandberg 2013 — *Sandberg Anders*. Feasibility of Whole Brain Emulation // Philosophy and Theory of Artificial Intelligence / Ed. Vincent C. Müller. Studies in Applied Philosophy, Epistemology and Rational Ethics. New York: Springer, 2013, 5, p. 251–264.

Sandberg, Bostrom 2006 — *Sandberg Anders*, *Bostrom Nick*. Converging Cognitive Enhancements // Annals of the New York Academy of Sciences, 2006, 1093, p. 201–227.

Sandberg, Bostrom 2008 — *Sandberg Anders*, *Bostrom Nick*. Whole Brain Emulation: A Roadmap. Technical Report 2008–3. Future of Humanity Institute, University of Oxford, 2008.

Sandberg, Bostrom 2011 — *Sandberg Anders*, *Bostrom Nick*. Machine Intelligence Survey. Technical Report 2011–1. Future of Humanity Institute, University of Oxford, 2011.

Sandberg, Savulescu 2011 — *Sandberg Anders*, *Savulescu Julian*. The Social and Economic Impacts of Cognitive Enhancement // Enhancing Human Capacities / Eds. Julian Savulescu, Ruud ter Meulen, Guy Kahane. Malden, MA: Wiley-Blackwell, 2011, 92–112.

Schaeffer 1997 — *Schaeffer Jonathan*. One Jump Ahead: Challenging Human Supremacy in Checkers. New York: Springer, 1997.

Schaeffer et al. 2007 — *Schaeffer J., Burch N., Bjornsson Y., Kishimoto A., Muller M., Lake R., Lu P., Sutphen S*. Checkers Is Solved // Science, 2007, 317 (5844), p. 1518–1522.

Schalk 2008 — *Schalk Gerwin*. Brain-Computer Symbiosis // Journal of Neural Engineering, 2008, 5 (1), p. P1–P15.

Schelling 1980 — *Schelling Thomas C*. The Strategy of Conflict. 2nd ed. Cambridge, MA: Harvard University Press, 1980.

Schultz 2000 — *Schultz T. R*. In Search of Ant Ancestors // Proceedings of the National Academy of Sciences of the United States of America, 2000, 97 (26), p. 14028–14029.

Schultz et al. 1997 — *Schultz W., Dayan P., Montague P. R*. A Neural Substrate of Prediction and Reward // Science, 1997, 275 (5306), p. 1593–1599.

Schwartz 1987 — *Schwartz Jacob T*. Limits of Artificial Intelligence // Encyclopedia of Artificial Intelligence / Eds. Stuart C. Shapiro, David Eckroth. New York: Wiley, 1987, 1, p. 488–503.

Schwitzgebel 2013 — *Schwitzgebel Eric*. If Materialism is True, the United States is Probably Conscious // Working Paper, 2013, February 8.

Sen, Williams 1982 — Utilitarianism and Beyond / Eds. Amartya Sen, Bernard Williams. New York: Cambridge University Press, 1982.

Shanahan 2010 — *Shanahan Murray*. Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds. New York: Oxford University Press, 2010.

Shannon 2012 — *Shannon Robert V*. Advances in Auditory Prostheses // Current Opinion in Neurology, 2012, 25 (1), p. 61–66.

Shapiro 1992 — *Shapiro Stuart C*. Artificial Intelligence // Encyclopedia of Artificial Intelligence. 2nd ed. New York: Wiley, 1992, 1, p. 54–57.

Sheppard 2002 — *Sheppard Brian*. World-Championship-Caliber Scrabble // Artificial Intelligence, 2002, 134 (1–2), p. 241–275.

Shoemaker 1969 — *Shoemaker Sydney*. Time Without Change // Journal of Philosophy, 1969, 66 (12), p. 363–381.

Shulman 2010 a — *Shulman Carl*. Omohundro's «Basic AI Drives» and Catastrophic Risks. San Francisco, CA: Machine Intelligence Research Institute, 2010.

Shulman 2010 b — *Shulman Carl*. Whole Brain Emulation and the Evolution of Superorganisms. San Francisco, CA: Machine Intelligence Research Institute, 2010.

Shulman 2012 — *Shulman Carl*. Could We Use Untrustworthy Human Brain Emulations to Make Trustworthy Ones? Paper presented at the AGI Impacts conference, 2012.

Shulman, Bostrom 2012 — *Shulman Carl, Bostrom Nick*. How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects // Journal of Consciousness Studies, 2012, 19 (7–8), p. 103–130.

Shulman, Bostrom 2014 — *Shulman Carl, Bostrom Nick*. Embryo Selection for Cognitive Enhancement: Curiosity or Game-Changer? // Global Policy, 2014, 5 (1), p. 85–92.

Shulman et al. 2009 — *Shulman Carl, Jonsson Henrik, Tarleton Nick*. Which Consequentialism? Machine Ethics and Moral Divergence // AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy

Conference, October lst–2nd, University of Tokyo, Japan. Proceedings / Eds. Carson Reynolds, Alvaro Cassinelli. AP-CAP, 2009, p. 23–25.

Sidgwick, Jones 2010 — *Sidgwick Henry, Jones Emily Elizabeth Constance*. The Methods of Ethics. Charleston, SC: Nabu Press, 2010.

Silver 2006 — *Silver Albert*. How Strong Is GNU Backgammon? // Backgammon Galore! 2006, September 16 (Retrieved 2013, October 26; http://www.bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu).

Simeral et al. 2011 — *Simeral J. D., Kim S. P., Black M. J., Donoghue J. P., Hochberg L. R*. Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelectrode Array // Journal of Neural Engineering, 2011, 8 (2), p. 025–027.

Simester, Knez 2002 — *Simester Duncan, Knez Marc*. Direct and Indirect Bargaining Costs and the Scope of the Firm // Journal of Business, 2002, 75 (2), p. 283–304.

Simon 1965 — *Simon Herbert Alexander*. The Shape of Automation for Men and Management. New York: Harper & Row, 1965.

Sinhababu 2009 — *Sinhababu Neil*. The Humean Theory of Motivation Reformulated and Defended // Philosophical Review, 2009, 118 (4), p. 465–500.

Slagle 1963 — *Slagle James R*. A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus // Journal of the ACM, 1963, 10 (4), p. 507–520.

Smeding et al. 2006 — *Smeding H. M., Speelman J. D., Koning-Haanstra M., Schuurman P. R., Nijssen P., van Laar T., Schmand B*. Neuropsychological Effects of Bilateral STN Stimulation in Parkinson Disease: A Controlled Study // Neurology, 2006, 66 (12), p. 1830–1836.

Smith 1987 — *Smith Michael*. The Humean Theory of Motivation // Mind: A Quarterly Review of Philosophy, 1987, 96 (381), p. 36–61.

Smith et al. 1989 — *Smith Michael, Lewis David, Johnston Mark*. Dispositional Theories of Value // Proceedings of the Aristotelian Society, 1989, 63, p. 89–174.

Sparrow 2013 — *Sparrow Robert*. In Vitro Eugenics // Journal of Medical Ethics. doi:10.1136/medethics-2012–101200. Published online, 2013, April 4 (http://jme.bmj.com/content/early/2013/02/13/medethics-2012–101200.full).

Stansberry, Kudritzki 2012 — *Stansberry Matt, Kudritzki Julian*. Uptime Institute 2012 Data Center Industry Survey. Uptime Institute, 2012.

Stapledon 1937 — *Stapledon Olaf*. Star Maker. London: Methuen, 1937.

Steriade et al. 1998 — *Steriade M., Timofeev I., Durmuller N., Grenier F*. Dynamic Properties of Corticothalamic Neurons and Local Cortical Interneurons Generating Fast Rhythmic (30–40 Hz) Spike Bursts // Journal of Neurophysiology, 1998, 79 (1), p. 483–490.

Stewart et al. 2008 — *Stewart P. W., Lonky E., Reihman J., Pagano J., Gump B. B., Darvill T*. The Relationship Between Prenatal PCB Exposure and Intelligence (IQ) in 9-Year-Old Children // Environmental Health Perspectives, 2008, 116 (10), p. 1416–1422.

Sun et al 2012 — *Sun W., Yu H., Shen Y., Banno Y., Xiang Z., Zhang Z*. Phylogeny and Evolutionary History of the Silkworm // Science China Life Sciences, 2012, 55 (6), p. 483–496.

Sundet et al. 2004 — *Sundet J., Barlaug D., Torjussen T*. The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century // Intelligence, 2004, 32 (4), p. 349–362.

Sutton, Barto 1998 — *Sutton Richard S., Barto Andrew G*. Reinforcement Learning: An Introduction. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 1998.

Talukdar et al. 2002 — *Talukdar D., Sudhir K., Ainslie A*. Investigating New Product Diffusion Across Products and Countries // Marketing Science, 2002, 21 (1), p. 97–114.

Teasdale, Owen 2008 — *Teasdale Thomas W., Owen David R*. Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect // Intelligence, 2008, 36 (2), p. 121–126.

Tegmark, Bostrom 2005 — *Tegmark Max, Bostrom Nick*. Is a Doomsday Catastrophe Likely? // Nature, 2005, 438, p. 754.

Teitelman 1966 — *Teitelman Warren*. Pilot: A Step Towards Man-Computer Symbiosis. PhD diss., Massachusetts Institute of Technology, 1966.

Temple 1986 — *Temple Robert K. G*. The Genius of China: 3000 Years of Science, Discovery, and Invention. 1st ed. New York: Simon & Schuster, 1986.

Tesauro 1995 — *Tesauro Gerald*. Temporal Difference Learning and TD-Gammon // Communications of the ACM, 1995, 38 (3), p. 58–68.

Tetlock 2005 — *Tetlock Philip E*. Expert Political Judgment: How Good is it? How Can We Know? Princeton, NJ: Princeton University Press, 2005.

Tetlock, Belkin 1996 — *Tetlock Philip E., Belkin Aaron*. Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives // Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives / Eds. Philip E. Tetlock, Aaron Belkin. Princeton, NJ: Princeton University Press, 1996, p. 1–38.

Thompson 1997 — *Thompson Adrian*. Artificial Evolution in the Physical World // Evolutionary Robotics: From Intelligent Robots to Artificial Life / Ed. Takashi Gomi. Er '97. Carp, ON: Applied AI Systems, 1997, p. 101–125.

Thrun et al. 2006 — *Thrun S., Montemerlo M., Dahlkamp H., Stavens D., Aron A., Diebel J., Fong P.* Stanley: The Robot That Won the DARPA Grand Challenge // Journal of Field Robotics, 2006, 23 (9), p. 661–692.

Trachtenberg et al. 2002 — *Trachtenberg J. T., Chen B. E., Knott G. W., Feng G., Sanes J. R., Welker E., Svoboda K*. Long-Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex // Nature, 2002, 420 (6917), p. 788–794.

Traub 2012 — *Traub Wesley A*. Terrestrial, Habitable-Zone Exoplanet Frequency from Kepler // Astrophysical Journal, 2012, 745 (1), p. 1–10.

Truman et al. 1993 — *Truman James W., Taylor Barbara J., Awad Timothy A*. Formation of the Adult Nervous System // The Development of Drosophila Melanogaster / Eds. Michael Bate, Alfonso Martinez Arias. Plainview, NY: Cold Spring Harbor Laboratory, 1993.

Tuomi 2002 — *Tuomi Ilkka*. The Lives and the Death of Moore's Law // First Monday, 2002, 7 (11).

Turing 1950 — *Turing A. M*. Computing Machinery and Intelligence // Mind, 1950, 59 (236), p. 433–460.

Turkheimer et al. 2003 — *Turkheimer Eric, Haley Andreana, Waldron Mary, D'Onofrio Brian, Gottesman Irving I*. Socioeconomic Status Modifies Heritability of IQ in Young Children // Psychological Science, 2003, 14 (6), p. 623–628.

Uauy, Dangour 2006 — *Uauy Ricardo, Dangour Alan D*. Nutrition in Brain Development and Aging: Role of Essential Fatty Acids // Supplement, Nutrition Reviews, 2006, 64 (5), p. S24–S33.

Ulam 1958 — *Ulam Stanislaw M*. John von Neumann // Bulletin of the American Mathematical Society, 1958, 64 (3), p. 1–49.

Uncertain Future, 2012 — Frequently Asked Questions. The Uncertain Future (Retrieved 2012, March 25; http://www.theuncertainfuture.com/faq.html).

US Congress, 1995 — U.S. Congress, Office of Technology Assessment. U.S.-Russian Cooperation in Space ISS-618. Washington, DC: U.S. Government Printing Office, 1995, April.

Van Zanden 2003 — *Van Zanden Jan Luiten*. On Global Economic History: A Personal View on an Agenda for Future Research. International Institute for Social History, 2003, July 23.

Vardi 2012 — *Vardi Moshe Y*. Artificial Intelligence: Past and Future // Communications of the ACM, 2012, 55 (1), p. 5.

Vassar, Freitas 2006 — *Vassar Michael, Freitas Robert A., Jr*. Lifeboat Foundation Nanoshield. Lifeboat Foundation, 2006 (Retrieved 2012, May 12; http://lifeboat.com/ex/nanoshield).

Vinge 1993 — *Vinge Vernor*. The Coming Technological Singularity: How to Survive in the Post-Human Era // Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace. NASA Conference Publication 10129. NASA Lewis Research Center, 1993, p. 11–22.

Visscher et al. 2008 — *Visscher P. M., Hill W. G., Wray N. R*. Heritability in the Genomics Era: Concepts and Misconceptions // Nature Reviews Genetics, 2008, 9 (4), p. 255–266.

Vollenweider et al. 1998 — *Vollenweider Franz, Gamma Alex, Liechti Matthias, Huber Theo*. Psychological and Cardiovascular Effects and Short-Term Sequelae of MDMA («Ecstasy») in MDMA-Naïve Healthy Volunteers // Neuropsychopharmachology, 1998, 19 (4), p. 241–251.

Wade 1976 — *Wade Michael J*. Group Selections Among Laboratory Populations of Tribolium // Proceedings of the National Academy of Sciences of the United States of America, 1976, 73 (12), p. 4604–4607.

Wainwright, Jordan 2008 — *Wainwright Martin J., Jordan Michael I*. Graphical Models, Exponential Families, and Variational Inference // Foundations and Trends in Machine Learning, 2008, 1 (1–2), p. 1–305.

Walker 2002 — *Walker Mark*. Prolegomena to Any Future Philosophy // Journal of Evolution and Technology, 2002, 10 (1).

Walsh 2001 — *Walsh Nick Paton*. Alter our DNA or robots will take over, warns Hawking // The Observer, 2001, September 1 (http://www.theguardian.com/uk/2001/sep/02/medicalscience.genetics).

Warwick 2002 — *Warwick Kevin*. I, Cyborg. London: Century, 2002.

Wehner et al. 2008 — *Wehner M., Oliker L., Shalf J*. Towards Ultra-High Resolution Models of Climate and Weather // International Journal of High Performance Computing Applications, 2008, 22 (2), p. 149–165.

Weizenbaum 1966 — *Weizenbaum Joseph*. Eliza: A Computer Program for the Study of Natural Language Communication Between Man And Machine // Communications of the ACM, 1966, 9 (1), p. 36–45.

Weizenbaum 1976 — *Weizenbaum Joseph*. Computer Power and Human Reason: From Judgment to Calculation. San FrancYork, CA: W. H. Freeman, 1976.

Werbos 1994 — *Werbos Paul John*. The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting. New York: Wiley, 1994.

White et al. 1986 — *White J. G., Southgate E., Thomson J. N., Brenner S*. The Structure of the Nervous System of the Nematode Caenorhabditis Elegans // Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,1986, 314 (1165), p. 1–340.

Whitehead 2003 — *Whitehead Hal*. Sperm Whales: Social Evolution in the Ocean. Chicago: University of Chicago Press, 2003.

Whitman et al. 1998 — *Whitman William B., Coleman David C., Wiebe William J*. Prokaryotes: The Unseen Majority // Proceedings of the National Academy of Sciences of the United States of America, 1998, 95 (12), p. 6578–6583.

Wiener 1960 — *Wiener Norbert*. Some Moral and Technical Consequences of Automation // Science, 1960, 131 (3410), p. 1355–1358.

Wikipedia, 2012 a — s. v. «Computer Bridge» // Wikipedia, 2012 (Retrieved 2013, June 30.; http://en.wikipedia.org/wiki/Computer_bridge).

Wikipedia, 2012 b — s. v. «Supercomputer» // Wikipedia, 2012 (Retrieved 2013, June 30; http://et.wikipedia.org/wiki/Superarvuti).

Williams 1966 — *Williams George C*. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton Science Library. Princeton, NJ: Princeton University Press, 1966.

Winograd 1972 — *Winograd Terry*. Understanding Natural Language. New York: Academic Press, 1972.

Wood 2007 — *Wood Nigel*. Chinese Glazes: Their Origins, Chemistry and Re-creation. London: A. & C. Black, 2007.

World Bank, 2008 — Global Economic Prospects: Technology Diffusion in the Developing World. Washington, DC: The International Bank for Reconstruction and Development / The World Bank, 2008.

World Robotics, 2011 — Executive Summary of l. World Robotics 2011: Industrial Robots; 2. World Robotics 2011: Service Robots // World Robotics, 2011 (Retrieved 2012, June 30; http://www.bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.pdf).

World Values Survey, 2008 — WVS 2005–2008 // World Values Survey, 2008 (Retrieved 2013, October 29; http://www.wvsevsdb.com/wvs/WVSAnalizeStudy.jsp).

Wright 2001 — *Wright Robert*. Nonzero: The Logic of Human Destiny. New York: Vintage, 2001.

Yaeger 1994 — *Yaeger Larry*. Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context // Proceedings of the Artificial Life III Conference / Ed. C. G. Langton. Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley, 1994, p. 263–298.

Yudkowsky 2001 — *Yudkowsky Eliezer*. Creating Friendly AI l.0: The Analysis and Design of Benevolent Goal Architectures. San Francisco, CA: Machine Intelligence Research Institute, 2001, June 15.

Yudkowsky 2002 — *Yudkowsky Eliezer*. The AI-Box Experiment, 2002 (Retrieved 2012, January 15; http://yudkowsky.net/singularity/aibox).

Yudkowsky 2004 — *Yudkowsky Eliezer*. Coherent Extrapolated Volition. San Francisco, CA: Machine Intelligence Research Institute, 2004, May.

Yudkowsky 2007 — *Yudkowsky Eliezer*. Levels of Organization in General Intelligence // Artificial General Intelligence / Eds. Ben Goertzel, Cassio Pennachin. Cognitive Technologies. Berlin: Springer, 2007, p. 389–501.

Yudkowsky 2008 a — *Yudkowsky Eliezer*. Artificial Intelligence as a Positive and Negative Factor in Global Risk // Global Catastrophic Risks / Eds. Nick Bostrom, Milan M. Cirkovic. New York: Oxford University Press, 2008, p. 308–345.

Yudkowsky 2008 b — *Yudkowsky Eliezer*. Sustained Strong Recursion // Less Wrong (blog), 2008, December 5.

Yudkowsky 2010 — *Yudkowsky Eliezer*. Timeless Decision Theory. San Francisco, CA: Machine Intelligence Research Institute, 2010.

Yudkowsky 2011 — *Yudkowsky Eliezer*. Complex Value Systems are Required to Realize Valuable Futures. San Francisco, CA: Machine Intelligence Research Institute, 2011.

Yudkowsky 2013 — *Yudkowsky Eliezer*. Intelligence Explosion Microeconomics. Technical Report 2013–1. Berkeley, CA: Machine Intelligence Research Institute, 2013.

Zahavi, Zahavi 1997 — *Zahavi Amotz, Zahavi Avishag*. The Handicap Principle: A Missing Piece of Darwin's Puzzle / Transl. N. Zahavi-Ely, M. P. Ely. New York: Oxford University Press, 1997.

Zalasiewicz et al. 2008 — *Zalasiewicz J., Williams M., Smith A., Barry T. L., Coe A. L., Bown P. R., Brenchley P*. Are We Now Living in the Anthropocene? // GSA Today, 2008, 18 (2), p. 4–8.

Zeira 2011 — *Zeira Joseph*. Innovations, Patent Races and Endogenous Growth // Journal of Economic Growth, 2011, 16 (2), p. 135–156.

Zuleta 2008 — *Zuleta Hernando*. An Empirical Note on Factor Shares // Journal of International Trade and Economic Development, 2008, 17 (3), p. 379–390.

Список сокращений

FLOPS (floating-point operation per second) — количество операций с плавающей запятой в секунду

IQ (intelligence quotient) — коэффициент интеллекта, или коэффициент умственного развития

ВВП — валовой внутренний продукт

ИИ — искусственный интеллект

ИИЧУ — искусственный интеллект человеческого уровня

КИИ — классический искусственный интеллект

КИМГМ — компьютерная имитационная модель головного мозга

КЭВ — когерентное экстраполированное волеизъявление

МД — моральная допустимость

МП — моральная правота

ПО — программное обеспечение

УИИ — универсальный искусственный интеллект

УИИЧУ — универсальный искусственный интеллект человеческого уровня

ЭКО — экстракорпоральное оплодотворение

Благодарности

Защитная оболочка, которой окружает себя человек во время работы над книгой, остается довольно проницаемой. Многие концепции и идеи, родившиеся в процессе создания текста, каким-то образом просочились во внешний мир. В результате, став предметом широкого обсуждения, они вернулись ко мне в виде многочисленных соображений, конечно, нашедших свое место в окончательном варианте «Искусственного интеллекта». Я дотошно отслеживал, чтобы ни одно имя цитируемых мною авторов не было пропущено в аппарате книги, однако боюсь, что не смогу перечислить всех, кто оказал влияние на ход моих мыслей, — таких людей множество.

В первую очередь я благодарен всем за активное участие в обсуждении темы — это Росс Андерсен, Стюарт Армстронг, Оуэн Коттон-Баррат, Ник Бекстед, Дэвид Чалмерс, Пол Кристиано, Милан Чиркович, Дэниел Деннет, Дэвид Дойч, Дэниел Дьюи, Эрик Дрекслер, Питер Эккерсли, Амнон Эден, Оуэн Эванс, Бенья Фалленштейн, Алекс Флинт, Карл Фрей, Иэн Голдин, Катя Грэйс, Дж. Сторрс Холл, Робин Хэнсон, Дэмис Хассабис, Джеймс Хьюз, Маркус Хаттер, Гарри Каспаров, Марцин Кульчицки, Шейн Легг, Моше Лукс, Уиллам Макаскилл, Эрик Мандельбаум, Джеймс Мартин, Лилиан Мартин, Роко Мийич, Винсент Мюллер, Элон Маск, Шон О'Хейгертах, Тоби Орд, Деннис Памлин, Дерек Парфит, Дэвид Пирс, Хью Прайс, Мартин Рис, Билл Роско, Стюарт Рассел, Анна Саламон, Лу Салкинд, Андерс Сандберг, Джулиан Савулеску, Йюрген Шмидхубер, Николас Шекел, Марри Шанахан, Ноэль Шарки, Карл Шульман, Питер Сингер, Дэн Стойческу, Яаан Таллинн, Александр Тамас, Макс Тегмарк, Роман Ямпольский и Элиезер Юдковский.

Я признателен тем, кто высказал подробные критические замечания, — это Милан Чиркович, Дэниел Дьюи, Оуэн Эванс, Ник Хэй, Кит Мэнсфилд,

Люк Мюельхаузер, Тоби Орд, Джесс Ридел, Андерс Сандберг, Марри Шанахан и Карл Шульман.

Хочу поблагодарить за советы и помощь в исследованиях Стюарта Армстронга, Дэниела Дьюи, Эрика Дрекслера, Александра Эрлера, Ребекку Роуч и Андерса Сандберга.

Я благодарен за помощь в подготовке рукописи Калебу Беллу, Мало Бургону, Робину Брандту, Лансу Бушу, Кэти Дугласс, Александру Эрлеру, Кристиану Ронну, Сьюзан Роджерс, Эндрю Снайдеру-Битти, Сесилии Тилли и Алексу Вермееру. Особенно хочу выразить признательность моему редактору Киту Мэнсфилду за его абсолютную поддержку на протяжении всей работы.

Прошу прощения у всех, чье имя должен был бы здесь упомянуть.

В заключение говорю огромное спасибо людям, профинансировавшим мой труд, а также друзьям и близким — без вашей поддержки я никогда не написал бы эту книгу.

Об авторе

Ник Бостром — профессор факультета философии Оксфордского университета, основатель и директор Института будущего человечества — междисциплинарного исследовательского центра, занимающегося изучением возможного влияния технологий на будущие глобальные катастрофы. Члены института — лучшие математики, философы и ученые.

Помимо философии, Бостром специализируется на нейробиологии, математической логике и физике. Он — автор более 200 научных публикаций и лауреат премии Eugene R. Gannon Award, которая каждый год вручается одному ученому в мире за достижения в области философии, математики и естественных наук.

Ник Бостром — самый молодой представитель топ-15 списка ведущих мыслителей мира по версии журнала Prospect. Его работы переведены на 22 языка.

Научно-популярное издание

Ник **Бостром**

Искусственный интеллект

Этапы. Угрозы. Стратегии