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

Оглавление

Предисловие партнера

> …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.

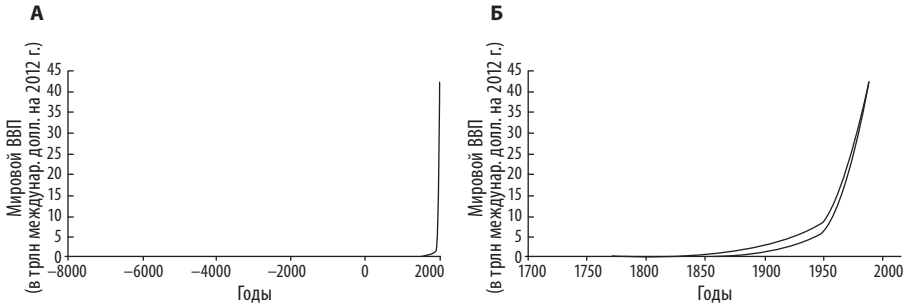


Рис. 1. Динамика мирового ВВП за длительный исторический период. На линейной шкале история мировой экономики отображена как линия, сначала почти сливающаяся с горизонтальной осью, а затем резко устремляющаяся вертикально вверх. **А.** Даже расширив временные границы до десяти тысяч лет в прошлое, мы 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%

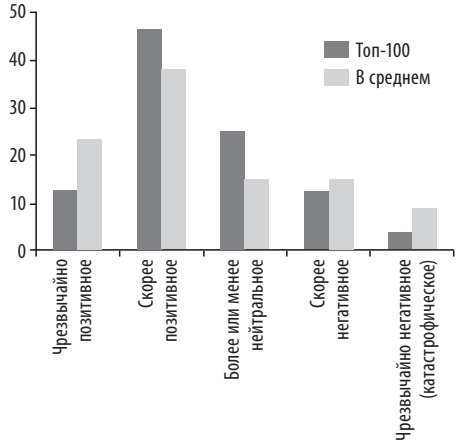
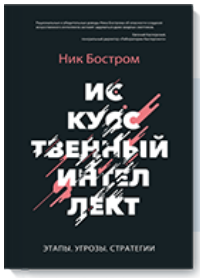


Рис. 2. Долговременное воздействие искусственного интеллекта человеческого уровня[83]

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.



издательство
МАНН, ИВАНОВ И ФЕРБЕР
Максимально полезные книги