ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Зачем эта книга? Почему именно сейчас?

Это книга о прошлом, настоящем и будущем нашего осмысления понятия «искусственный интеллект» (ИИ) и о попытках его создания. Эта тема важна не потому, что ИИ быстро становится обычным явлением в настоящем, а потому, что это господствующая технология будущего. Могущественные государства начинают осознавать этот факт, уже некоторое время известный крупнейшим мировым корпорациям. Мы не можем с точностью предсказать, как быстро будет развиваться технология и по какому пути она пойдет. Тем не менее мы должны строить планы, исходя из возможности того, что машины далеко обойдут человека в способности принятия решений в реальном мире. Что тогда?

Все, что может предложить цивилизация, является продуктом нашего интеллекта; обретение доступа к существенно превосходящим интеллектуальным возможностям стало бы величайшим событием в истории. Цель этой книги — объяснить, почему оно может стать последним событием цивилизации и как нам исключить такой исход.

Общий план книги

Книга состоит из трех частей. Первая часть (главы с первой по третью) исследует понятие интеллекта, человеческого и машинного. Материал не требует специальных знаний, но для интересующихся дополнен четырьмя приложениями, в которых объясняются базовые концепции, лежащие в основе сегодняшних систем ИИ. Во второй части (главы с четвертой по шестую) рассматривается ряд проблем, вытекающих из наделения машин интеллектом. Я уделяю особое внимание контролю — сохранению абсолютной власти над машинами, возможности которых превосходят наши. Третья часть (главы с седьмой по десятую) предлагает новое понимание ИИ, предполагающее, что машины всегда будут служить на благо человечеству. Книга адресована широкому кругу читателей, но, надеюсь, пригодится и специалистам по ИИ, заставив их пересмотреть свои базовые предпосылки.

1

ЧТО, ЕСЛИ МЫ ДОБЬЕМСЯ СВОЕГО?

Много лет назад мои родители жили в английском городе Бирмингеме возле университета. Решив уехать из города, они продали дом Дэвиду Лоджу, профессору английской литературы, на тот момент уже известному романисту. Я с ним так и не встретился, но познакомился с его творчеством, прочитав книги «Академический обмен» и «Тесный мир», в которых главными героями были вымышленные ученые, приезжающие из вымышленной версии Бирмингема в вымышленную версию калифорнийского Беркли. Поскольку я был реальным ученым из реального Бирмингема, только что переехавшим в реальный Беркли, создавалось впечатление, что некто из «Службы совпадений» подает мне сигнал.

Меня особенно поразила одна сцена из книги «Тесный мир». Главное действующее лицо, начинающий литературовед, выступая на крупной международной конференции,

обращается к группе корифеев: «Что, если все согласятся с вами?» Вопрос вызывает ступор, потому что участников больше устраивает интеллектуальная битва, чем раскрытие истины и достижение понимания. Мне тогда пришло в голову, что крупнейшим деятелям в сфере ИИ можно задать тот же вопрос: «Что, если вы добьетесь своего?» Ведь их целью всегда являлось создание ИИ человеческого или сверхчеловеческого уровня, но никто не задумывался о том, что произойдет, если нам это удастся.

Через несколько лет мы с Питером Норвигом начали работать над учебником по ИИ, первое издание которого вышло в 1995 г.[1] Последний раздел этой книги называется «Что произойдет, если у нас получится?». В нем обрисовывается возможность хорошего и плохого исходов, но не делается конкретного вывода. К моменту выхода третьего издания в 2010 г. многие наконец задумались о том, что сверхчеловеческий ИИ необязательно благо, но эти люди находились по большей части за пределами магистральной линии исследования ИИ. К 2013 г. я пришел к убеждению, что это не просто очень важная тема, но, возможно, основной вопрос, стоящий перед человечеством.

В ноябре 2013 г. я выступал с лекцией в Даличской картинной галерее, знаменитом художественном музее в южной части Лондона. Аудитория состояла главным образом из пенсионеров — не связанных с наукой, просто интересующихся интеллектуальными вопросами, — и мне пришлось избегать любых специальных терминов. Мне это показалось подходящей возможностью впервые опробовать свои идеи на публике. Объяснив, что такое ИИ, я огласил пятерку кандидатов на звание «величайшего события в будущем человечества»:

1. Мы все умираем (удар астероида, климатическая катастрофа, пандемия и т. д.).

2. Живем вечно (медицина решает проблему старения).
3. Осваиваем перемещение со сверхсветовой скоростью и покоряем Вселенную.
4. Нас посещает превосходящая инопланетная цивилизация.
5. Мы создаем сверхразумный ИИ.

Я предположил, что пятый вариант, создание сверхразумного ИИ, станет победителем, поскольку это позволило бы нам справиться с природными катастрофами, обрести вечную жизнь и освоить перемещения со сверхсветовыми скоростями, если подобное в принципе возможно. Для нашей цивилизации это был бы громадный скачок. Появление сверхразумного ИИ во многих отношениях аналогично прибытию превосходящей инопланетной цивилизации, но намного более вероятно. Что, пожалуй, самое важное, ИИ, в отличие от инопланетян, в какой-то степени находится в нашей власти.

Затем я предложил слушателям представить, что произойдет, если мы получим от инопланетной цивилизации сообщение, что она явится на Землю через 30–50 лет. Слово «светопреставление» слишком слабо, чтобы описать последствия. В то же время наша реакция на ожидаемое появление сверхразумного ИИ является, как бы это сказать, — индифферентной, что ли? (В последующей лекции я проиллюстрировал это в форме электронной переписки, см. рис. 1.) В итоге я так объяснил значимость сверхразумного ИИ: «Успех в этом деле стал бы величайшим событием в истории человечества… и, возможно, последним ее событием».

Через несколько месяцев, в апреле 2014 г., когда я был на конференции в Исландии, мне позвонили с Национального общественного радио с просьбой дать интервью о фильме «Превосходство», только что вышедшем на экраны в США. Я читал краткое содержание и обзоры, но фильма не видел, поскольку

От кого: Превосходящая инопланетная цивилизация <sac12@sirius.canismajor.u>
Кому: humanity@UN.org
Тема: контакт
Предупреждаем, мы прибудем через 30–50 лет.

От кого: humanity@UN.org
Кому: Превосходящей инопланетной цивилизации <sac12@sirius.canismajor.u>
Тема: нет в офисе: Re: контакт
Человечества в настоящее время нет в офисе. Мы ответим на ваше сообщение, когда вернемся ☺

Рис. 1. Маловероятный обмен электронными письмами после первого контакта с нами превосходящей инопланетной цивилизации

жил в то время в Париже, где он должен был выйти в прокат только в июне. Однако я только что включил в маршрут своего возвращения из Исландии посещение Бостона, чтобы принять участие в собрании Министерства обороны. В общем, из бостонского аэропорта Логан я поехал в ближайший кинотеатр, где шел этот фильм. Я сидел во втором ряду и наблюдал за тем, как в профессора из Беркли, специалиста по ИИ в исполнении Джонни Деппа, стреляют активисты — противники ИИ, напуганные перспективой — вот именно — появления сверхразумного ИИ. Я невольно съежился в кресле. (Очередной сигнал «Службы совпадений»?) Прежде чем герой Джонни Деппа умирает, его мозг загружается в квантовый суперкомпьютер и быстро превосходит человеческий разум, угрожая захватить мир.

19 апреля 2014 г. обзор «Превосходства», написанный в соавторстве с Максом Тегмарком, Фрэнком Уилчеком

и Стивеном Хокингом, вышел в *Huffington Post*. В нем была фраза из моего выступления в Даличе о величайшем событии в человеческой истории. Так я публично связал свое имя с убеждением в том, что моя сфера исследования несет возможную угрозу моему собственному биологическому виду.

Как мы к этому пришли

Идея ИИ уходит корнями в седую древность, но ее «официальным» годом рождения считается 1956 г. Два молодых математика, Джон Маккарти и Марвин Минский, убедили Клода Шеннона, успевшего прославиться как изобретатель теории информации, и Натаниэля Рочестера, разработчика первого коммерческого компьютера IBM, вместе с ними организовать летнюю программу в Дартмутском колледже. Цель формулировалась следующим образом:

> Исследование будет вестись на основе предположения, что любой аспект обучения или любой другой признак интеллекта можно, теоретически, описать настолько точно, что возможно будет создать машину, его воспроизводящую. Будет предпринята попытка узнать, как научить машины использовать язык, формировать абстрактные понятия и концепции, решать задачи такого типа, которые в настоящее время считаются прерогативой человека, и совершенствоваться. Мы считаем, что по одной или нескольким из этих проблем возможен значительный прогресс, если тщательно подобранная группа ученых будет совместно работать над ними в течение лета.

Незачем говорить, что времени потребовалось значительно больше: мы до сих пор трудимся над всеми этими задачами.

В первые лет десять после встречи в Дартмуте в разработке ИИ произошло несколько крупных прорывов, в том числе создание алгоритма универсального логического мышления Алана Робинсона[2] и шахматной программы Артура Самуэля, которая сама научилась обыгрывать своего создателя[3]. В работе над ИИ первый пузырь лопнул в конце 1960-х гг., когда начальные результаты в области машинного обучения и машинного перевода оказались не соответствующими ожиданиям. В отчете, составленном в 1973 г. по поручению правительства Великобритании, делался вывод: «Ни по одному из направлений этой сферы исследований совершенные на данный момент открытия не имели обещанных радикальных последствий»[4]. Иными словами, машины просто не были достаточно умными.

К счастью, в 11-летнем возрасте я не подозревал о существовании этого отчета. Через два года, когда мне подарили программируемый калькулятор Sinclair Cambridge, я просто захотел сделать его разумным. Однако при максимальной длине программы в 36 строк «Синклер» был недостаточно мощным для ИИ человеческого уровня. Не смирившись перед неудачей, я добился доступа к гигантскому суперкомпьютеру CDC 6600[5] в Королевском колледже Лондона и написал шахматную программу — стопку перфокарт 60 см высотой. Не слишком толковую, но это было не важно. Я знал, чем хочу заниматься.

К середине 1980-х гг. я стал профессором в Беркли, а ИИ переживал бурное возрождение благодаря коммерческому потенциалу так называемых экспертных систем. Второй «ИИ-пузырь» лопнул, когда оказалось, что эти системы не отвечают многим задачам, для которых предназначены. Опять-таки машины просто не были достаточно умными. В сфере ИИ настал ледниковый период. Мой курс по ИИ в Беркли, ныне привлекающий 900 с лишним студентов, в 1990 г. заинтересовал всего 25 слушателей.

Сообщество разработчиков ИИ усвоило урок: очевидно, чем умнее, тем лучше, но, чтобы этого добиться, нам нужно покорпеть над основами. Появился выраженный уклон в математику. Были установлены связи с давно признанными научными дисциплинами: теорией вероятности, статистикой и теорией управления. Зерна сегодняшнего прогресса были посажены во время того «ледниковья», в том числе начальные разработки крупномасштабных систем вероятностной логики и того, что стало называться *глубоким обучением*.

Около 2011 г. методы глубокого обучения начали демонстрировать огромные достижения в распознавании речи и визуальных объектов, а также машинного перевода — трех важнейших нерешенных проблем в исследовании ИИ. В 2016 и 2017 гг. программа AlphaGo, разработанная компанией DeepMind, обыграла бывшего чемпиона по игре го Ли Седоля и действующего чемпиона Кэ Цзе. По ранее сделанным оценкам некоторых экспертов, это событие могло произойти не раньше 2097 г. или вообще никогда[6].

Теперь ИИ почти ежедневно попадает на первые полосы мировых СМИ. Созданы тысячи стартапов, питаемые потоками венчурного финансирования. Миллионы студентов занимаются на онлайн-курсах по ИИ и машинному обучению, а эксперты в этой области зарабатывают миллионы долларов. Ежегодные инвестиции из венчурных фондов, от правительств и крупнейших корпораций исчисляются десятками миллиардов долларов — за последние пять лет в ИИ вложено больше денег, чем за всю предшествующую историю этой области знания. Достижения, внедрение которых не за горами, например машины с полным автопилотом и интеллектуальные персональные помощники, по всей видимости, окажут заметное влияние на мир в следующем десятилетии. Огромные экономические и социальные выгоды, которые обещает ИИ, создают мощный импульс для его исследования.

Что будет дальше

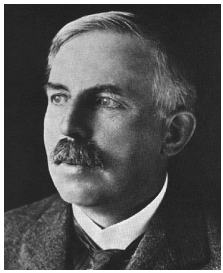
Означает ли этот стремительный прогресс, что нас вот-вот поработят машины? Нет. Прежде чем мы получим нечто, напоминающее машины со сверхчеловеческим разумом, должно произойти немало кардинальных прорывов.

Научные революции печально знамениты тем, что их трудно предсказать. Чтобы это оценить, бросим взгляд на историю одной из научных областей, способной уничтожить человечество, — ядерной физики.

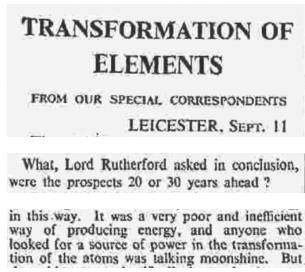
В первые годы XX в., пожалуй, не было более видного физика-ядерщика, чем Эрнест Резерфорд, первооткрыватель протона, «человек, который расщепил атом» (рис. 2а). Как и его коллеги, Резерфорд долгое время знал о том, что ядра атомов заключают в себе колоссальную энергию, но разделял господствующее убеждение, что овладеть этим источником энергии невозможно.

11 сентября 1933 г. Британская ассоциация содействия развитию науки проводила ежегодное собрание в Лестере. Лорд Резерфорд открыл вечернее заседание. Как и прежде, он остудил жар надежд на атомную энергию: «Всякий, кто ищет источник энергии в трансформации атомов, гонится за миражом». На следующее утро речь Резерфорда была напечатана в лондонской газете *Times* (рис. 2б).

Лео Силард (рис. 2в), венгерский физик, только что бежавший из нацистской Германии, остановился в лондонском отеле «Империал» на Рассел-сквер. За завтраком он прочитал статью в *The Times*. Размышляя над речью Резерфорда, он вышел пройтись и открыл нейтронную цепную реакцию[7]. «Неразрешимая» проблема высвобождения ядерной энергии была решена, по сути, менее чем за 24 часа. В следующем году Силард подал секретную заявку на патент ядерного реактора. Первый патент на атомное оружие был выдан во Франции в 1939 г.



(а)



TRANSFORMATION OF ELEMENTS

FROM OUR SPECIAL CORRESPONDENTS

LEICESTER, SEPT. 11

What, Lord Rutherford asked in conclusion, were the prospects 20 or 30 years ahead ?

in this way. It was a very poor and inefficient way of producing energy, and anyone who looked for a source of power in the transformation of the atoms was talking moonshine. But

(б)

(в)

Рис. 2. (а) Лорд Резерфорд, физик-ядерщик.
(б) Фрагменты статьи в *The Times* от 12 сентября 1933 г. о речи, с которой Резерфорд выступил накануне.
(в) Лео Силард, физик-ядерщик

Мораль этой истории — держать пари на человеческую изобретательность безрассудно, особенно если на кону наше будущее. В сообществе разработчиков ИИ складывается своего рода культура отрицания, доходящая даже до отрицания возможности достижения долгосрочных целей ИИ. Как если бы водитель автобуса, в салоне которого сидит все человечество, заявил: «Да, я делаю все возможное, чтобы мы въехали на вершину горы, но, уверяю вас, бензин кончится прежде, чем мы туда попадем!»

Я не утверждаю, что успех в создании ИИ *гарантирован*, и считаю очень маловероятным, что это случится в ближайшие годы. Представляется тем не менее разумным подготовиться к самой возможности. Если все сложится хорошо, это возвестит золотой век для человечества, но мы должны взглянуть правде в лицо: мы собираемся использовать нечто намного более могущественное, чем люди. Как добиться, чтобы оно никогда, ни при каких условиях не взяло верх над нами?

Чтобы составить хотя бы какое-то представление о том, с каким огнем мы играем, рассмотрим алгоритмы выбора контента в социальных сетях. Они не особо интеллектуальны,

но способны повлиять на весь мир, поскольку оказывают непосредственное воздействие на миллиарды людей. Обычно подобные алгоритмы направлены на максимизацию вероятности того, что пользователь кликнет мышью на представленные элементы. Решение простое — демонстрировать те элементы, которые пользователю нравится кликать, правильно? Неправильно. Решение заключается в том, чтобы менять предпочтения пользователя, делая их более предсказуемыми. Более предсказуемому пользователю можно подсовывать элементы, которые он с большой вероятностью кликнет, повышая прибыль таким образом. Люди с радикальными политическими взглядами отличаются большей предсказуемостью в своем выборе. (Вероятно, имеется и категория ссылок, на которые с высокой долей вероятности станут переходить убежденные центристы, но нелегко понять, что в нее входит.) Как любая рациональная сущность, алгоритм обучается способам изменения своего окружения — в данном случае предпочтений пользователя, — чтобы максимизировать собственное вознаграждение[8]. Возможные последствия включают возрождение фашизма, разрыв социальных связей, лежащих в основе демократий мира, и, потенциально, конец Европейского союза и НАТО. Неплохо для нескольких строчек кода, пусть и действовавшего с небольшой помощью людей. Теперь представьте, на что будет способен *действительно* интеллектуальный алгоритм.

Что пошло не так?

Историю развития ИИ движет одно-единственное заклинание: «Чем интеллектуальнее, тем лучше». Я убежден, что это ошибка, и дело не в туманных опасениях, что нас превзойдут, а в самом нашем понимании интеллекта.

Понятие интеллекта является определяющим для нашего представления о самих себе — поэтому мы называем себя

Homo sapiens, или «человек разумный». По прошествии двух с лишним тысяч лет самопознания мы пришли к пониманию интеллекта, которое может быть сведено к следующему утверждению:

> *Люди разумны настолько, насколько можно ожидать, что наши действия приведут к достижению поставленных нами целей.*

Все прочие характеристики разумности — восприятие, мышление, обучение, изобретательство и т. д. — могут быть поняты через их вклад в нашу способность успешно действовать. С самого начала разработки ИИ интеллектуальность машин определялась аналогично:

> *Машины разумны настолько, насколько можно ожидать, что их действия приведут к достижению поставленных ими целей.*

Поскольку машины, в отличие от людей, не имеют собственных целей, мы говорим им, каких целей нужно достичь. Иными словами, мы строим оптимизирующие машины, ставим перед ними цели, и они принимаются за дело.

Этот общий подход не уникален для ИИ. Он снова и снова применяется в технологических и математических схемах нашего общества. В области теории управления, которая разрабатывает системы управления всем, от авиалайнеров до инсулиновых помп, работа системы заключается в минимизации *функции издержек*, обычно дающих некоторое отклонение от желаемого поведения. В сфере экономики механизмы политики призваны максимизировать *пользу* для индивидов, *благосостояние* групп и *прибыль* корпораций[9]. В исследовании операций, направлении, решающем комплексные логистические и производственные проблемы,

решение максимизирует ожидаемую *сумму вознаграждений* во времени. Наконец, в статистике обучающиеся алгоритмы строятся с таким расчетом, чтобы минимизировать ожидаемую *функцию потерь*, определяющую стоимость ошибки прогноза.

Очевидно, эта общая схема, которую я буду называть *стандартной моделью*, широко распространена и чрезвычайно действенна. К сожалению, *нам не нужны машины, интеллектуальные в рамках стандартной модели*.

На оборотную сторону стандартной модели указал в 1960 г. Норберт Винер, легендарный профессор Массачусетского технологического института и один из ведущих математиков середины XX в. Винер только что увидел, как шахматная программа Артура Самуэля научилась играть намного лучше своего создателя. Этот опыт заставил его написать провидческую, но малоизвестную статью «Некоторые нравственные и технические последствия автоматизации»[10]. Вот как он формулирует главную мысль:

> Если мы используем для достижения своих целей механического посредника, в действие которого не можем эффективно вмешаться… нам нужна полная уверенность в том, что заложенная в машину цель является именно той целью, к которой мы действительно стремимся.

«Заложенная в машину цель» — это те самые задачи, которые машины оптимизируют в стандартной модели. Если мы вводим ошибочные цели в машину, более интеллектуальную, чем мы сами, она достигнет цели и мы проиграем. Описанная мною деградация социальных сетей — просто цветочки, результат оптимизации неверной цели во всемирном масштабе, в сущности, неинтеллектуальным алгоритмом. В главе 5 я опишу намного худшие результаты.

Этому не приходится особенно удивляться. Тысячелетиями мы знали, как опасно получить именно то, о чем мечтаешь. В любой сказке, где герою обещано исполнить три желания, третье всегда отменяет два предыдущих.

В общем представляется, что движение к созданию сверхчеловеческого разума не остановить, но успех может обернуться уничтожением человеческой расы. Однако не все потеряно. Мы должны найти ошибки и исправить их.

Можем ли мы что-то исправить

Проблема заключается в самом базовом определении ИИ. Мы говорим, что машины разумны, поскольку можно ожидать, что их действия приведут к достижению *их* целей, но не имеем надежного способа добиться того, чтобы *их* цели совпадали с *нашими*.

Что, если вместо того, чтобы позволить машинам преследовать *их* цели, потребовать от них добиваться *наших* целей? Такая машина, если бы ее можно было построить, была бы не только *интеллектуальной*, но и *полезной* для людей. Попробуем следующую формулировку:

> *Машины полезны настолько, насколько можно ожидать, что их действия достигнут наших целей.*

Пожалуй, именно к этому нам все время следовало стремиться.

Разумеется, тут есть трудность: наши цели заключены в нас (всех 8 млрд человек, во всем их великолепном разнообразии), а не в машинах. Тем не менее возможно построить машины, полезные именно в таком понимании. Эти машины неизбежно будут не уверены в наших целях — в конце концов, мы сами в них не уверены, — но, оказывается, это

свойство, а не ошибка (то есть хорошо, а не плохо). Неуверенность относительно целей предполагает, что машины неизбежно будут полагаться на людей: спрашивать разрешения, принимать исправления и позволять себя выключить.

Исключение предпосылки, что машины должны иметь определенные цели, означает, что мы должны будем изъять и заменить часть предпосылок ИИ — базовые определения того, что мы пытаемся создать. Это также предполагает перестройку значительной части суперструктуры — совокупности идей и методов по разработке ИИ. В результате возникнут новые отношения людей и машин, которые, я надеюсь, позволят нам благополучно прожить следующие несколько десятилетий.

2

РАЗУМНОСТЬ ЛЮДЕЙ И МАШИН

Если вы зашли в тупик, имеет смысл вернуться назад и выяснить, в какой момент вы свернули не в ту сторону. Я заявил, что стандартная модель ИИ, в которой машины оптимизируют фиксированную цель, поставленную людьми, — это тупик. Проблема не в том, что у нас может *не получиться* хорошо выполнить работу по созданию ИИ, а в том, что мы может добиться *слишком большого успеха*. Само определение успеха применительно к ИИ ошибочно.

Итак, пройдем по собственным следам в обратном направлении вплоть до самого начала. Попытаемся понять, как сложилась наша концепция разумности и как получилось, что она была применена к машинам. Тогда появится шанс предложить лучшее определение того, что следует считать хорошей системой ИИ.

Разумность

Как устроена Вселенная? Как возникла жизнь? Где ключи к пониманию этого? Эти фундаментальные вопросы заслуживают размышлений. Но кто их задает? Как я на них отвечаю?

Как может горстка материи — несколько килограммов розовато-серого бланманже, которое мы называем мозгом, — воспринимать, понимать, прогнозировать и управлять невообразимо огромным миром? Очень скоро мозг начинает исследовать сам себя.

Тысячелетиями мы пытаемся понять, как работает наш ум. Первоначально это делалось из любопытства, ради самоконтроля и вполне прагматичной задачи решения математических задач. Тем не менее каждый шаг к объяснению того, как работает ум, является и шагом к воссозданию возможностей ума в искусственном объекте — то есть к созданию ИИ.

Чтобы разобраться в том, как создать разумность, полезно понять, что это такое. Ответ заключается не в тестах на IQ и даже не в тесте Тьюринга, а попросту во взаимосвязи того, что мы воспринимаем, чего хотим и что делаем. Грубо говоря, сущность разумна настолько, насколько ее действия могут привести к получению желаемого при условии, что желание было воспринято.

Эволюционные корни

Возьмем самую обыкновенную бактерию, например *E. coli*. У нее имеется полдесятка жгутиков — длинных тонких, как волоски, усиков, вращающихся у основания по часовой или против часовой стрелки. (Этот двигатель сам по себе потрясающая штука, но сейчас речь не о нем.) Плавая в жидкости у себя дома — в нижнем отделе вашего кишечника, — *E. coli* вращает жгутики то по часовой стрелке и «пританцовывает» на месте, то против, отчего они сплетаются в своего рода пропеллер, и бактерия плывет по прямой. Таким образом, *E. coli* может перемещаться произвольным образом — то плыть, то останавливаться, — что позволяет ей находить

и потреблять глюкозу, вместо того чтобы оставаться неподвижной и погибнуть от голода.

Если бы на этом все заканчивалось, мы не назвали бы *E. coli* сколько-нибудь разумной, потому что ее действия совершенно не зависели бы от среды. Она не принимала бы никаких решений, только выполняла определенные действия, встроенные эволюцией в ее гены. Но это не все. Если *E. coli* ощущает увеличение концентрации глюкозы, то дольше плывет и меньше задерживается на месте, а чувствуя меньшую концентрацию глюкозы — наоборот. Таким образом, то, что она делает (плывет к глюкозе), повышает ее шансы достичь желаемого (по всей видимости, больше глюкозы), причем она действует с опорой на воспринимаемое (увеличение концентрации глюкозы).

Возможно, вы думаете: «Но ведь и такое поведение встроила в ее гены эволюция! Как это делает ее разумной?» Такое направление мысли опасно, поскольку и в ваши гены эволюция встроила базовую конструкцию мозга, но вы едва ли станете отрицать собственную разумность на этом основании. Дело в том, что нечто заложенное эволюцией в гены *E. coli*, как и в ваши, представляет собой механизм изменения поведения бактерии под влиянием внешней среды. Эволюция не знает заранее, где будет глюкоза или ваши ключи, поэтому организм, наделенный способностью найти их, получает еще одно преимущество.

Разумеется, *E. coli* не гигант мысли. Насколько мы знаем, она не помнит, где была, и если переместится из точки А в точку Б и не найдет глюкозы, то, скорее всего, просто вернется в А. Если мы создадим среду, где привлекательное увеличение концентрации глюкозы ведет к месту содержания фенола (яда для *E. coli*), бактерия так и будет следовать вслед за ростом концентрации. Она совершенно не учится. У нее нет мозга, за все отвечает лишь несколько простых химических реакций.

Огромным шагом вперед стало появление *потенциала действия* — разновидности электрической сигнализации, возникшей у одноклеточных организмов около 1 млрд лет назад. Впоследствии многоклеточные организмы выработали специализированные клетки, *нейроны*, которые с помощью электрических потенциалов быстро — со скоростью до 120 м/с, или 430 км/ч — передают сигналы в организме. Связи между нейронами называются *синапсами*. Сила синаптической связи определяет меру электрического возбуждения, проходящего от одного нейрона к другому. Изменяя силу синаптических связей, животные учатся[1]. Обучаемость дает громадное эволюционное преимущество, поскольку позволяет животному адаптироваться к широкому спектру условий. Кроме того, обучаемость ускоряет темп самой эволюции.

Первоначально нейроны были сгруппированы в *нервные узлы*, которые распределялись по всему организму и занимались координацией деятельности, скажем, питания и выделения, или согласованным сокращением мышечных клеток в определенной области тела. Изящные пульсации медузы — результат действия нервной сети. У медузы нет мозга.

Мозг возник позднее, вместе со сложными органами чувств, такими как глаза и уши. Через несколько сот миллионов лет после появления медузы с ее нервными узлами появились мы, люди, существа с большим головным мозгом — 100 млрд (10^{11}) нейронов и квадриллион (10^{15}) синапсов. Медленное в сравнении с электрическими цепями «время цикла» в несколько миллисекунд на каждое изменение состояния является быстрым по сравнению с большинством биологических процессов. Человеческий мозг часто описывается своими владельцами как «самый сложный объект во Вселенной», что, скорее всего, неверно, но хорошее оправдание тому факту, что мы до сих пор очень слабо представляем себе, как он работает. Мы очень много знаем о биохимии нейронов

и синапсов в анатомических структурах мозга, но о нейронной реализации *когнитивного* уровня — обучении, познании, запоминании, мышлении, планировании, принятии решений и т. д. — остается по большей части гадать[2]. (Возможно, это изменится с углублением нашего понимания ИИ или создания все более точных инструментов измерения мозговой активности.) Итак, читая в СМИ, что такое-то средство реализации ИИ «работает точно так же, как человеческий мозг», можно подозревать, что это чье-то предположение или чистый вымысел.

В сфере *сознания* мы в действительности не знаем ничего, поэтому и я ничего не стану об этом говорить. Никто в сфере ИИ не работает над наделением машин сознанием, никто не знает, с чего следовало бы начинать такую работу, и никакое поведение не имеет в качестве предшествующего условия сознание. Допустим, я даю вам программу и спрашиваю: «Представляет ли она угрозу для человечества?» Вы анализируете код и видите — действительно, если его запустить, код составит и осуществит план, результатом которого станет уничтожение человеческой расы, как шахматная программа составила и осуществила бы план, в результате которого смогла бы обыграть любого человека. Предположим далее, что я говорю, что этот код, если его запустить, еще и создает своего рода машинное сознание. Изменит ли это ваш прогноз? Ни в малейшей степени. Это *не имеет совершенно никакого значения*[3]. Ваш прогноз относительно его действия останется точно таким же, потому что основывается на коде. Все голливудские сюжеты о том, как машины таинственным образом обретают сознание и проникаются ненавистью к людям, упускают из вида главное: важны способности, а не осознанность.

У мозга есть важное когнитивное свойство, которое мы *начинаем* понимать, а именно — *система вознаграждения*. Это

интересная сигнальная система, основанная на дофамине, которая связывает с поведением положительные и отрицательные стимулы. Ее действие открыл шведский нейрофизиолог Нильс-Аке Хилларп и его сотрудники в конце 1950-х гг. Она заставляет нас искать положительные стимулы, например сладкие фрукты, повышающие уровень дофамина; она же заставляет нас избегать отрицательные стимулы, скажем, опасность и боль, снижающие уровень дофамина. В каком-то смысле она действует так же, как механизм поиска глюкозы у бактерии *E. coli*, но намного сложнее. Система вознаграждения обладает «встроенными» методами обучения, так что наше поведение со временем становится более эффективным в плане получения вознаграждения. Кроме того, она делает возможным отложенное вознаграждение, благодаря чему мы учимся желать, например, деньги, обеспечивающие отдачу в будущем, а не сию минуту. Мы понимаем, как работает система вознаграждения в нашем мозге, в том числе потому, что она напоминает метод *обучения с подкреплением*, разработанный в сфере исследования ИИ, для которого у нас имеется основательная теория[4].

С эволюционной точки зрения мы можем считать систему вознаграждения мозга аналогом механизма поиска глюкозы у *E. coli*, способом повышения эволюционной приспособленности. Организмы, более эффективные в поиске вознаграждения — а именно: в нахождении вкусной пищи, избегании боли, занятии сексом и т. д., — с большей вероятностью передают свои гены потомству. Организму невероятно трудно решить, какое действие в долгосрочной перспективе скорее всего приведет к успешной передаче его генов, поэтому эволюция упростила нам эту задачу, снабдив встроенными указателями.

Однако эти указатели несовершенны. Некоторые способы получения вознаграждения *снижают* вероятность того, что наши гены будут переданы потомству. Например,

принимать наркотики, пить огромное количество сладкой газировки и играть в видеоигры по 18 часов в день представляется контрпродуктивным с точки зрения продолжения рода. Более того, если бы вы получили прямой электрический доступ к своей системе вознаграждения, то, по всей вероятности, занимались бы самостимуляцией без конца, пока не умерли бы[5].

Рассогласование вознаграждающих сигналов и эволюционной необходимости влияет не только на отдельных индивидов. На маленьком острове у берегов Панамы живет карликовый трехпалый ленивец, как оказалось, страдающий зависимостью от близкого к валиуму вещества в своем рационе из мангровых листьев и находящийся на грани вымирания[6]. Таким образом, целый вид может исчезнуть, если найдет экологическую нишу, где сможет поощрять свою систему вознаграждения нездоровым образом.

Впрочем, за исключением подобных случайных неудач, обучение максимизации вознаграждения в естественной среде обычно повышает шансы особи передать свои гены и пережить изменения окружающей среды.

Эволюционный ускоритель

Обучение способствует не только выживанию и процветанию. Оно еще и *ускоряет эволюцию*. Каким образом? В конце концов, обучение не меняет нашу ДНК, а эволюция заключается в изменении ДНК с поколениями. Предположение, что между обучением и эволюцией существует связь, независимо друг от друга высказали в 1896 г. американский психолог Джеймс Болдуин[7] и британский этолог Конви Ллойд Морган[8], но в те времена оно не стало общепринятым.

Эффект Болдуина, как его теперь называют, можно понять, если представить, что эволюция имеет выбор между

созданием *инстинктивного* организма, любая реакция которого зафиксирована заранее, и *адаптивного* организма, который учится, как ему действовать. Теперь предположим, для примера, что оптимальный инстинктивный организм можно закодировать шестизначным числом, скажем, 472116, тогда как в случае адаптивного организма эволюция задает лишь 472, и организм сам должен заполнить пробел путем обучения на протяжении жизни. Очевидно, если эволюция должна позаботиться лишь о выборе трех первых цифр, ее работа значительно упрощается; адаптивный организм, получая через обучение последние три цифры, за одну жизнь делает то, на что эволюции потребовалось бы много поколений. Таким образом, способность учиться позволяет идти эволюционно коротким путем при условии, что адаптивный организм сумеет выжить в процессе обучения. Компьютерное моделирование свидетельствует о реальности эффекта Болдуина[9]. Влияние культуры лишь ускоряет процесс, потому что организованная цивилизация защищает индивидуальный организм, пока тот учится, и передает ему информацию, которую в ином случае индивиду пришлось бы добывать самостоятельно.

Описание эффекта Болдуина является увлекательным, но неполным: оно предполагает, что обучение и эволюция обязательно работают в одном направлении, а именно, что направление обучения, вызванное любым сигналом внутренней обратной связи в организме, с точностью соответствует эволюционной приспособленности. Как мы видели на примере карликового трехпалого ленивца, это не так. В лучшем случае встроенные механизмы обучения дают лишь самое общее представление о долгосрочных последствиях любого конкретного действия для эволюционной приспособленности. Более того, возникает вопрос: как вообще возникла система вознаграждения? Ответ: разумеется,

в процессе эволюции, усвоившей тот механизм обратной связи, который хоть сколько-нибудь соответствовал эволюционной приспособленности[10]. Очевидно, механизм обучения, который заставлял бы организм удаляться от потенциальных брачных партнеров и приближаться к хищникам, не просуществовал бы долго.

Таким образом, мы должны поблагодарить эффект Болдуина за то, что нейроны, с их способностью к обучению и решению задач, широко распространены в животном царстве. В то же время важно понимать, что эволюции на самом деле все равно, есть у вас мозг или интересные мысли. Эволюция считает вас лишь *агентом*, то есть кем-то, кто действует. Такие достославные характеристики интеллекта, как логическое рассуждение, целенаправленное планирование, мудрость, остроумие, воображение и креативность, могут быть принципиально важны для разумности агента, а могут и не быть. Идея ИИ невероятно захватывает в том числе потому, что предлагает возможный путь к пониманию этих механизмов. Может быть, нам удастся узнать, как эти характеристики интеллекта делают возможным разумное поведение, а также почему без них невозможно достичь по-настоящему разумного поведения.

Рациональность для одного

С самых истоков древнегреческой философии концепция разума связывалась со способностью воспринимать, мыслить логически и действовать *успешно*[11]. В течение столетий эта концепция расширилась и уточнилась.

Аристотель среди прочих изучал понятие успешного рассуждения — методы логической дедукции, которые ведут к верному выводу при условии верной предпосылки. Он также исследовал процесс принятия решения о том,

как действовать, иногда называемый *практическим* рассуждением. Философ считал, что предполагается логическое заключение о том, что определенная последовательность действий приводит к желаемой цели*:

> Решение наше касается не целей, а средств, ведь врач принимает решения не о том, будет ли он лечить, и ритор — не о том, станет ли он убеждать… но, поставив цель, он заботится о том, каким образом и какими средствами ее достигнуть; и если окажется несколько средств, то прикидывают, какое самое простое и наилучшее; если же достижению цели служит одно средство, думают, *как* ее достичь при помощи этого средства и *что* будет средством для этого средства, покуда не дойдут до первой причины, находят которую последней… И, если наталкиваются на невозможность [достижения], отступаются (например, если нужны деньги, а достать их невозможно); когда же это представляется возможным, тогда и берутся за дело[12].

Можно сказать, что этот фрагмент задает направление следующих 2000 лет западной мысли о рациональности. В нем говорится, что «цель» — то, чего хочет данный человек, — фиксирована и задана, а также что рациональным является такое действие, которое, согласно логическому выводу о последовательности действий, самым «простым и наилучшим» образом приводит к цели.

Предположение Аристотеля выглядит разумно, но не исчерпывает рационального поведения. Главное, в нем отсутствует неопределенность. В реальном мире наблюдается склонность реальности вторгаться в наши действия, и лишь немногие из них или их последовательностей гарантированно

---

* Цит. в пер. Н. Брагинской. — *Прим. пер.*

достигают поставленной цели. Например, я пишу это предложение в дождливое воскресенье в Париже, а во вторник в 14:15 из аэропорта Шарля де Голля вылетает мой самолет в Рим. От моего дома до аэропорта около 45 минут, и я планирую выехать в аэропорт около 11:30, то есть с большим запасом, но из-за этого мне, скорее всего, придется не меньше часа просидеть в зоне вылета. Значит ли это, что я *гарантированно* успею на рейс? Вовсе нет. Может возникнуть ужасная пробка или забастовка таксистов; такси, в котором я еду, может попасть в аварию; или водителя задержат за превышение скорости и т. д. Я мог бы выехать в аэропорт в понедельник, на целый день раньше. Это значительно снизило бы шанс опоздать на рейс, но перспектива провести ночь в зоне вылета меня не привлекает. Иными словами, мой план включает *компромисс* между уверенностью в успехе и стоимостью этой уверенности. План приобретения дома предполагает аналогичный компромисс: купить лотерейный билет, выиграть миллион долларов, затем купить дом. Этот план является самым «простым и наилучшим» путем к цели, но маловероятно, чтобы он оказался успешным. Однако между легкомысленным планом покупки дома и моим трезвым и обоснованным планом приезда в аэропорт разница лишь в степени риска. Оба представляют собой ставку, но одна ставка выглядит более рациональной.

Оказывается, ставка играет главную роль в обобщении предположения Аристотеля с тем, чтобы включить неопределенность. В 1560-х гг. итальянский математик Джероламо Кардано разработал первую математически точную теорию вероятности, используя в качестве основного примера игру в кости. (К сожалению, эта работа была опубликована лишь в 1663 г.[13]) В XVII в. французские мыслители, в том числе Антуан Арно и Блез Паскаль, начали — разумеется, в интересах математики — изучать вопрос рационального принятия

решений в азартных играх[14]. Рассмотрим следующие две ставки:

А: 20% вероятности выиграть $10.
Б: 5% вероятности выиграть $100.

Предложение, выдвинутое математиками, скорее всего, совпадает с решением, которое приняли бы вы: сравнить *ожидаемую ценность* ставок, то есть среднюю сумму, которую можно рассчитывать получить с каждой ставки. В случае А ожидаемая ценность составляет 20% от $10, или $2. В случае Б — 5% от $100, или $5. Так что, согласно этой теории, ставка Б лучше. В теории есть смысл, поскольку, если делать одну и ту же ставку снова и снова, игрок, следующий правилу, в конце концов выиграет больше, чем тот, кто ему не следует.

В XVIII в. швейцарский математик Даниил Бернулли заметил, что это правило, по-видимому, не работает для больших денежных сумм[15]. Рассмотрим, например, такие две ставки:

А: 100% вероятности получить $10 000 000
(ожидаемая ценность $10 000 000).
Б: 1% вероятности получить $1 000 000 100
(ожидаемая ценность $10 000 001).

Большинство читателей этой книги, как и ее автор, предпочли бы ставку А, несмотря на то что ожидаемая ценность призывает к противоположному выбору! Бернулли предположил, что ставки оцениваются не по ожидаемой денежной ценности, а по ожидаемой *полезности*. Полезность — способность приносить человеку пользу или выгоду — является, по его мысли, внутренним, субъективным качеством,