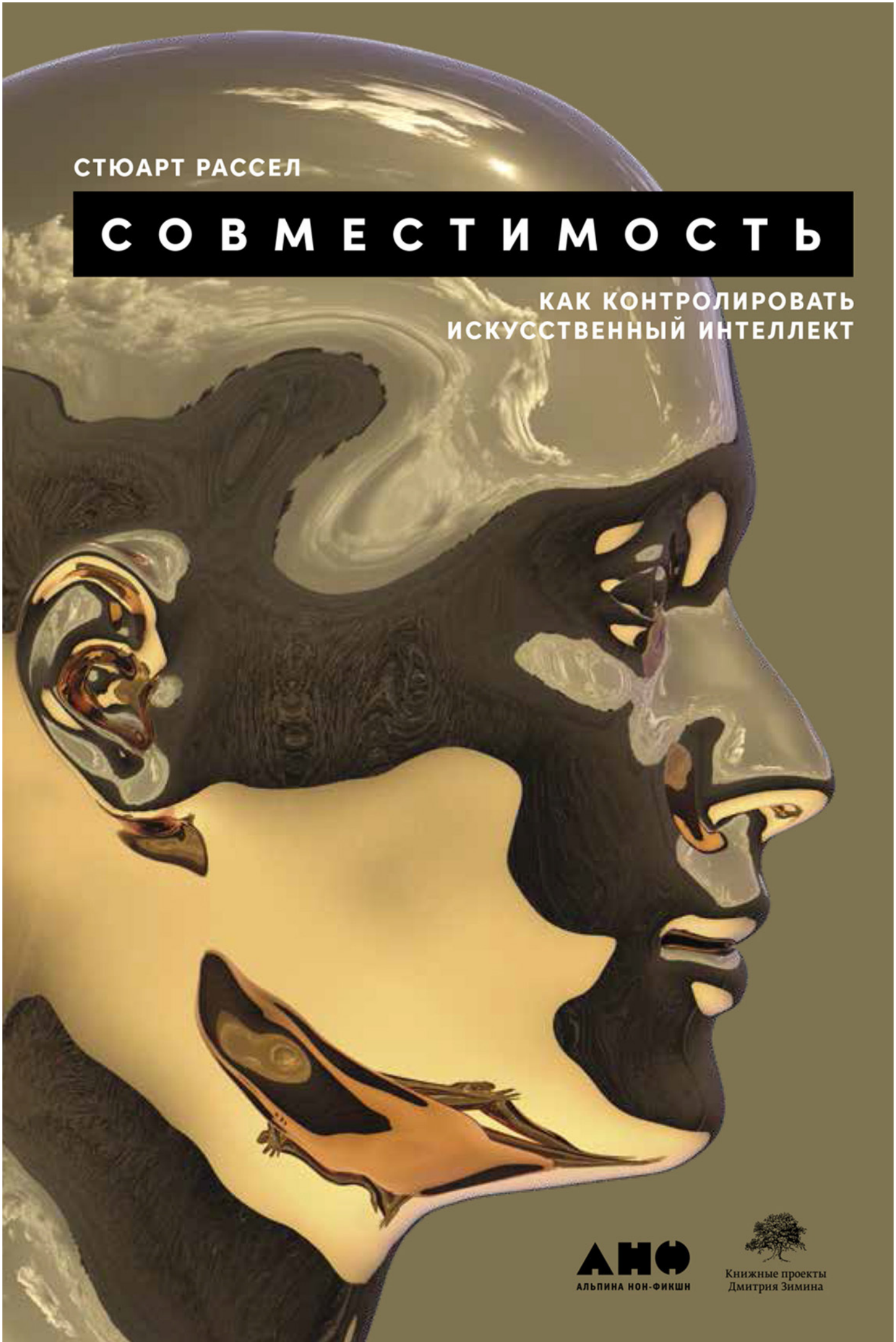


СТЮАРТ РАССЕЛ

СОВМЕСТИМОСТЬ

КАК КОНТРОЛИРОВАТЬ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

АНФ
АЛЬПИНА НОН-ФИКШН

Книжные проекты
Дмитрия Зимина

Стюарт Рассел

Совместимость. Как контролировать искусственный интеллект

«Альпина Диджитал»

2019

Рассел С.

Совместимость. Как контролировать искусственный интеллект / С. Рассел — «Альпина Диджитал», 2019

ISBN 978-5-0013-9370-2

В массовом сознании сверхчеловеческий искусственный интеллект – технологическое цунами, угрожающее не только экономике и человеческим отношениям, но и самой цивилизации. Конфликт между людьми и машинами видится неотвратимым, а его исход предопределенным. Выдающийся исследователь ИИ Стюарт Рассел утверждает, что этого сценария можно избежать. В своей новаторской книге автор рассказывает, каким образом люди уже научились использовать ИИ, в диапазоне от смертельного автономного оружия до манипуляций нашими предпочтениями, и чему еще смогут его научить. Если это случится и появится сверхчеловеческий ИИ, мы столкнемся с сущностью, намного более могущественной, чем мы сами. Как гарантировать, что человек не окажется в подчинении у сверхинтеллекта? Для этого, полагает Рассел, искусственный интеллект должен строиться на новых принципах. Машины должны быть скромными и альтруистичными и решать наши задачи, а не свои собственные. О том, что это за принципы и как их реализовать, читатель узнает из этой книги, которую самые авторитетные издания в мире назвали главной книгой об искусственном интеллекте. 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. Введение понятия полезности – невидимого свойства – для объяснения человеческого поведения посредством математической теории было потрясающим для своего времени. Тем более что, в отличие от денежных сумм, ценность разных ставок и призов с точки зрения полезности недоступна для прямого наблюдения. Первыми, кто действительно выиграет от появления роботов в доме, станут престарелые и немощные, которым полезный робот может обеспечить определенную

степень независимости, недостижимую иными средствами. Даже если робот выполняет ограниченный круг заданий и имеет лишь зачаточное понимание происходящего, он может быть очень полезным. Очевидно, действия лояльных машин должны будут ограничиваться правилами и запретами, как действия людей ограничиваются законами и социальными нормами. Некоторые специалисты предлагают в качестве решения безусловную ответственность.

ISBN 978-5-0013-9370-2

Содержание

Стюарт Рассел
Совместимость. Как контролировать искусственный интеллект

Переводчик *Наталья Колпакова*
Научный редактор *Борис Миркин, д-р техн. наук*
Редактор *Антон Никольский*
Издатель *П. Подкосов*
Руководитель проекта *И. Серёгина*
Корректоры *И. Астапкина, М. Миловидова*
Компьютерная верстка *А. Фоминов*
Дизайн обложки *Ю. Буга*

Фото на обложке Shutterstock

Рассел С.
Совместимость. Как контролировать искусственный интеллект / Тони Салдана; Стюарт Рассел; Пер. с англ. – М.: Альпина нон-фикшн, 2021.
ISBN 978-5-0013-9370-2

⁂

Посвящается
Лой, Гордону, Люси, Джорджу и Айзеку

Книжные проекты
Дмитрия Зимина

Эта книга издана в рамках программы «Книжные проекты Дмитрия Зимина» и продолжает серию «Библиотека «Династия». Дмитрий Борисович Зимин – основатель компании «Вымпелком» (Beeline), фонда некоммерческих программ «Династия» и фонда «Московское время». Программа «Книжные проекты Дмитрия Зимина» объединяет три проекта, хорошо знакомые читательской аудитории: издание научно-популярных переводных книг «Библиотека «Династия», издательское направление фонда «Московское время» и премию в области русскоязычной научно-популярной литературы «Просветитель». Подробную информацию о «Книжных проектах Дмитрия Зимина» вы найдете на сайте ziminbookprojects.ru.

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»:

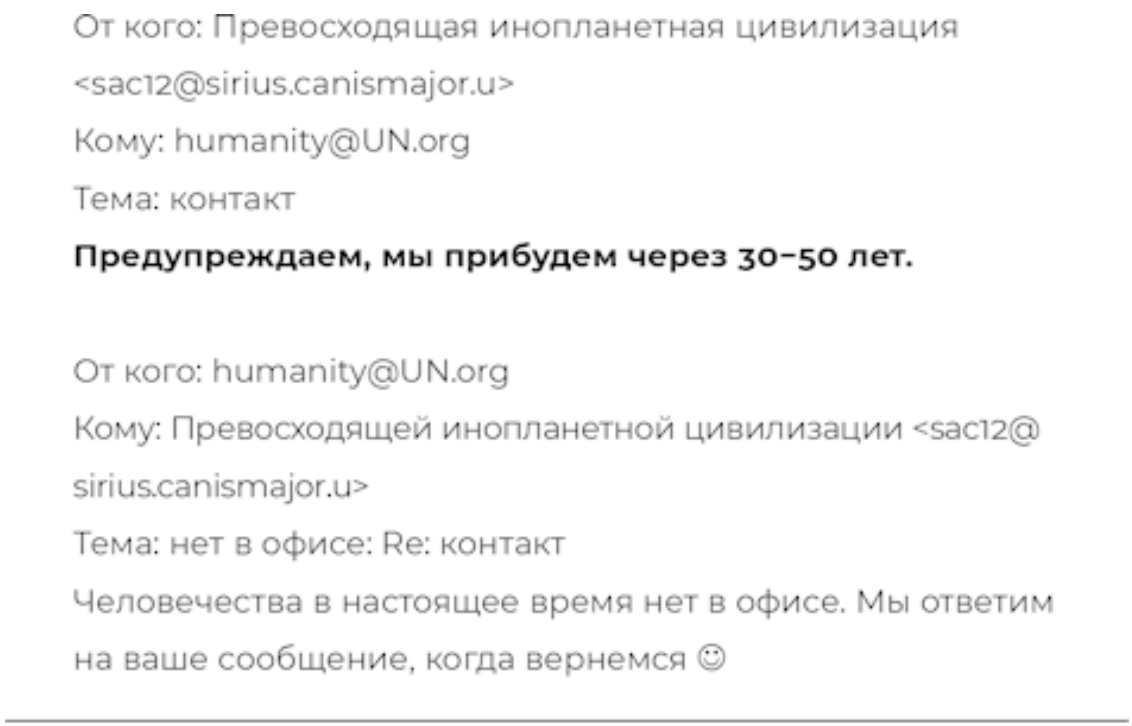
1. Мы все умираем (удар астероида, климатическая катастрофа, пандемия и т. д.).
2. Живем вечно (медицина решает проблему старения).
3. Осваиваем перемещение со сверхсветовой скоростью и покоряем Вселенную.
4. Нас посещает превосходящая инопланетная цивилизация.
5. Мы создаем сверхразумный ИИ.

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.

[1] Первое издание моего учебника по ИИ, написанного в соавторстве с Питером Норвигом, в настоящее время директором Google по науке: Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1st ed. (Prentice Hall, 1995).

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 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.


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

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-

---

[2] Робинсон разработал алгоритм *разрешения*, который может, при наличии времени, доказать любое логическое следствие из комплекса логических утверждений первого порядка. В отличие от предыдущих алгоритмов, он не требует преобразования в пропозиционную логику. J. Alan Robinson, “A machine-oriented logic based on the resolution principle,” *Journal of the ACM* 12 (1965): 23–41.

[3] Артур Самуэль, американский первопроходец компьютерной эры, начал карьеру в IBM. В статье, посвященной его работе с шашками, впервые был использован термин *машинное обучение*, хотя Алан Тьюринг еще в 1947 г. говорил о «машине, способной учиться на опыте». Arthur Samuel, “Some studies in machine learning using the game of checkers”, *IBM Journal of Research and Development* 3 (1959): 210–29.

[4] Так называемый Отчет Лайтхилла привел к отмене финансирования исследования ИИ везде, кроме Эдинбургского и Сассекского университетов: Michael James Lighthill, “Artificial intelligence: A general survey,” in *Artificial Intelligence: A Paper Symposium* (Science Research Council of Great Britain, 1973).

[5] CDC 6600 занимал целую комнату, а его стоимость была эквивалентна $20 млн. Для своего времени он был невероятно мощным, хотя и в миллион раз менее мощным, чем iPhone.

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.

[6] После победы DeepBlue над Каспаровым по крайней мере один комментатор предсказал, что в го подобное произойдет не раньше чем через сто лет: George Johnson, “To test a powerful computer, play an ancient game,” *The New York Times*, July 29, 1997.

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.

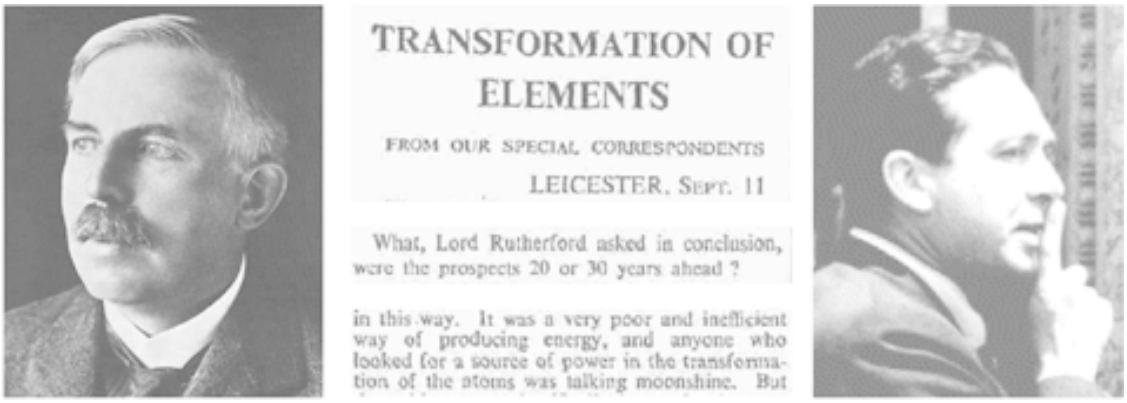


Рис. 2. (а) Лорд Резерфорд, физик-ядерщик.
(б) Фрагменты статьи в *The Times* от 12 сентября 1933 г. о речи, с которой Резерфорд выступил накануне.
(в) Лео Силард, физик-ядерщик

[7] Очень легкое для понимания описание развития ядерной технологии см. в: Richard Rhodes, *The Making of the Atomic Bomb* (Simon & Schuster, 1987).

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.

[8] Простой алгоритм контролируемого обучения может не обладать таким эффектом, если не имеет оболочки в виде платформы A/B тестирования (обычного инструмента онлайнового маркетинга). Алгоритмы решения проблемы многорукого бандита и алгоритмы обучения с подкреплением окажут это воздействие, если будут работать с явным представлением состояния пользователя или неявным представлением в плане истории взаимодействий с пользователем.

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:

---

[9] Некоторые считают, что корпорации, ориентированные на максимизацию прибыли, уже являются вышедшими из-под контроля искусственными сущностями. См., например: Charles Stross, “Dude, you broke the future!” (keynote, 34th Chaos Communications Congress, 2017). См. также: Ted Chiang, “Silicon Valley is turning into its own worst fear,” *Buzzfeed*, December 18, 2017. Эта мысль углубленно исследуется в сб.: Daniel Hillis, “The first machine intelligences,” in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (Penguin Press, 2019).

[10] Для своего времени статья Винера была редким примером расхождения с господствующим представлением, что любой технологический прогресс во благо: Norbert Wiener, “Some moral and technical consequences of automation,” *Science* 131 (1960): 1355–58.

> 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[11]. 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[12]. (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*[13]. Ваш прогноз

---

[11] Сантьяго Рамон-и-Кахаль в 1894 г. предположил, что изменения синапсов являются признаком обучения, но эта гипотеза была экспериментально подтверждена только в конце 1960-х гг. См.: Timothy Bliss and Terje Lomo, " Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path," *Journal of Physiology* 232 (1973): 331–56.

[12] Краткое введение см. в статье: James Gorman, "Learning how little we know about the brain," *The New York Times*, November 10, 2014. См. также: Tom Siegfried, "There's a long way to go in understanding the brain," *ScienceNews*, July 25, 2017. Специальный выпуск журнала *Neuron* в 2014 г. (vol. 94, pp. 933–1040) дает общее представление о множестве подходов к пониманию головного мозга.

[13] Наличие или отсутствие сознания – активного субъективного опыта – безусловно, принципиально важно для нашего отношения к машинам с точки зрения морали. Даже если бы мы знали достаточно, чтобы сконструировать сознающие машины или обнаружить тот факт, что нам это удалось, то столкнулись бы со множеством серьезных нравственных проблем, к решению большинства из которых не готовы.

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[14].

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[15].

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[16]. 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.

---

[14] Данная статья одной из первой установила четкую связь между алгоритмами обучения с подкреплением и нейрофизиологической регистрацией: Wolfram Schultz, Peter Dayan, and P. Read Montague, “A neural substrate of prediction and reward,” *Science* 275 (1997): 1593–99.

[15] Исследования внутричерепной стимуляции проводились в надежде найти средства лечения различных психических болезней. См., например: Robert Heath, “Electrical self-stimulation of the brain in man,” *American Journal of Psychiatry* 120 (1963): 571–77.

[16] Пример биологического вида, который может исчезнуть из-за зависимости: Bryson Voirin, “Biology and conservation of the pygmy sloth, Bradypus pygmaeus,” *Journal of Mammalogy* 96 (2015): 703–7.

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[17] и британский этолог Конви Ллойд Морган[18], 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[19]. 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[20]. 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 планирова-

---

[17] Появление понятия *эффект Болдуина* в эволюции обычно связывается со следующей статьей: James Baldwin, “A new factor in evolution,” *American Naturalist* 30 (1896): 441–51.

[18] Основная идея эффекта Болдуина также описывается в работе: Conwy Lloyd Morgan, *Habit and Instinct* (Edward Arnold, 1896).

[19] Современный анализ и компьютерная реализация, демонстрирующие эффект Болдуина: Geoffrey Hinton and Steven Nowlan, “How learning can guide evolution,” *Complex Systems* 1 (1987): 495–502.

[20] Дальнейшее раскрытие эффекта Болдуина в компьютерной модели, включающей эволюцию внутренней цепи сигнализации о вознаграждении: David Ackley and Michael Littman, “Interactions between learning and evolution,” in *Artificial Life II*, ed. Christopher Langton et al. (Addison-Wesley, 1991).

ние, 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*[21]. 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[22]:

> 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[23].

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 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,

---

[21] Здесь я указываю на корни нашего сегодняшнего понимания разума, а не описываю древнегреческое понятие *нус*, или «ум», имеющее много связанных друг с другом значений.

[22] Цит. в пер. Н. Брагинской. – *Прим. пер.*

[23] Цит. по: Aristotle, *Nicomachean Ethics*, Book III, 3, 1112b.

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.[24]) 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[25]. 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[26]. 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, связанным, но не совпадающим с денежной ценностью. Главное, полезность отличается *убывающей доходностью по отношению к деньгам*. Это означает, что полезность данной суммы денег не строго пропорциональна сумме, но возрастает медленнее ее. Например, полезность владения суммой в $1 000 000 100 намного меньше сотни полезностей владения $10 000 000. Насколько меньше? Спросите об этом себя! Какими должны быть шансы выиграть $1 млрд, чтобы это заставило вас отказаться от гаран-

[24] Кардано, один из первых европейских математиков, занимавшихся отрицательными числами, разработал раннюю математическую трактовку вероятности в играх. Он умер в 1576 г., за 87 лет до опубликования своего труда: Gerolamo Cardano, *Liber de ludo aleae* (Lyons, 1663).

[25] Работу Арно, впервые изданную анонимно, часто называют «Логикой Пор-Рояля» [по названию монастыря Пор-Рояль, аббатом которого являлся Антуан Арно. – *Прим. пер.*]: Antoine Arnauld, *La logique, ou l'art de penser* (Chez Charles Savreux, 1662). См. также: Blaise Pascal, *Pensées* (Chez Guillaume Desprez, 1670).

[26] Понятие полезности: Daniel Bernoulli, "Specimen theoriae novae de mensura sortis," *Proceedings of the St. Petersburg Imperial Academy of Sciences* 5 (1738): 175–92. Идея Бернулли о полезности вытекает из рассмотрения случая с купцом Семпронием, делающим выбор между перевозкой ценного груза одним судном или его разделением между двумя судами из соображения, что каждое судно имеет 50 %-ную вероятность затонуть в пути. Ожидаемая денежная полезность двух решений одинакова, но Семпроний, очевидно, предпочитает решение с двумя судами.

тированных $10 млн? Я задал этот вопрос своим студентам, и они ответили, что около 50 %, из чего следует, что ставка Б должна иметь ожидаемую ценность $500 млн, чтобы сравниться с желательностью ставки А. Позвольте повторить: ставка Б была бы в 50 раз выше ставки А в денежном выражении, но обе ставки имели бы равную полезность.

Введение понятия полезности – невидимого свойства – для объяснения человеческого поведения посредством математической теории было потрясающим для своего времени. Тем более что, в отличие от денежных сумм, ценность разных ставок и призов с точки зрения полезности недоступна для прямого наблюдения. Полезность приходится *выводить* из *предпочтений*, демонстрируемых индивидом. Пройдет два столетия, прежде чем практические выводы из этой идеи будут полностью разработаны и она станет общепринятой среди статистиков и экономистов.

В середине XX в. Джон фон Нейман (великий математик, в честь которого названа архитектура компьютеров – «архитектура фон Неймана»[27]) и Оскар Моргенштерн опубликовали *аксиоматическую* основу теории полезности[28]. Имеется в виду следующее: поскольку предпочтения, выражаемые индивидом, отвечают определенным базовым аксиомам, которым должен отвечать любой рациональный агент, выбор, сделанный этим индивидом, *неизбежно* может быть описан как максимизирующий ожидаемое значение функции полезности. Короче говоря, *рациональный агент действует так, чтобы максимизировать ожидаемую полезность*.

Трудно переоценить важность этого вывода. Во многих отношениях поиск ИИ заключается в том, чтобы выяснить, как именно строить рациональные машины.

Давайте подробнее рассмотрим аксиомы, которым, предположительно, должны удовлетворять рациональные сущности. Одна из них называется *транзитивностью*: если вы отдаете предпочтение А перед Б и Б перед В, то вы отдаете предпочтение А перед В. Это кажется вполне разумным! (Если пицца с сосисками нравится вам больше стандартной пиццы, а стандартная больше пиццы с ананасом, то представляется обоснованным предположить, что, выбирая между пиццей с сосисками и пиццей с ананасом, вы остановитесь на первой.) Вот еще одна аксиома, *монотонность*: если вы отдаете предпочтение призу А перед призом Б и можете выбирать между лотереями, единственными возможными выигрышами в которых являются А и Б, то предпочтете лотерею с наивысшей вероятностью выиграть приз А, а не Б. Опять-таки разумно!

Предпочтения касаются не только пиццы и денежных лотерей. Они могут быть связаны с чем угодно, в частности со всей будущей жизнью, вашей и других людей. Применительно к предпочтениям, касающимся последовательностей событий во времени, часто делается еще одно допущение – о так называемой *стационарности*: если два разных будущих, А и Б, начинаются с одного и того же события и вы отдаете предпочтение А перед Б, то будете предпочитать А и после того, как это событие произойдет. Это звучит разумно, но имеет на удивление значимое следствие: полезность любой цепи событий есть сумма вознаграждений, связанных с каждым событием (возможно, уценивающихся со временем на своего рода процентную ставку)[29]. Несмотря на повсеместную распространенность предположения о «полезности как сумме воз-

---

[27] По большинству свидетельств, сам фон Нейман не изобретал эту архитектуру, но его имя значилось на начальном варианте текста влиятельного отчета, описывающего вычислительную машину с запоминаемой программой EDVAC.

[28] Работа фон Неймана и Моргенштерна во многих отношениях является фундаментом современной экономической теории: John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, 1944).

[29] Предположение, что полезность есть сумма дисконтируемых вознаграждений, было сделано в форме математически приемлемой гипотезы Полом Самуэльсоном: Paul Samuelson, “A note on measurement of utility,” *Review of Economic Studies* 4 (1937): 155–61. Если s_0, s_1, … – последовательность состояний, то полезность в этой модели есть $U(s_0, s_1, \ldots) = \sum_t \gamma^t R(s_t)$, где γ – коэффициент дисконтирования, а R – функция вознаграждения, описывающая желательность состояния. Наивное применение этой модели редко согласуется с оценкой реальными индивидами желательности нынешнего и будущего вознаграждений. Тщательный анализ см. в статье: Shane Frederick, George Loewenstein, and Ted O’Donoghue, “Time discounting and time preference: A critical review,” *Journal of Economic Literature* 40 (2002): 351–401.

награждений» – восходящего по меньшей мере к XVIII в., к «гедонистическому исчислению» Джереми Бентама, основателя утилитаризма, – допущение стационарности, на котором оно основано, необязательно является свойством рационального агента. Стационарность исключает также вероятность того, что чьи-либо предпочтения могут меняться со временем, тогда как наш опыт свидетельствует об обратном.

Несмотря на разумность аксиом и важность выводов, которые из них следуют, на теорию полезности обрушивается шквал критики с тех самых пор, как она получила широкую известность. Некоторые отвергают ее за то, что она, предположительно, сводит все к деньгам и эгоизму. (Некоторые французские авторы презрительно называли эту теорию «американской»[30], несмотря на то что она уходит корнями во французскую мысль.) Действительно, что может быть разумнее, чем мечтать прожить жизнь в самоотречении, желая лишь уменьшить страдания других. Альтруизм заключается попросту в том, чтобы придавать существенный вес благополучию других при оценке любого конкретного будущего.

Другой комплекс возражений связан с трудностью получения необходимой оценки ценности возможностей и полезностей и их перемножения для расчета ожидаемой полезности. При этом просто смешиваются две разные вещи: выбор рационального действия и выбор его *путем вычисления ожидаемых полезностей*. Например, если вы пытаетесь ткнуть пальцем себе в глаз, веко опускается, чтобы защитить глазное яблоко; это рационально, но никакие расчеты ожидаемой полезности этому не сопутствуют. Можете также представить, что катитесь на велосипеде без тормозов вниз по склону и имеете возможность выбирать, врезаться в одну бетонную стену на скорости 16 км/ч или в другую, ниже по склону, на скорости 32 км/ч. Что вы предпочтете? Если вы выбрали 16 км/ч, мои поздравления! Вы вычисляли ожидаемую полезность? Вряд ли. Тем не менее выбор скорости 16 км/ч рационален. Это следует из двух базовых предположений: во-первых, что вы предпочитаете менее серьезные травмы более серьезным, во-вторых, что при любой тяжести травмы увеличение скорости столкновения повышает вероятность превышения этого уровня. Из этих двух предположений математически следует (совершенно без вычисления конкретных числовых значений), что столкновение на скорости 16 км/ч имеет более высокую ожидаемую полезность, чем столкновение на скорости 32 км/ч[31]. В общем, максимизация ожидаемой полезности необязательно требует вычисления каких-либо ожиданий или полезностей. Это чисто *внешнее* описание рациональной сущности.

Еще одна критика теории рациональности лежит в определении места принятия решений, то есть что рассматривается в качестве агентов. Кажется очевидным, что агентами являются люди. Но как быть с семьями, племенами, корпорациями, цивилизациями, государствами? Если обратиться к социальным насекомым, таким как муравьи, можно рассматривать индивидуального муравья как интеллектуального агента, или же интеллект связан со всей муравьиной колонией, с неким синтетическим мозгом, состоящим из мозгов и тел многих муравьев, взаимосвязанных феромонными сигналами, в отличие от сигналов электрических? С эволюционной точки зрения так думать о колонии муравьев, вероятно, более продуктивно, так как муравьи тесно связаны. Отдельно взятые муравьи, как и другие социальные насекомые, по-видимому, не обладают инстинктом самосохранения, в отличие от инстинкта сохранения колонии: они всегда вступают в битву против захватчиков, даже ценой собственной жизни. Иногда и люди поступают так же, чтобы защитить совсем чужих людей. Виду полезно наличие определенной доли индивидуумов, способных пожертвовать собой в бою, или отправиться в экспедиции в неизвестные земли, или воспитывать чужое потомство. В подобных случаях ана-

[30] Морис Алле, французский экономист, предложил сценарий принятия решения, в котором человек последовательно нарушает аксиомы фон Неймана – Моргенштерна: Maurice Allais, “Le comportement de l’homme rationnel devant le risque: Critique des postulats et axiomes de l’école américaine,” *Econometrica* 21 (1953): 503–46.

[31] Введение в анализ принятия неколичественных решений см. в: Michael Wellman, “Fundamental concepts of qualitative probabilistic networks,” *Artificial Intelligence* 44 (1990): 257–303.

лиз рациональности, основанный на интересах одного индивида, очевидно упускает из виду нечто существенное.

Другие принципиальные возражения против теории полезности носят эмпирический характер – они опираются на экспериментальные свидетельства, заставляющие предположить, что люди иррациональны. Мы систематически не угождаем аксиомам[32]. Я сейчас не ставлю своей целью отстоять теорию полезности как формальную модель человеческого поведения. Действительно, люди не всегда могут вести себя рационально. Наши предпочтения распространяются на всю собственную дальнейшую жизнь, жизни детей и внуков, а также других существ, которые живут сейчас или будут жить в дальнейшем. Тем не менее мы не можем даже сделать правильные ходы на шахматной доске, в крохотном и простом пространстве с четкими правилами и очень коротким горизонтом планирования. Причина не в иррациональности наших *предпочтений*, а в *сложности* проблемы принятия решения. В огромной мере наша когнитивная структура занята тем, что компенсирует несоответствие маленького медленного мозга непостижимо громадной сложности проблемы принятия решения, с которой мы постоянно сталкиваемся.

Таким образом, в то время как было бы весьма неразумно основывать теорию выгодного для нас ИИ на предположении, что люди рациональны, можно вполне заключить, что взрослый человек имеет довольно последовательные предпочтения относительно своей дальнейшей жизни. А именно – *если бы вы имели возможность посмотреть два фильма, каждый из которых достаточно подробно описывает вашу возможную будущую жизнь, вы могли бы сказать, какой вариант предпочитаете, или выразить безразличие к обоим*[33].

Это, возможно, чересчур сильное заявление, если наша единственная цель – гарантировать, чтобы развитие интеллектуальных машин не обернулось катастрофой для человеческой расы. Сама идея *катастрофы* предполагает жизнь, со всей определенностью не являющуюся предпочитаемой. Таким образом, чтобы избежать катастрофы, нам достаточно заявить, что взрослые люди способны опознать катастрофическое будущее, если оно показано подробно. Конечно, предпочтения людей имеют намного более детальную и, предположительно, проверяемую структуру, чем простое «отсутствие катастрофы лучше, чем катастрофа».

В действительности теория благотворного ИИ может принять во внимание непоследовательность человеческих предпочтений, но непоследовательную часть предпочтений невозможно удовлетворить, и ИИ здесь совершенно бессилен. Предположим, например, что ваши предпочтения в отношении пиццы нарушают аксиому транзитивности:

РОБОТ. Добро пожаловать домой! Хотите пиццу с ананасами?
ВЫ. Нет, пора бы знать, что я больше люблю обычную.
РОБОТ. Хорошо, обычная пицца уже готовится!
ВЫ. Нет уж, мне больше хочется пиццу с сосисками.
РОБОТ. Прошу прощения! Пожалуйста, вот пицца с сосисками!
ВЫ. Вообще-то, лучше уж с ананасами, чем с сосисками.
РОБОТ. Это мой промах, вот вам с ананасами!

---

[32] Я вернусь к рассмотрению свидетельств человеческой иррациональности в главе 9. Основные работы по данной теме: Allais, “Le comportement”; Daniel Ellsberg, *Risk, Ambiguity, and Decision* (PhD thesis, Harvard University, 1962); Amos Tversky and Daniel Kahneman, “Judgment under uncertainty: Heuristics and biases,” *Science* 185 (1974): 1124–31.

[33] Следует понимать, что это мысленный эксперимент, который невозможно поставить на практике. Выбор разных вариантов будущего никогда не предстает во всех деталях, и люди никогда не имеют роскошной возможности подробнейшим образом исследовать и оценить эти варианты, прежде чем выбирать. Мы получаем лишь краткие резюме, скажем, «библиотекарь» или «шахтер». Когда человек делает такой выбор, то в действительности ему предлагается сравнить два распределения вероятности по полным вариантам будущего, один из которых начинается с выбора «библиотекарь», а другой – с выбора «шахтер», причем каждое распределение предполагает оптимальные действия со стороны данного человека в рамках каждого будущего. Очевидно, сделать такой выбор непросто.

ВЫ. Я ведь уже сказал, что мне больше нравится обычная пицца, а не с ананасами.

Нет такой пиццы, которой робот мог бы вас осчастливить, потому что вы всегда предпочитаете какую-нибудь другую. Робот может удовлетворить только последовательную часть ваших предпочтений – например, если вы предпочитаете все три вида пиццы отсутствию пиццы. В этом случае услужливый робот мог бы подать вам любую из трех пицц, таким образом удовлетворив ваше предпочтение избежать «отсутствия пиццы» и предоставив вам на досуге обдумывать свои раздражающе непоследовательные предпочтения относительно ее ингредиентов.

Рациональность на двоих

Базовая идея, что рациональный агент действует так, чтобы максимизировать ожидаемую полезность, достаточно проста, даже если в действительности добиться этого сложно до невозможности. Теория, однако, применима только в случае, если агент действует в одиночку. При более чем одном агенте идея, что возможно – хотя бы в принципе – приписать вероятности разным результатам его действий, становится проблематичной. Дело в том, что теперь имеется часть мира – другой агент, – пытающаяся предугадать, какое действие вы собираетесь предпринять, и наоборот, поэтому становится неочевидной оценка вероятности того, как намерена вести себя эта часть мира. В отсутствии же вероятностей определение рационального действия как максимизации ожидаемой полезности неприменимо.

Таким образом, как только подключается кто-то еще, агенту требуется другой способ принятия рациональных решений. Здесь вступает в действие *теория игр*. Несмотря на название, теория игр необязательно занимается играми в обычном понимании; это попытка распространить понятие рациональности на ситуации с участием многих агентов. Очевидно, что это важно для наших целей, поскольку мы (пока) не планируем строить роботов, которые будут жить на необитаемых планетах других звездных систем; мы собираемся поместить роботов в наш мир, населенный нами.

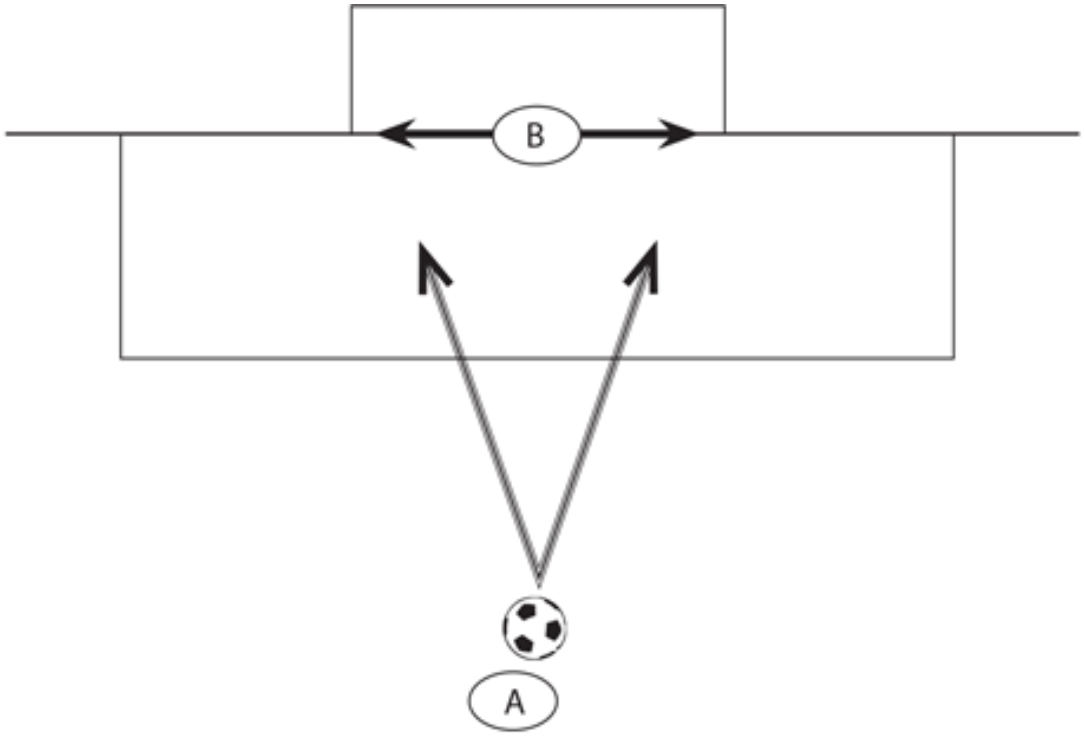


Рис. 3. Алиса готовится бить пенальти, играя против Боба

Чтобы прояснить, зачем нам нужна теория игр, рассмотрим простой пример: Алиса и Боб играют во дворе в футбол (рис. 3). Алиса готовится пробить пенальти, Боб стоит на воротах. Алиса собирается направить мяч справа или слева от Боба. Поскольку она правша, для нее проще и надежнее бить вправо от Боба. У Алисы мощный удар, и Боб знает, что должен броситься в одну либо в другую сторону – у него не будет времени подождать и узнать, куда летит мяч. Боб мог бы рассуждать так: «У Алисы больше шансов забить гол, если она пробьет справа от меня, поскольку она правша, значит, это она и выберет, и мне нужно броситься вправо». Однако Алиса не дурочка, она может представить этот ход рассуждений Боба и тогда пробьет влево. Поскольку Боб тоже не дурак и поймет, что замыслила Алиса, то бросится влево. Но Алиса умна и способна представить, что Боб думает именно так… В общем, вы поняли. Иными совами, если у Алисы есть рациональный выбор, Боб тоже может его обнаружить, предвосхитить и помешать Алисе забить гол, так что выбор, в принципе, не может быть рациональным.

Еще в 1713 г. – опять-таки в ходе анализа азартных игр – был найден выход из этого затруднительного положения[34]. Хитрость состоит в том, чтобы выбирать не какое-либо действие, а *рандомизированную стратегию*. Например, Алиса может выбрать стратегию «бить правее Боба с вероятностью 55 % и левее с вероятностью 45 %». Боб может выбрать «кидаться вправо с вероятностью 60 % и влево с вероятностью 40 %». Каждый мысленно бросает монету с соответствующей тенденцией перед каждым действием, чтобы не отклониться от своих намерений. Действуя *непредсказуемо*, Алиса и Боб избегают ограничений, описанных в предыду-

[34] Первое упоминание о рандомизированной стратегии в играх: Pierre Rémond de Montmort, *Essay d'analyse sur les jeux de hazard*, 2nd ed. (Chez Jacques Quillau, 1713). В книге упоминается некий монсеньор де Вальдграв в качестве автора оптимального рандомизированного решения для карточной игры Ле Гер. Сведения о личности Вальдграва раскрываются в статье: David Bellhouse, "The problem of Waldegrave," *Electronic Journal for History of Probability and Statistics* 3 (2007).

щем абзаце. Даже если Боб выяснит, в чем состоит рандомизированная стратегия Алисы, он бессилен справиться с ней, если у него нет «хрустального шара».

Следующий вопрос: какими *должны быть* вероятности? Рационален ли выбор Алисы, 55 % на 45 %? Конкретные значения зависят от того, насколько выше точность Алисы при ударе направо от Боба, насколько успешно Боб берет мяч, когда кидается вправо, и т. д. (Полный анализ см. в сносках[35].) Общий критерий, впрочем, очень прост:

1. Стратегия Алисы – лучшая, которую она может выбрать при условии, что Боб неподвижен.

2. Стратегия Боба – лучшая, которую он может выбрать при условии, что Алиса неподвижна.

Если выполняются оба условия, мы говорим, что стратегии находятся в равновесии. Такого рода равновесие называется *равновесием Нэша* в честь Джона Нэша, который в 1950 г. в возрасте 22 лет доказал, что оно существует для любого числа агентов с любыми рациональными предпочтениями, независимо от правил игры. После нескольких десятилетий борьбы с шизофренией Нэш выздоровел и в 1994 г. получил за эту работу Нобелевскую премию за достижения в экономических науках.

В футбольном матче Алисы и Боба равновесие лишь одно. В других случаях их может быть несколько. Таким образом, концепция равновесия Нэша, в отличие от решений на основе ожидаемой полезности, не всегда ведет к уникальным рекомендациям о том, как действовать.

Что еще хуже, бывают ситуации, когда равновесие Нэша может приводить к крайне нежелательным результатам. Одним из таких случаев является знаменитая «*дилемма заключенного*», название которой дал в 1950 г. научный руководитель Нэша Альберт Таккер[36]. Игра представляет собой абстрактную модель печально распространенных в реальном мире ситуаций, когда взаимодействие было бы лучше во всех смыслах, но люди тем не менее выбирают взаимное уничтожение.

Вот как работает «дилемма заключенного». Алиса и Боб подозреваются в преступлении и оказываются в одиночном заключении. У каждого есть выбор: признать вину и заложить подельника или отказаться давать показания[37]. Если оба откажутся, то будут обвинены в менее серьезном преступлении и отсидят два года; если оба сознаются, то получат более серьезное обвинение и сядут на 10 лет; если один сознается, а второй запирается, то сознавшийся выходит на свободу, а второй садится на 20 лет.

Итак, Алиса размышляет: «Если Боб решит признаться, то и мне следует признаваться (10 лет лучше, чем 20); если он планирует запираться, то мне лучше заговорить (выйти на свободу лучше, чем провести два года в тюрьме); так или иначе, нужно признаваться». Боб мыслит так же. В результате оба дают признательные показания и сидят 10 лет, тогда как, совместно отказавшись признавать вину, они могли бы отсидеть только два года. Проблема в том, что совместный отказ не является равновесием Нэша, потому что у каждого есть стимул предать другого и освободиться путем признания.

---

[35] Задача полностью определяется, если задать вероятность того, что Алиса забивает гол в каждом из следующих четырех случаев: если она бьет вправо от Боба, и Боб бросается вправо или влево, и если она бьет влево от Боба, и он бросается вправо или влево. В данном случае эти вероятности составляют 25, 70, 65 % и 10 % соответственно. Предположим, что стратегия Алисы – бить вправо от Боба с вероятностью p и влево с вероятностью 1 – p, тогда как Боб бросается вправо с вероятностью q и влево с вероятностью 1 – q. Выигрыш Алисы: $U_A = 0{,}25\ pq + 0{,}70\ p\ (1 - q) + 0{,}65\ (1 - p)q + 0{,}10\ (1 - p)\ (1 - q)$, Боба: $U_B = -U_A$. В равновесии $\partial U_A/\partial p = 0$ and $\partial U_B/\partial q = 0$, что дает $p = 0{,}55$ и $q = 0{,}60$.

[36] Исходную задачу теории игр предложили Меррил Флуд и Мелвин Дрешер в RAND Corporation. Такер увидел матрицу выигрышей, зайдя к ним в кабинет, и предложил сопроводить ее «историей».

[37] Специалисты теории игр обычно говорят, что Алиса и Боб смогли *сотрудничать* друг с другом (отказались давать показания) или предать подельника. Мне эти определения кажутся вводящими в заблуждение, поскольку «сотрудничество друг с другом» не тот выбор, который каждый агент может сделать индивидуально, а также из-за влияния общепринятого выражения «сотрудничать с полицией», когда за сотрудничество можно получить более легкий приговор и т. д.

Заметьте, что Алиса могла бы рассуждать следующим образом: «Как бы я ни мыслила, Боб тоже будет размышлять. В конце концов мы выберем одно и то же. Поскольку совместный отказ лучше совместного признания, нам нужно молчать». Эта разновидность рассуждения признает, что, будучи рациональными агентами, Алиса и Боб сделают согласующийся выбор, а не два независимых. Это лишь один из многих подходов, опробованных в теории игр в попытке получить менее удручающие решения «дилеммы заключенного»[38].

Другой знаменитый пример нежелательного равновесия – *трагедия общих ресурсов*, впервые проанализированная в 1833 г. английским экономистом Уильямом Ллойдом[39], хотя дал ей название и привлек к ней внимание всего мира эколог Гаррет Хардин в 1968 г.[40] Проблема возникает, если несколько человек могут использовать ограниченный и медленно восполняемый ресурс – например, общее пастбище или рыбный пруд. В отсутствие любых социальных или юридических ограничений единственное равновесие Нэша для эгоистичных (неальтруистичных) агентов заключается в том, чтобы каждый потреблял как можно больше, что вело бы к быстрому исчерпанию ресурса. Идеальное решение, при котором все пользуются ресурсом так, чтобы общее потребление было устойчивым, не является равновесием, поскольку у каждого индивида есть стимул хитрить и брать больше справедливой доли – перекладывая издержки на других. На практике, конечно, люди предпринимают меры во избежание этой ситуации, создавая такие механизмы, как квоты и наказания или схемы ценообразования. Они могут это сделать, потому что не ограничены в решении о том, сколько потреблять; кроме того, они имеют возможность принять решение осуществлять *коммуникацию*. Расширяя проблему принятия решения подобным образом, мы находим выходы, лучшие для каждого.

Эти и многие другие примеры иллюстрируют тот факт, что распространение теории рациональных решений на несколько агентов влечет за собой много видов интересного и сложного поведения. Это крайне важно еще и потому, очевидно, что людей на свете больше одного. Скоро к ним присоединятся интеллектуальные машины. Незачем говорить, что мы должны достичь взаимной кооперации, влекущей за собой пользу для людей, а не взаимное уничтожение.

---

[38] Интересное решение на основе доверия для дилеммы заключенного и других игр см.: Joshua Letchford, Vincent Conitzer, and Kamal Jain, “An ‘ethical’ game-theoretic solution concept for two-player perfect-information games,” in *Proceedings of the 4th International Workshop on Web and Internet Economics*, ed. Christos Papadimitriou and Shuzhong Zhang (Springer, 2008).

[39] Источник трагедии *общих ресурсов*: William Forster Lloyd, *Two Lectures on the Checks to Population* (Oxford University, 1833).

[40] Современное рассмотрение темы в контексте глобальной экологии: Garrett Hardin, “The tragedy of the commons,” *Science* 162 (1968): 1243–48.

Компьютеры

Рациональное определение интеллектуальности – первый компонент в создании интеллектуальных машин. Вторым компонентом является машина, в которой это определение может быть реализовано. По причинам, которые скоро станут очевидными, эта машина – компьютер. Это *могло бы быть* нечто другое, например мы могли бы попытаться сделать интеллектуальные машины на основе сложных химических реакций или путем захвата биологических клеток[41], но устройства, созданные для вычислений, начиная с самых первых механических калькуляторов всегда казались своим изобретателям естественным вместилищем разума.

Мы сегодня настолько привыкли к компьютерам, что едва замечаем их невероятные возможности. Если у вас есть десктоп, ноутбук или смартфон, посмотрите на него: маленькая коробочка с возможностью набора символов. Одним лишь набором вы можете создавать программы, превращающие коробочку в нечто другое, например, способное волшебным образом синтезировать движущиеся изображения океанских кораблей, сталкивающихся с айсбергами, или других планет, населенных великанами. Набираете еще что-то, и коробочка переводит английский текст на китайский язык; еще что-то – она слушает и говорит, еще – побеждает чемпиона мира по шахматам.

Способность осуществлять любой процесс, который приходит вам в голову, называется *универсальностью*. Эту концепцию ввел Алан Тьюринг в 1936 г.[42] Универсальность означает, что нам не нужны отдельные машины для вычислений, машинного перевода, шахмат, распознавания речи или анимации: все это делает одна машина. Ваш ноутбук, в сущности, подобен огромным серверам крупнейших IT-компаний – даже тех, которые оборудованы причудливыми специализированными тензорными процессорами для машинного обучения. Он также по сути идентичен всем компьютерным устройствам, которые еще будут изобретены. Ноутбук может выполнять те же самые задачи при условии, что ему хватает памяти; это лишь занимает намного больше времени.

Статья Тьюринга, где вводилось понятие универсальности, стала одной из важнейших когда-либо написанных статей. В ней он рассказал о простом вычислительном устройстве, способном принимать в качестве входного сигнала описание любого другого вычислительного устройства вместе с входным сигналом этого второго устройства и, симулируя операции второго устройства на своем входе, выдавать тот же результат, что выдало второе устройство. Теперь мы называем это первое устройство *универсальной машиной Тьюринга*. Чтобы доказать его универсальность, Тьюринг ввел точные определения двух новых типов математических объектов: машин и программ. Вместе машина и программа определяют последовательность событий, а именно – последовательность изменений состояния в машине и в ее памяти.

В истории математики новые типы объектов возникают довольно редко. Математика началась с чисел на заре письменной истории. Затем, около 2000 г. до н. э., древние египтяне и вавилоняне стали работать с геометрическими объектами (точками, линиями, углами, областями и т. д.). Китайские математики в течение I тыс. до н. э. ввели матрицы, тогда как группы математических объектов появились лишь в XIX в. Новые объекты Тьюринга – машины и программы – возможно, самые мощные математические объекты в истории. Ирония заключается

---

[41] Весьма вероятно, что, даже если бы мы попытались создать интеллектуальные машины на основе химических реакций или биологических клеток, объединения этих элементов оказались бы реализацией машины Тьюринга нетрадиционным способом. Вопрос о том, является ли объект универсальным компьютером, никак не связан с вопросом о том, из чего он сделан.

[42] Эпохальная статья Тьюринга дает определение понятию, в настоящее время известному как *машина Тьюринга*, основополагающему в компьютерной науке. *Entscheidungsproblem*, или *проблема принятия решения*, в названии статьи есть проблема выбора следования в логике первого порядка: Alan Turing, “On computable numbers, with an application to the Entscheidungsproblem,” *Proceedings of the London Mathematical Society*, 2nd ser., 42 (1936): 230–65.

в том, что сфера математики по большей части не сумела этого признать и с 1940-х гг. и до настоящего времени компьютеры и вычисления остаются в большинстве крупнейших университетов вотчиной инженерных факультетов.

Возникшая область знания – компьютерная наука – последующие 70 лет бурно развивалась, создав великое множество новых понятий, конструкций, методов и применений, а также семь из восьми самых ценных компаний в мире.

Центральным для компьютерной науки является понятие *алгоритма* – точно определенного метода вычисления чего-либо. Сейчас алгоритмы являются привычным элементом повседневной жизни. Алгоритм вычисления квадратного корня в карманном калькуляторе получает на входе число и выдает на выходе квадратный корень этого числа; алгоритм игры в шахматы принимает позицию на доске и выдает ход; алгоритм поиска маршрута получает стартовое местоположение, целевую точку и карту улиц и выдает более быстрый путь из отправной точки к цели. Алгоритмы можно описывать на английском языке или в виде математической записи, но, чтобы они были выполнены, их нужно закодировать в виде программ на *языке программирования*. Сложные алгоритмы можно построить, используя простые в качестве кирпичей, так называемые *подпрограммы*, – например, машина с автопилотом может использовать алгоритм поиска маршрута как подпрограмму, благодаря чему будет знать, куда ехать. Так, слой за слоем, строятся бесконечно сложные программные системы.

Аппаратная часть компьютера также важна, поскольку более быстрые компьютеры с большей памятью позволяют быстрее выполнять алгоритм и включать больше информации. Прогресс в этой сфере хорошо известен, но по-прежнему не укладывается в голове. Первый коммерческий программируемый электронный компьютер, Ferranti Mark I, мог выполнять около 1000 (10^3) команд в секунду и имел примерно 1000 байт основной памяти. Самый быстрый компьютер начала 2019 г., Summit, Национальной лаборатории Ок-Ридж в Теннесси выполняет около 10^{18} команд в секунду (в 1000 трлн раз быстрее) и имеет $2{,}5 \times 10^{17}$ байт памяти (в 250 трлн раз больше). Этот прогресс стал результатом совершенствования электронных устройств и развития стоящей за ними физики, что позволило добиться колоссальной степени миниатюризации.

Хотя сравнение компьютера и головного мозга, в общем, лишено смысла, замечу, что показатели Summit слегка превосходят емкость человеческого мозга, который, как было сказано, имеет порядка 10^{15} синапсов и «цикл» примерно в 0,01 секунды с теоретическим максимумом около 10^{17} «операций» в секунду. Самым существенным различием является потребление энергии: Summit использует примерно в миллион раз больше энергии.

Предполагается, что закон Мура, эмпирическое наблюдение, что количество электронных компонентов чипа удваивается каждые два года, продолжит выполняться примерно до 2025 г., хотя и немного медленнее. Сколько-то лет скорости ограничены большим количеством тепла, выделяемого при быстрых переключениях кремниевых транзисторов; более того, невозможно значительно уменьшить размеры цепей, поскольку провода и соединения (на 2019 г.) уже не превышают длины в 25 атомов и толщины от пяти до десяти атомов. После 2025 г. нам придется использовать более экзотические физические явления, в том числе устройства отрицательной емкости[43]

[43] Хорошее исследование отрицательной емкости от одного из ее изобретателей: Sayeef Salahuddin, “Review of negative capacitance transistors”, in *International Symposium on VLSI Technology, Systems and Application* (IEEE Press, 2016).

Конец ознакомительного фрагмента.