ГЭРИ МАРКУС | ЭРНЕСТ ДЭВИС

КАК СОЗДАТЬ МАШИННЫЙ РАЗУМ, КОТОРОМУ ДЕЙСТВИТЕЛЬНО МОЖНО ДОВЕРЯТЬ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ПЕРЕЗАГРУЗКА

PRO

Эрнест Дэвис

Искусственный интеллект: перезагрузка. Как создать машинный разум, которому действительно можно доверять

«Альпина Диджитал»

2019

Дэвис Э.

Искусственный интеллект: перезагрузка. Как создать машинный разум, которому действительно можно доверять / Э. Дэвис — «Альпина Диджитал», 2019

ISBN 978-5-20-600030-6

Работы по развитию искусственного интеллекта сегодня занимают умы тысяч людей по всему миру. Инженерам уже удалось достичь потрясающих результатов в этом направлении, но искусственному интеллекту пока еще очень далеко до уровня человеческого. Перед учеными стоит крайне сложная задача создать безопасную для людей мыслящую машину, которой мы сможем доверить аспекты жизни, связанные с образованием, медициной, строительством, транспортом. В своей книге известные исследователи в области ИИ объясняют, что нужно сделать, чтобы умные роботы вышли на новый уровень. Как наделить машины здравым смыслом и глубоким умом? Каковы перспективы современной науки в сфере ИИ? Как новое поколение ИИ может сделать нашу жизнь лучше и как снизить риски, связанные с его развитием?

ISBN 978-5-20-600030-6

Содержание

Гэри Маркус, Эрнест Дэвис
Искусственный интеллект: перезагрузка. Как создать машинный разум, которому действительно можно доверять

Перевод В. Скворцов
Редактор А. Марченкова
Руководители проекта А. Марченкова, Ю. Семенова
Дизайн обложки А. Маркович
Корректоры Н. Витько, Е. Якимова
Верстка Б. Руссо

* * *

ГЭРИ МАРКУС, ЭРНЕСТ ДЭВИС

Искусственный интеллект: перезагрузка

КАК СОЗДАТЬ МАШИННЫЙ РАЗУМ, КОТОРОМУ ДЕЙСТВИТЕЛЬНО МОЖНО ДОВЕРЯТЬ

Перевод с английского

МОСКВА
2021

Моим детям, Александру и Хлое,
которые научили меня так многому,
и моей жене Афине, понимающей,
как и я, как здорово учиться у детей.

Гэри

Моей жене Бьянке,
ставшей любовью на всю жизнь.

Эрни

Глава 1
Осторожно: разрыв

В течение двадцати лет машины смогут научиться выполнять любую работу, доступную для человека.
Херб Саймон, один из первопроходцев в области искусственного интеллекта

ПЕРВЫЙ РЕБЕНОК (в долгой утомительной поездке): Еще долго, Папа Смурф?
ОТЕЦ: Уже скоро.
ВТОРОЙ РЕБЕНОК (через несколько часов): Еще долго, Папа Смурф?
ОТЕЦ: Уже скоро.
Из комикса The Smurfs («Смурфики»)

С самого момента своего зарождения искусственный интеллект (ИИ) очень много сулил и очень мало давал. Уже в 1950-х и 1960-х годах такие первопроходцы, как Марвин Мински, Джон Маккарти и Херб Саймон, искренне верили[1], что до конца XX века со всеми проблемами, встающими на пути разработки ИИ, будет фактически покончено. «В течение одного поколения, – писал Марвин Мински в 1967 году в своем знаменитом послании, – проблема создания искусственного интеллекта будет по большому счету решена». Пятьдесят лет спустя эти обещания так и не исполнились, однако заявления подобного рода продолжали появляться. В 2002 году футуролог Рэй Курцвейл публично заявил, что к 2029 году ИИ «превзойдет естественный человеческий интеллект». В ноябре 2018 года Илья Суцкевер, соучредитель Open AI, крупного исследовательского института по изучению искусственного интеллекта, высказал мнение, что универсальный искусственный интеллект «уже в ближайшей перспективе следует воспринимать всерьез как возможность». Хотя теоретически допустимо, что Курцвейл и Суцкевер могут оказаться правы, весьма велики шансы, что ничего такого и не случится. Достижение нового уровня – универсального искусственного интеллекта с гибкостью человеческого мышления – для нынешнего поколения отнюдь не находится на расстоянии вытянутой руки; это долгий путь, который потребует очень серьезного прогресса в целом ряде основополагающих областей знания. Иначе говоря, это не просто «чуть более» того, чего науке удалось добиться за последние несколько лет: мы покажем, что универсальный ИИ – нечто совершенно иное.

Даже если не все так оптимистичны, как Курцвейл и Суцкевер, амбициозные обещания продолжают фигурировать везде, где задействован искусственный интеллект: от медицинских технологий до беспилотных автомобилей. Чаще всего обещанное не исполняется. Например, в 2012 году мы много слышали о том, как увидим «в ближайшем будущем автономные автомобили». В 2016 году IBM заявила, что Watson, система искусственного интеллекта, сумевшая принять участие в телевикторине Jeopardy! произведет «революцию в здравоохранении», потому что якобы «когнитивные системы Watson Health могут понимать, рассуждать, учиться и взаимодействовать», и что «с помощью последних достижений в области когнитивных вычислений мы можем достичь большего, чем мы когда-либо считали возможным». Ком-

[1] Minsky 1967: 2 (цитируется буквально). Маккарти утверждает: «Мы полагаем, что в решении одной или целого ряда [перечисленных выше] проблем можно достичь значительного прогресса, если специально выбранная группа [квалифицированных] ученых поработает над этим в течение буквально нескольких месяцев»: см. McCarthy, Minsky, Rochester, and Shannon 1955. Херб Саймон (Simon, 1965: 96) дословно процитирован, как указано в эпиграфе к настоящей главе.

пания IBM стремилась решить любые проблемы, начиная от фармакологии и радиологии и заканчивая диагностикой и лечением рака, используя Watson для чтения медицинской литературы и выработки рекомендаций, которые врачи-люди могли бы упустить из виду. Вторя им, Джеффри Хинтон, один из самых выдающихся исследователей искусственного интеллекта, заявил примерно в то же время: «Совершенно очевидно, что мы должны прекратить подготовку [людей-]радиологов».

В 2015 году Facebook запустил свой амбициозный и широко освещаемый проект, известный как «M», – чат-бот, который должен был удовлетворить любые потребности пользователя, от бронирования столика в ресторане до планирования очередного отпуска.

Но ничего из этого пока что не произошло. Да, автономные транспортные средства когда-нибудь могут стать безопасными и способными ездить по любым дорогам. Чат-боты, которые будут способны удовлетворить любые потребности, наверное, тоже однажды превратятся в обычное явление; что-то подобное может рано или поздно произойти и в медицине, когда появятся суперинтеллектуальные врачи-роботы. Но пока что все это остается лишь фантазией, а не свершившимся фактом.

Существующие беспилотные автомобили по-прежнему способны функционировать только в условиях автомагистралей, причем люди-водители все равно должны в них находиться и оставаться начеку, поскольку программное обеспечение этих машин слишком ненадежно для нестандартных ситуаций. В 2017 году Джон Крафчик, генеральный директор Waymo, дочерней компании Google, которая почти десять лет работает над беспилотными автомобилями, хвастался, что вскоре у Waymo появятся беспилотные автомобили… без водителей (что уже весьма забавно); впрочем, этого все равно не случилось. Год спустя, как выразился журнал *Wired*, бравада исчезла, а водители (в качестве резервного средства безопасности) – нет. Никто на самом деле и не думает, что автомобили без водителя готовы самостоятельно ездить в городах или хотя бы в плохую погоду по шоссе, и ранний оптимизм сменился общим признанием того, что мы находимся на расстоянии как минимум десятилетия от подобного прорыва – а возможно, для этого потребуется куда больше времени.

Использование системы IBM Watson в сфере здравоохранения также потеряло популярность. В 2017 году сотрудничество с IBM в области диагностики и борьбы с раковыми заболеваниями прекратил онкологический центр MD Anderson. Совсем недавно стало известно, что некоторые рекомендации IBM Watson оказались «небезопасными и неверными». Проект 2016 года по использованию Watson для диагностики редких заболеваний в Марбурге (Германия), в Центре редких и недиагностируемых заболеваний (Marburg's Center for Rare and Undiagnosed Diseases), продержался менее двух лет и полностью остановился, поскольку «эффективность работы системы была неприемлема». Например, в одном случае IBM Watson исследовала пациента, страдавшего от болей в груди, и пропустила диагнозы, которые были бы очевидны даже студенту-первокурснику, например сердечный приступ, стенокардию или разрыв аорты. Когда проблемы, связанные с системой Watson, стали все больше и больше проникать в общественное сознание, проект «M» от Facebook тихо прикрыли, и произошло это всего через три года после гордых заявлений о его универсальной полезности.

Несмотря на уже солидную историю несостоявшихся свершений, риторика вокруг искусственного интеллекта продолжает оставаться почти мессианской. Так, Эрик Шмидт, бывший генеральный директор Google, заявил, что развитие ИИ решит проблемы изменения климата, бедности, войны и рака. Основатель компании XPRIZE Питер Диамандис сделал аналогичные заявления в своей книге «Изобилие» (Abundance), утверждая, что могучий искусственный интеллект (стоит ему только появиться) «без сомнения, умчит нас прямо к "пирамиде Изобилия"». В начале 2018 года генеральный директор Google Сундар Пичаи уверял, что «искусственный интеллект – одна из самых важных вещей, над которыми работает человечество… более основополагающая, чем … электричество или огонь». (Менее чем через год после этого

выступления Google была вынуждена признать в сообщении для инвесторов, что продукты и услуги, «которые включают в себя или используют искусственный интеллект и машинное обучение, могут вызвать новые или усугубить существующие этические, технологические, юридические и другие проблемы».)

Многие мыслители всерьез беспокоятся и по поводу потенциальных опасностей, таящихся в искусственном интеллекте, причем способы, которыми это делается, явно демонстрируют отрыв суждений от реальности. Один из недавних научно-популярных бестселлеров оксфордского философа Ника Бострома описывает перспективу завоевания мира некой сверхразведкой в таких словах, будто это действительно может стать серьезной угрозой в обозримом будущем. На страницах The Atlantic Генри Киссинджер предполагает, что риски, связанные с искусственным интеллектом, способны оказаться настолько большими, что «человеческая история может пойти по пути инков, столкнувшихся с непостижимой и даже внушающей им священный страх испанской культурой». Илон Маск считает, что работа над совершенствованием ИИ – это «обряд заклинания демонов», по своей опасности «страшнее ядерного оружия», а покойный гений физики Стивен Хокинг предупреждал, что искусственный интеллект может сделаться «самым худшим событием в истории нашей цивилизации».

Но о каком именно искусственном интеллекте все они говорят? Возвращаясь в реальный мир, мы видим, что современные роботы едва справляются с тем, чтобы повернуть обычную дверную ручку, а знаменитая «Тесла», управляемая ИИ в режиме «автопилот», врезается сзади в припаркованные машины скорой помощи (только в 2018 году такое случалось как минимум четырежды). Это все равно, что люди в XIV веке переживали бы о скором наступлении гибельной эры дорожно-транспортных происшествий, хотя в то время гораздо полезнее было бы беспокоиться о приличной гигиене.

Одна из причин, по которой люди часто переоценивают возможности искусственного интеллекта, заключается в том, что сообщения, появляющиеся в СМИ, часто до такой степени преувеличивают его возможности, что любое самое скромное продвижение в технологиях начинает выглядеть как «прорыв тысячелетия». Рассмотрим следующую пару заголовков, описывающих «невероятный прогресс» в области машинного чтения.

> Отныне роботы смогут читать лучше, чем люди, подвергая риску существование миллионов рабочих мест.
> ***Newsweek, 15 января 2018 года***

> Компьютеры становятся лучшими читателями, чем мы сами.
> ***CNN Money, 16 января 2018 года***

Первое из этих утверждений является куда более вопиющим преувеличением, чем второе, но оба они представляют собой откровенную дичь, подавая незначительный прогресс в области компьютерного чтения как новость мировой значимости. Начнем с того, что в действительности в эксперименте не был задействован ни один робот, а сам тест оценивал лишь один крошечный аспект машинного чтения. Речь даже не шла о каком-либо понимании текста искусственным интеллектом, не говоря уже о самой отдаленной угрозе каким бы то ни было рабочим местам.

А случилось, собственно, вот что. Две компании, Microsoft и Alibaba, только что создали программы, которые добились незначительного (и не внезапного) прогресса (82,65 % точности против предыдущего показателя в 82,136 %) в конкретном тестировании одного узкого аспекта чтения, известного как SQuAD (the Stanford Question Answering Dataset, то есть набор вопросов и ответов, разработанный Стэндфордским университетом). Вероятно, мы можем здесь говорить о достижении уровня человеческой эффективности в этой конкретной задаче, в которой искусственный интеллект раньше немного отставал от людей, но одна из компаний выпустила

по этому поводу пресс-релиз, который сделал незначительное достижение звучащим почти революционно, объявив о создании «искусственного интеллекта, который может читать документ и отвечать на вопросы о нем так же хорошо, как и человек».

Реальность была намного менее будоражащей. Компьютерам показывали короткие отрывки текста, взятые из задания, предназначенного для исследовательских целей, и затем задавали вопросы о них. Подвох был в том, что в каждом случае правильные ответы находились прямо в тексте, что превращало задание не более чем в подчеркивание нужных слов. Незатронутой оставалась реальная проблема машинного чтения: обнаружение значений слов или предложений, которые подразумеваются, но не видны в явной форме.

Предположим, например, что мы даем вам лист бумаги с небольшим отрывком текста:

> Двое детей, Хлоя и Александр, пошли гулять. Они оба увидели собаку и дерево. Еще Александр увидел кошку и показал ее Хлое. А та пошла эту кошку погладить[2].

Ответить на вопросы типа «Кто пошел погулять?», естественно, очень легко, ведь ответ («Хлоя и Александр») прямо прописан в тексте. Однако любой компетентный (на самом деле – просто обычный) читатель должен так же легко ответить на вопросы, ответы на которые отсутствуют в тексте в утвердительной форме, например: «Видела ли Хлоя кошку?» или «Испугала ли кошка детей?» Если вы не можете этого сделать, значит, вы просто не обратили внимания на то, о чем шла речь. Поскольку SQuAD не включал в себя никаких вопросов подобного рода, то он не являлся по-настоящему серьезным тестом на способность к чтению; и на самом деле новые системы искусственного интеллекта попросту не смогли бы с ним справиться. Чтобы продемонстрировать различие между машиной и человеком, Гэри предложил этот тест своей дочери Хлое, которой тогда было четыре с половиной года. Настоящая Хлоя без труда сделала вывод о том, что Хлоя вымышленная действительно видела кошку. (Ее старший брат, которому тогда еще не исполнилось шести лет, пошел еще дальше, размышляя о том, что произойдет, если собака на самом деле окажется кошкой, – ни одна из форм нынешнего искусственного интеллекта не сможет даже близко подойти к этому.)

Практически каждый раз, когда один из мировых технологических гигантов выпускает пресс-релиз, мы имеем повторение того, о чем шла речь выше: незначительный прогресс изображается во многих (к счастью, не во всех) СМИ как настоящая революция. Например, пару лет назад Facebook представила абсолютно сырую программу, которая читала простые рассказы и отвечала на вопросы о них. За этим последовало множество восторженных заголовков, таких как «Представители Facebook полагают, что компания разгадала секрет того, как сделать чат-боты менее тупыми» (*Slate*) и «Facebook AI Software учится и отвечает на вопросы. Программное обеспечение, способное прочитать краткий пересказ "Властелина колец" и ответить на вопросы о нем, может кардинально улучшить поиск в Facebook» (*Technology Review*).

Тут действительно можно было бы говорить о настоящем прорыве – будь все это правдой. Программа, которая могла бы усвоить книгу Толкина хотя бы в версии *Reader's Digest* или *Cliffs-Notes* (не говоря уже о полноразмерных произведениях), была бы серьезным достижением в области искусственного интеллекта.

Но, увы, программы, действительно способной на такие подвиги, что-то нигде не видно. Тот пересказ, который на самом деле читала система Facebook, представлял собой всего лишь следующие строки:

---

[2] Кажущиеся еще более простыми вопросы типа «Что увидел Александр?» были бы целиком за допустимыми для компьютеров пределами, потому что ответ на них (собака, дерево и кошка) требует выделения двух несмежных фрагментов текста, в то время как SQuAD облегчал машинам работу, ограничивая вопросы теми, на которые можно ответить, используя связанный текстовый фрагмент.

> Бильбо отправился в пещеру. Голлум обронил там кольцо. Бильбо взял кольцо. Бильбо вернулся в Шир. Бильбо оставил кольцо там. Фродо получил кольцо. Фродо отправился на Роковую Гору. Фродо бросил кольцо туда. Саурон умер. Фродо вернулся в Шир. Бильбо отправился в Серые Гавани. Конец.

И даже при таком примитивном раскладе все, что могла сделать программа, – это отвечать на элементарные вопросы, ответы на которые содержались непосредственно в приведенных выше предложениях, например: «Где кольцо?», «Где сейчас Бильбо?» и «Где сейчас Фродо?» И забудьте о вопросах наподобие «Почему Фродо бросил кольцо?».

Конечная цель шумихи, поднятой в средствах массовой информации и сильно преувеличивающей технологический прогресс, заключается в том, чтобы общественность поверила, что проблема создания искусственного интеллекта гораздо ближе к решению, чем есть на самом деле.

Всякий раз, когда вы слышите об очередном успехе, достигнутом искусственным интеллектом, попробуйте задать, скажем, шесть вопросов из следующего списка.

1. Если отбросить риторику, что на самом деле совершила система искусственного интеллекта в этот раз?

2. Насколько универсальным оказался результат? Например, задание якобы на тестирование чтения включает в себя все составляющие нормального чтения или только незначительные и частные его аспекты?

3. Создана ли демонстрационная версия, на которой я могу протестировать систему, пользуясь собственными примерами? Если ее нет, успех выглядит более чем сомнительным.

4. Если исследователи (или их представители в прессе) утверждают, что система искусственного интеллекта что-то умеет лучше, чем люди, то о каких людях идет речь и насколько система превосходит подобных людей?

5. Насколько успех в решении конкретной задачи, о которой сообщается в новом исследовании, ведет нас к созданию универсального, подлинного искусственного интеллекта?

6. Насколько устойчива система, о которой пишут в прессе? Может ли она хорошо работать с другими наборами данных без огромной работы по предварительной их подготовке? Например, может ли игровой автомат, который овладел игрой в шахматы, успешно играть в приключенческую игру типа Zelda? Может ли система распознавания животных правильно идентифицировать существо, которое она никогда раньше не воспринимала как животное? Будет ли система автопилота, которая обучалась в дневное время на шоссе с указателями, способна ездить ночью, или по снегу, или если на ее карте нет указателя объезда?

Эта книга не просто о том, как не быть слишком легковерным человеком, но и о том, почему искусственный интеллект до сих пор развивается далеко не самым правильным образом, и, наконец, о том, что следовало бы сделать, чтобы создать такие мыслящие машины, которые смогли бы работать надежно и устойчиво и были бы способны функционировать в сложном и постоянно меняющемся мире так, чтобы мы могли спокойно доверять им наши дома, наших родителей и детей, наше медицинское обслуживание и, в конечном счете, всю нашу жизнь.

Нельзя отрицать и того, что искусственный интеллект в последние несколько лет впечатляет нас по-новому почти каждый день, порой даже творит чудеса. Значительные успехи появились в самых разных областях, от компьютерных игр до распознавания речи и идентификации лиц. Вот пример нового проекта, который нам искренне нравится: молодая компания Zipline использует (в умеренных дозах) искусственный интеллект, чтобы управлять беспилот-

ными аппаратами, доставляющими донорскую кровь пациентам в Африке, – почти фантастическое решение, о котором не могло быть и речи несколько лет назад.

Успех в области искусственного интеллекта, о котором мы говорим, был обусловлен главным образом двумя факторами: во-первых, достижениями в аппаратном обеспечении, которые позволяют увеличить объем памяти и ускорить вычисления (часто благодаря использованию множества машин, работающих параллельно); во-вторых, большими данными – огромными наборами, содержащими гигабайты, терабайты или более информации, чего не было еще несколько лет назад; например, такие базы, как ImageNet – библиотека из 15 млн маркированных изображений, которая сыграла ключевую роль в обучении систем ИИ компьютерному зрению, проект Wikipedia и, наконец, огромные коллекции документов, которые вместе и составляют то, что мы называем Всемирной паутиной.

Вместе с большими данными появился и алгоритм для сбора этих данных, называемый глубоким обучением, – своеобразный, весьма мощный статистический механизм, суть которого мы объясним и проанализируем в главе 3. Глубокое обучение оказалось в центре практически любого серьезного прорыва в области искусственного интеллекта за последние несколько лет, от сверхчеловеческого DeepMind, победившего человека в го, и шахматной системы AlphaZero до новейших инструментов Google, способных синтезировать речь и разговоры (Google Duplex). В каждом случае рецептом победы были большие данные плюс глубокое обучение плюс более мощное и быстрое оборудование.

Глубокое обучение использовалось с большим успехом и для широкого круга практических задач, от диагностики рака кожи до прогнозирования подземных толчков и выявления мошенничества с кредитными картами. Оно нашло применение в изобразительном искусстве, в музыке, в огромном числе коммерческих проектов от расшифровки речи до маркировки фотографий и организации новостных лент в интернете. Вы можете использовать глубокое обучение для идентификации растений, для автоматического улучшения цвета неба на фотографиях и даже для раскрашивания старых черно-белых изображений.

Вместе с ошеломляющим успехом глубокого обучения искусственный интеллект превратился в огромный бизнес. Гигантские информационные корпорации, подобные Google и Facebook, ведут грандиозные сражения за талантливых ученых, нередко предлагая сотрудникам с докторскими степенями такую зарплату, какую мы могли бы представить разве что у профессиональных спортсменов. В 2018 году билеты на самую важную научную конференцию по глубокому обучению были распроданы за двенадцать минут. Хотя мы будем постоянно доказывать, что создать искусственный интеллект с гибкостью мышления на уровне человека гораздо сложнее, чем думают многие, нет никаких сомнений в том, что в последнее десятилетие достигнут реальный прогресс в частных сферах применения ИИ. Поэтому вполне закономерно, что широкую публику так волнует все, что связано с данной областью.

Естественно, это волнует и правительства самых разных государств. Такие страны, как Франция, Россия, Канада и Китай, взяли на себя огромные обязательства по развитию искусственного интеллекта. Один только Китай планирует к 2030 году инвестировать в эту сферу 150 млрд долларов. По оценкам Глобального института McKinsey, общее экономическое воздействие искусственного интеллекта можно оценить в 13 трлн долларов, что сопоставимо (по относительному уровню влияния) с паровым двигателем в XIX веке и информационными технологиями в XXI. Тем не менее это не гарантирует того, что мы находимся на правильном пути.

Действительно, даже теперь, когда данных намного больше, компьютеры стали существенно быстрее, а инвестиции увеличились в несколько раз, важно понимать, что чего-то фундаментального во всем этом по-прежнему не хватает. Несмотря на бесспорный прогресс, машины во многих отношениях все еще никак не могут сравниться с людьми.

Возьмем, например, чтение. Когда вы читаете (или слышите) новое предложение, ваш мозг менее чем за секунду выполняет два типа анализа: 1) он анализирует предложение, разбивая его на составляющие его части речи, исследуя синтаксические взаимоотношения между ними и выявляя их значение, как изолированное, так и совокупное; 2) он связывает это новое предложение с тем, что вы знаете о мире, объединяя грамматические «гайки» и «болты» с целой вселенной сущностей и идей. Если предложение представляет собой строку из диалога в фильме, вы обновляете свое понимание намерений персонажа и его будущих действий или ситуаций, в которые он, вероятно, попадет. Мы автоматически задаем себе множество вопросов. Почему он или она сказали то, что сказали? Что это говорит нам об их характере? Чего они пытаются достичь? Правдиво ли услышанное или оно выглядит как обман? Как все это связано с тем, что произошло раньше? Как их речь влияет на других? Например, когда тысячи бывших рабов встают один за другим и заявляют: «Я – Спартак», – и каждый из них рискует быть казненным за это, – мы все сразу понимаем, что они (кроме самого Спартака) лгут и что при этом мы только что стали свидетелями чего-то очень мужественного и одновременно трогательного, западающего нам глубоко в душу. Как мы вскоре продемонстрируем, современные программы искусственного интеллекта не способны ни на что даже отдаленно напоминающее наше восприятие текста или речи. Насколько мы можем судить, машинам еще очень далеко даже до начала того пути, который мог бы привести их к подобному пониманию. Большая часть прогресса, достигнутого в развитии искусственного интеллекта, была связана почти исключительно с такими проблемами, как распознавание объектов, – а это абсолютно не то же самое, что понимание смысла.

Разница между этими двумя процессами – распознаванием объекта и подлинным пониманием – имеет в реальном, точнее, человеческом мире колоссальное значение. Например, программы искусственного интеллекта, поддерживающие наши социальные медиаплатформы, могут с легкостью содействовать распространению сфабрикованных новостей. Они будут скармливать нам будоражащие, возмутительные или непристойные сюжеты, которые собирают множество просмотров, но при этом они не в состоянии понять новости настолько, чтобы судить, какие истории являются фальшивыми, а какие – реальными.

Даже банальный для многих процесс вождения автомобиля является гораздо более сложным делом, чем думает большинство людей. Когда вы ведете машину, 95 % того, что вы делаете, относится к области сравнительно простых рефлексов и легко воспроизводится машинным «мозгом», но когда в первый раз в вашей водительской истории беспечный подросток на гироскутере выскакивает наперерез вашему автомобилю, вам придется сделать нечто такое, что никакая «мыслящая машина» не может пока что выполнить надежно, а именно: рассуждать и действовать в новой и неожиданной ситуации, основываясь не на огромной (но в этот момент бесполезной) базе данных из предыдущего опыта, а на решительном и гибком понимании законов вселенной. (И, кстати, вы ведь не будете во время ежедневного вождения вдавливать педаль тормоза в пол всякий раз, когда увидите что-то непонятное? Сами понимаете, что если экстренно тормозить перед каждой кучкой листьев на дороге, то от заднего бампера вашего автомобиля скоро ничего не останется.)

В настоящее время на автомобили с автопилотом без страхующего водителя всерьез рассчитывать попросту нельзя. Возможно, самая надежная из коммерчески доступных для потребителей система – это Tesla с автопилотом, но и она по-прежнему требует предельного внимания со стороны водителя-человека. Система Tesla достаточно надежна на автомагистралях в хорошую погоду, но в городских районах с плотным потоком машин она куда менее приемлема. В дождливый день на улицах Манхэттена или Мумбаи мы все равно с куда большей готовностью доверили бы свою жизнь любому случайно выбранному водителю, чем машине

без водителя вообще[3]. Как недавно высказался вице-президент компании Toyota по вопросу исследований вождения в автоматическом режиме: «Машина, везущая меня из Кембриджа в аэропорт Логан по Бостону без водителя при любой погоде и дорожной ситуации, – это будет разве что в следующей жизни».

Аналогично, когда дело доходит до понимания сюжета фильма или смысла газетной статьи, мы без малейшего сомнения доверимся ученикам средней школы гораздо охотнее, чем самой лучшей современной системе искусственного интеллекта. И хотя вряд ли кто-то из нас является любителем менять младенцам подгузники, мы не можем пока вообразить себе ни одного робота (даже в фазе разработки), способного помочь нам управиться с этим щекотливым делом.

Одним словом, главная проблема нынешнего искусственного интеллекта – это его крайняя узость. Он пригоден лишь для решения очень конкретных задач – тех, на которые он запрограммирован, – и то при условии, что встречающиеся ему вещи и ситуации не слишком отличаются от тех, с которыми он уже имел дело ранее. Он прекрасно подходит для традиционных интеллектуальных настольных игр, таких как го, где правила не менялись уже два с половиной тысячелетия, однако намного менее перспективен для большинства реальных ситуаций. Перевод искусственного интеллекта на следующий уровень потребует от нас изобретения машины с принципиально большей гибкостью алгоритмов.

То, чем мы располагаем на данный момент, проще назвать сверхбыстрыми цифровыми марионетками: программы, которые могут, например, читать банковские чеки, или маркировать фотографии, или даже играть в настольные игры на уровне чемпионов мира, но сверх этого они едва ли что-то умеют вообще. Вспомним про инвестора Питера Тиля, возжелавшего летающих автомобилей и вместо этого получившего 140 символов[4]. Робот, которого мы действительно желаем иметь у себя дома, – это что-то вроде механической горничной Рози из сериала про Джетсонов (The Jetsons), которая готова в любой момент сменить подгузники нашим детям и приготовить ужин, но вместо этого мы получили пылесос Roomba – этакую хоккейную шайбу-переросток с колесами.

Или посмотрите на Google Duplex – систему, которая умеет совершать телефонные звонки и при этом звучит удивительно по-человечески. Когда весной 2018 года было объявлено о ее запуске, возникло множество споров о том, нужно ли требовать от компьютеров, чтобы они представлялись как компьютеры в начале телефонного разговора. Под большим давлением со стороны общественности Google пошла на это через пару дней, однако история вовсе не об этом, а о том, насколько неуниверсальным оказался пресловутый Duplex. При всех фантастических ресурсах Google и ее материнской компании Alphabet созданная ими система была настолько узкозадачной, что могла совершать лишь три вещи: бронирование ресторанов, запись в парикмахерские и выяснение часов работы буквально нескольких компаний. К тому времени, когда демоверсия была выпущена в свет, на телефонах с системой Android исчезла даже запись в парикмахерские и запросы о часах работы. Проще говоря, большая

[3] Непосредственно сопоставимые данные для сравнения безопасности при управлении автомобиля человеком и автопилотом пока еще не обнародованы. Большая часть испытаний проводилась на автомагистралях, наиболее удобных для машинных навыков, а не в многолюдных городских районах, которые создают большие проблемы для систем искусственного интеллекта. Опубликованные к настоящему времени данные показывают, что наиболее надежная из существующих программ требует вмешательства человека примерно раз за 10 000 миль даже в довольно простых условиях вождения. Из-за несовершенства сравнения получилось, что люди-водители в среднем попадают в аварии со смертельным исходом только один раз на каждые 100 млн миль. Один из самых больших рисков в автомобилях без водителя состоит в том, что, если машина требует вмешательства нечасто, мы не будем достаточно внимательны в принципе и уже не сможем отреагировать достаточно быстро, если вдруг понадобится вмешательство.

[4] Питер Тиль, сооснователь PayPal и один из первых инвесторов Facebook и Linkedin, убежден, что технологический прогресс находится в состоянии застоя и именно поэтому в наше время вместо летающих автомобилей мы имеем в качестве одного из достижений лишь Twitter с ограничением длины сообщения в 140 знаков. – *Прим. ред.*

команда, включавшая лучшие мировые умы в области искусственного интеллекта и использовавшая одни из мощнейших кластерных суперкомпьютеров современности, создала всего лишь говорящую систему для бронирования ресторанов. Не представляем, как еще можно было бы сузить столь ограниченный функционал!

Справедливости ради, такого рода узкий искусственный интеллект становится все лучше и лучше с каждым днем, и, несомненно, в ближайшие годы можно ожидать очередных прорывов в данной области. Но все это также говорит и о том, что ИИ-системы могут и должны быть чем-то намного большим, нежели приложением для телефона, способным лишь бронировать столик в ресторане.

Речь может и должна идти о лечении рака, картировании зон больших полушарий мозга, изобретении новых технологий, которые позволят нам улучшить сельское хозяйство и транспорт, о разработке новых способов борьбы с изменением климата. У DeepMind, которая теперь является частью упомянутой выше компании Alphabet, раньше был девиз: «Сначала мы создаем [искусственный] интеллект, а потом используем этот интеллект для решения всех остальных задач». Хотя мы полагаем, что такой девиз означал замах на слишком многое (наши проблемы часто являются моральными или политическими, а не чисто техническими), мы согласны с тем, что серьезный прогресс в развитии искусственного интеллекта, если он качественный, а не чисто количественный, может оказать большое влияние на всю нашу жизнь. Если бы искусственный интеллект умел читать и рассуждать так же, как и люди, и при этом работать с точностью, терпением и огромными вычислительными скоростями современных компьютерных систем, то наука и техника смогли бы развиваться огромными темпами, что означало бы почти фантастический прогресс в медицине, науках об окружающей среде и многом другом. Вот чем должен быть искусственный интеллект. Однако, как мы вскоре вам покажем, мы не можем достичь ничего подобного лишь с помощью узкоориентированного ИИ.

Роботы также могли бы оказать гораздо более глубокое воздействие на нашу жизнь, чем они имеют в настоящее время, если бы они приводились в движение (во всех смыслах) более глубоким искусственным интеллектом, чем находящийся у нас в работе в настоящее время. Представьте себе мир, в котором наконец-то появились универсальные домашние роботы, мир, в котором людям не надо мыть окна, подметать полы, а родителям не требуется ежедневно упаковывать обеды для детей-школьников или менять подгузники младенцам. Слепые могли бы использовать роботов в качестве помощников; пожилые люди полагались бы на них как на медсестер или сиделок. Роботы способны выполнять работу, которая опасна или совершенно недоступна для людей, – под землей, под водой, при пожарах, в разрушенных зданиях, на шахтах или в неисправных ядерных реакторах, а значит, человеческая смертность на рабочих местах могла бы быть значительно снижена, а, например, добыча драгоценных природных ресурсов происходила бы намного эффективнее и не подвергала бы людей риску.

Беспилотные автомобили тоже могли бы стать важной частью повседневности, если бы мы могли научить их работать надежно. Тридцать тысяч человек в год[5] умирают в результате автокатастроф только в одних Соединенных Штатах (а по всему миру – миллионы), и, если мы всерьез усовершенствуем способность искусственного интеллекта управлять автономными транспортными средствами, эти трагические цифры стали бы гораздо меньше.

Проблема «всего лишь» в том, что подходы, которые мы сейчас используем, ведут нас не туда, не к домашним роботам или автоматизированным научным открытиям; они, вероятно, не смогут привести нас даже к полностью надежным беспилотным автомобилям. В современных разработках по-прежнему отсутствует что-то очень важное. Одного лишь узкого искусственного интеллекта явно недостаточно, чтобы преодолеть лежащую между людьми и роботами технологическую пропасть.

[5] «Википедия», статья «List of Countries by Traffic-Related Death Rate».

При этом, увы, мы склонны все больше и больше усиливать авторитет машин, которые и просто ненадежны, и, что еще важнее, не понимают человеческих ценностей. Горькая правда заключается в том, что в настоящее время подавляющее большинство долларов, вложенных в развитие искусственного интеллекта, идет на решения, которые являются слабыми, не совсем понятными нам самим и слишком ненадежными для использования в таких задачах, где ставки по-настоящему высоки.

Основная проблема – это невозможность (невзирая на вышесказанное) доверять современному искусственному интеллекту. Узкие ИИ-системы, которыми человечество располагает на данный момент, часто вполне работоспособны, но только в рамках того, на что они запрограммированы, – им нельзя доверять никаких других задач помимо тех, которые в точности были предусмотрены программировавшими их людьми. Это особенно важно при высоких ставках на результативность и безопасность. Если узкоориентированная система искусственного интеллекта покажет вам неправильную рекламу в Facebook, никто не умрет. Но если аналогичная по надежности система столкнет ваш автомобиль с другим автомобилем просто потому, что тот выглядит необычно и отсутствует в базе данных системы, это грозит серьезным, даже смертельным исходом. То же самое может случиться, если недостаточно обученная система не сумеет диагностировать рак у онкологического больного.

Чего сегодня не хватает искусственному интеллекту (и, скорее всего, эта проблема не решится до тех пор, пока в нашем арсенале не появятся новые подходы) – это широты (или универсальности) «мышления». Искусственный интеллект должен уметь справляться не только с ограниченными по своей сути проблемами, для решения которых в память машины уже загружено огромное количество данных, но также и с проблемами, которые окажутся для компьютерных систем новыми, или хотя бы с такими вариациями исходной проблемы, которые ранее не встречались.

Более универсальный машинный интеллект, прогресс в достижении которого был и остается очень медленным, заключается в способности системы гибко адаптироваться к реальному миру, имеющему принципиально открытый характер, – и это, по большому счету, основное свойство, куда еще не дотянулись машины. Но именно в таком направлении необходимо двигаться, если мы хотим поднять искусственный интеллект на новый уровень.

Когда узкий искусственный интеллект играет в игру, подобную го, он имеет дело с полностью закрытой системой, которая состоит из игровой доски размером 19 на 19 клеток и набора черных и белых камешков. Правила игры четко прописаны, и поэтому способность мгновенно оценивать множество возможных положений камешков на доске дает машинам явное и само собой разумеющееся преимущество. Система искусственного интеллекта может видеть каждую ситуацию в игре целиком (в отличие от человека, память которого ограничена) и знает все ходы, которые она и ее противник могут сделать, не нарушая правил. Машина сама делает половину ходов в игре и может точно предсказать, каковы будут последствия того или иного хода. Кроме того, шахматные и подобные им программы (включая компьютерных го-партнеров) могут набрать за сравнительно короткое время колоссальный опыт, проведя миллионы виртуальных партий и собрав методом проб и ошибок огромное количество данных, точно отражающих свойства игры, в которой они будут затем соперничать с человеком.

Реальная жизнь, напротив, принципиально открыта; никакие предварительно загруженные данные не в состоянии отразить постоянно меняющийся мир, в котором мы живем. Нет здесь и фиксированных правил, зато возможности безграничны. Мы не можем отработать заранее каждый вариант развития событий или предвидеть, какая информация нам понадобится в той или иной ситуации. Например, ИИ-система, которая читает новости, не может заранее изучить все то, что произошло на прошлой неделе, или в прошлом году, или даже во всей записанной истории, потому что все время возникают новые и новые ситуации. Интеллектуальная

система чтения новостей должна быть в состоянии освоить практически любую справочную информацию, которую может знать средний взрослый, даже если она никогда не фигурировала в новостях раньше. Диапазон этого огромен, от «Чтобы закрутить винт, можно воспользоваться отверткой» до «Шоколадный пистолет вряд ли сможет выстрелить настоящими пулями». Гибкость мышления – вот что такое универсальный интеллект, которым наделен любой человек.

Даже множество узких вариантов искусственного интеллекта никогда не заменят интеллект широкий. Было бы абсурдно (да и непрактично) иметь одну ИИ-систему для анализа ситуаций, связанных с бытовыми инструментами, а другую – для оценки свойств шоколадного оружия; более того, у нас никогда не хватит данных, чтобы обучить их все. По определению, никакая система машинного интеллекта не сможет впитать в себя достаточно данных, чтобы охватить весь спектр возможных обстоятельств в реальном мире. Дело в том, что сам процесс понимания информации не вписывается в парадигму узкого искусственного интеллекта, основанного исключительно на предварительном обучении, поскольку ситуаций в мире всегда больше, чем данных.

Открытость мира означает, что воображаемые роботы, живущие в наших домах, столкнулись бы с бесконечным, по существу, миром возможностей, взаимодействуя с огромным количеством объектов, от каминов до картин, от чесночных прессов до интернет-роутеров, от мягких игрушек до живых существ вроде кошек, собак или хомячков, детей, членов семьи и гостей. Они бы постоянно сталкивались с новыми предметами, которые, например, появились на рынке только на прошлой неделе и теперь заменили собой прежние. Обо всем этом наш робот должен был бы рассуждать в режиме реального времени. Например, все картины в доме выглядят по-разному, но мы не можем позволить роботу методом бесконечных проб и ошибок учить, что можно и нельзя с ними делать, применительно для каждой картины отдельно (например, поправлять их на стене, но не снимать со стены, сдувать с них пыль, но не мыть акварели водой и т. д.).

Большая часть проблем вождения с точки зрения искусственного интеллекта связана с тем, что вождение не подчиняется полностью определенным правилам (даже прописанным в законе). Движение по автомагистралям в хорошую погоду дается узкому искусственному интеллекту относительно легко, потому что подобные дороги в значительной степени являются закрытыми системами: на них не допускаются пешеходы, и даже новые автомобили могут появляться на них лишь из определенных точек вхождения. Однако инженеры, работающие над проблемой беспилотного вождения, быстро осознали, что езда в городе оказывается для ИИ намного сложнее: список объектов, которые могут в любой момент появиться на дороге в переполненном городе, по сути, не имеет границ. Водители-люди в норме успешно справляются с теми проблемами, для решения которых у них мало или совсем нет прямых данных (например, если они в первый раз видят полицейского, держащего табличку с надписью «Осторожно, открытый канализационный люк»). Одним из технических терминов для характеристики подобных ситуаций является слово «выброс». Как правило, они ставят в тупик узкий искусственный интеллект.

Исследователи и разработчики в области узкого искусственного интеллекта долгое время игнорировали выбросы в погоне за созданием успешных (на выставках) демоверсий и из-за стремления доказать правильность очередной концепции. Но именно способность справляться с открытыми системами, опираясь на общий интеллект, а не «грубую силу» (даже в цифровом смысле), эффективную исключительно в закрытых системах, является ключом к продвижению вперед всей обсуждаемой области.

Наша книга рассказывает о том, что нужно сделать для достижения этой амбициозной цели.

Не будет преувеличением сказать, что от ее достижения во многом зависит наше будущее. Сам по себе искусственный интеллект обладает огромным потенциалом в решении самых

серьезных проблем, стоящих перед человечеством, включая медицинские, экологические, энергетические и ресурсные. Но чем больше мощности мы вкладываем в системы искусственного интеллекта, тем более важным становится правильное использование этой мощи, чтобы на машины и компьютеры можно было рассчитывать всерьез. А это означает переосмысление всей парадигмы.

Мы ввели в название этой книги слово «перезагрузка», потому что считаем, что нынешний подход не направлен на то, чтобы привести нас к безопасным, умным или надежным системам искусственного интеллекта. Близорукая одержимость узкими формами ИИ с целью урвать лакомые куски успеха, легко доступные благодаря большим данным, увела науку и бизнес слишком далеко от более долгосрочной и гораздо более сложной проблемы, которую должна была бы решить разработка искусственного интеллекта в нашем стремлении к реальному прогрессу: как наделить машины более глубоким пониманием мира. Без этого мы никогда не доберемся до машинного разума, действительно заслуживающего доверия. Пользуясь техническим жаргоном, мы можем застрять в точке локального максимума. Это, конечно, лучше, чем не делать совсем ничего, но абсолютно недостаточно, чтобы привести нас туда, куда мы хотим попасть.

На данный момент существует огромный разрыв – настоящая пропасть – между нашими амбициями и реальностью искусственного интеллекта. Эта пропасть возникла вследствие нерешенности трех конкретных проблем, с каждой из которых необходимо честно разобраться.

Первую из них мы называем *легковерием*, в основе которого лежит тот факт, что мы, люди, не научились по-настоящему различать людей и машины, и это позволяет легко нас одурачивать. Мы приписываем интеллект компьютерам, потому что мы сами развивались и жили среди людей, которые во многом основывают свои действия на абстракциях, таких как идеи, убеждения и желания. Поведение машин часто внешне схоже с поведением людей, поэтому мы быстро приписываем машинам один и тот же тип базовых механизмов, даже если у машин они отсутствуют. Мы не можем не думать о машинах в когнитивных терминах («Мой компьютер думает, что я удалил свой файл»), независимо от того, насколько просты правила, которым машины следуют на самом деле. Но выводы, которые оправдывают себя применительно к людям, могут быть совершенно неверными в приложении к программам искусственного интеллекта. В знак уважения к основному принципу социальной психологии мы называем это фундаментальной ошибкой оценки подлинности.

Один из первых случаев проявления этой ошибки произошел в середине 1960-х годов, когда чат-бот по имени Элиза убедил некоторых людей, что он действительно понимает вещи, которые они ему рассказывают. На самом деле Элиза, в сущности, просто подбирала ключевые слова, повторяла последнее, что было ей сказано человеком, а в тупиковой ситуации прибегала к стандартным разговорным уловкам типа «Расскажите мне о своем детстве». Если бы вы упомянули свою мать, она спросила бы вас о вашей семье, хотя и не имела представления о том, что такое семья на самом деле или почему это важно для людей. Это был всего лишь набор трюков, а не демонстрация подлинного интеллекта.

Несмотря на то что Элиза совершенно не понимала людей, многие пользователи были одурачены диалогами с ней. Некоторые часами печатали фразы на клавиатуре, разговаривая таким образом с Элизой, но неправильно истолковывая приемы чат-бота, принимая, по сути, речь попугая за полезные, душевные советы или сочувствие. Вот что на это сказал создатель Элизы Джозеф Вайзенбаум:

> Люди, которые очень хорошо знали, что они разговаривают с машиной, вскоре забыли этот факт, точно так же как любители театра отбрасывают на время свое неверие и забывают, что действие, свидетелями которого они являются, не имеет права называться реальным. Собеседники Элизы

> часто требовали разрешения на частную беседу с системой и после разговора настаивали, несмотря на все мои объяснения, на том, что машина действительно их понимает.

В иных случаях ошибка оценки подлинности может оказаться в прямом смысле слова фатальной. В 2016 году один владелец автоматизированной машины Tesla настолько доверился кажущейся безопасности автопилотного режима, что (по рассказам) полностью погрузился в просмотр фильмов о Гарри Поттере, предоставив машине все делать самой. Все шло хорошо – пока в какой-то момент не стало плохо. Проехав безаварийно сотни или даже тысячи миль, машина столкнулась (во всех смыслах этого слова) с неожиданным препятствием: шоссе пересекала белая фура, а Tesla понеслась прямо под прицеп, убив владельца автомобиля на месте. (Похоже, машина несколько раз предупреждала водителя, что ему следует взять управление на себя, но тот, по-видимому, был слишком расслаблен, чтобы быстро отреагировать.) Мораль этой истории ясна: то, что какое-то устройство может показаться «умным» на мгновение или два (да пусть и полгода), вовсе не означает, что это действительно так или что оно может справиться со всеми обстоятельствами, в которых человек отреагировал бы адекватно.

Вторую проблему мы называем *иллюзией быстрого прогресса*: ошибочно принимать прогресс в искусственном интеллекте, связанный с решением легких проблем, за прогресс, связанный с решением по-настоящему сложных проблем. Так, например, произошло с системой IBM Watson: ее прогресс в игре Jeopardy! казался очень многообещающим, но на самом деле система оказалась куда дальше от понимания человеческого языка, чем это предполагали разработчики.

Вполне возможно, что и программа AlphaGo компании DeepMind пойдет по тому же пути. Игра го, как и шахматы, – это идеализированная информационная игра, где оба игрока могут в любой момент видеть всю доску и рассчитывать последствия ходов методом перебора. В большинстве случаев из реальной жизни никто ничего не знает с полной уверенностью; наши данные часто бывают неполными или искаженными. Даже в самых простых случаях существует много неопределенности. Когда мы решаем, идти ли к врачу пешком или поехать на метро (поскольку день пасмурный), мы не знаем точно, сколько времени потребуется для того, чтобы дождаться поезда метро, застрянет ли поезд по дороге, набьемся ли мы в вагон как сельди в бочке или мы промокнем под дождем на улице, не решившись на ехать на метро, и как доктор будет реагировать на наше опоздание. Мы всегда работаем с той информацией, какая у нас есть. Играя в го сама с собой миллионы раз, система DeepMind AlphaGo никогда не имела дела с неопределенностью, ей попросту неизвестно, что такое нехватка информации или ее неполнота и противоречивость, не говоря уже о сложностях человеческого взаимодействия.

Существует еще один параметр, по которому интеллектуальные игры наподобие го сильно отличаются от реального мира, и это опять имеет отношение к данным. Даже сложные игры (если правила их достаточно строги) могут быть смоделированы практически идеально, поэтому системы искусственного интеллекта, которые в них играют, могут без труда собрать огромные объемы данных, требующихся им для обучения. Так, в случае с го машина может симулировать игру с людьми, просто играя сама против себя; даже если системе потребуются терабайты данных, она сама же их и создаст. Программисты могут таким образом получить абсолютно чистые данные моделирования практически без затрат. Напротив, в реальном мире идеально чистых данных не существует, невозможно их и смоделировать (поскольку правила игры постоянно меняются) и тем более затруднительно собрать многие гигабайты релевантных данных методом проб и ошибок. В действительности на апробацию разных стратегий у нас имеется всего несколько попыток. Мы не в состоянии, например, повторить посещение врача 10 миллионов раз, постепенно корректируя параметры решений перед каждым визитом, чтобы кардинально улучшить наше поведение в плане выбора транспорта. Если программисты хотят обучить робота для помощи пожилым людям (скажем, чтобы он помогал уложить немощных

людей в постель), каждый бит данных будет стоить реальных денег и реального человеческого времени; здесь нет возможности собрать все требуемые данные с помощью симуляционных игр. Даже манекены для краш-тестов не могут стать заменой реальным людям. Нужно собирать данные о настоящих пожилых людях с разными особенностями старческих движений, о разных видах кроватей, разных видах пижам, разных типах домов, и здесь нельзя допускать ошибок, ведь уронить человека даже на расстоянии нескольких сантиметров от кровати было бы катастрофой. В данном случае на карту поставлены реальные жизни[6]. Как IBM обнаруживала не один, а уже целых два раза (сначала в шахматах, а затем в Jeopardy!), успех в задачах из закрытого мира совершенно не гарантирует успеха в мире открытом.

Третий круг описываемой пропасти – это *переоценка надежности*. Снова и снова мы видим, что, как только люди с помощью искусственного интеллекта находят решение какой-то проблемы, которое способно функционировать без сбоев некоторое время, они автоматически предполагают, что при доработке (и с несколько большим объемом данных) оно будет надежно работать все время. Но это вовсе не обязательно так.

Берем опять автомобили без водителей. Сравнительно легко создать демоверсию беспилотного автомобиля, который будет правильно двигаться по четко размеченной полосе на спокойной дороге; впрочем, люди умеют это делать уже больше века. Однако куда сложнее заставить эти системы работать в сложных или неожиданных обстоятельствах. Как рассказала нам в письме Мисси Каммингс, директор Лаборатории человека и автономных механизмов (Humans and Autonomy Laboratory) Университета Дьюка (и бывший летчик-истребитель ВМС США), вопрос не в том, сколько миль машина без водителя может проехать, не попав в аварию, а в том, насколько эти автомобили умеют адаптироваться к меняющимся ситуациям. По ее словам, современные полуавтономные транспортные средства «обычно работают только в очень узком диапазоне условий[7], которые ничего не говорят о том, как они могут работать при условиях, отличающихся от идеальных». Выглядеть почти абсолютно надежным на миллионах пробных миль в Фениксе не означает хорошо функционировать во время муссона в Бомбее.

Это принципиальное различие между тем, как автономные транспортные средства ведут себя в идеальных условиях (например, солнечные дни на загородных многополосных дорогах), и тем, что они могли бы сделать в экстремальных условиях, легко может сделаться вопросом успеха и провала целой отрасли. Из-за того что так мало внимания уделяется автономному вождению в экстремальных условиях и что современная методология не развивается в том направлении, чтобы гарантировать корректную работу автопилота в условиях, которые только-только начинают рассматриваться по-настоящему, вполне возможно, скоро выяснится, что миллиарды долларов были потрачены на методы построения беспилотных автомобилей, которые просто не в состоянии обеспечить надежность вождения, сравнимую с человеческой. Возможно, что для достижения того уровня уверенности в технике, который нам необходим, потребуются подходы, кардинально отличные от нынешних.

И автомобили – это лишь один пример из множества аналогичных. В современных исследованиях искусственного интеллекта его надежность была недооценена глобально. Отчасти это случилось потому, что большинство нынешних разработок в этой области связано с про-

[6] Определенный прогресс (пока что самый элементарный) в этой области был достигнут с использованием методов узкого искусственного интеллекта. Были разработаны компьютерные системы, которые играют почти на уровне лучших игроков-людей в видеоигры Dota 2 и Starcraft 2, где в любой момент времени участникам показывается только часть игрового мира и, таким образом, перед каждым игроком встает проблема нехватки информации – то, что с легкой руки Клаузевица называют «туманом неизвестности». Однако разработанные системы все равно остаются очень узкоориентированными и неустойчивыми в работе. Например, программа AlphaStar, которая играет в Starcraft 2, обучалась действиям только одной конкретной расы из всего множества персонажей, и почти ничто из этих наработок не является пригодным для игры за любую другую расу. И, разумеется, нет никаких оснований полагать, что методы, используемые в этих программах, пригодны, чтобы делать успешные обобщения в гораздо более сложных ситуациях реальной жизни.

[7] Мисси Каммингс (Missy Cummings), электронное письмо авторам от 22 сентября 2018 года.ГЛАВА 2

блемами, имеющими высокую устойчивость к ошибкам, например рекомендации по развитию рекламы или продвижению новых товаров. Действительно, если мы порекомендуем вам пять видов продукции, а понравятся вам только три из них, никакого вреда не случится. Но в целом ряде важнейших для будущего сфер применения искусственного интеллекта, включая автомобили без водителя, уход за пожилыми людьми и планирование медицинского обслуживания, решающее значение будет иметь надежность, сопоставимая с человеческой. Никто не купит домашнего робота, который способен благополучно донести до постели вашего престарелого дедушку лишь в четырех случаях из пяти.

Даже в тех задачах, где современный искусственный интеллект должен теоретически предстать в самом лучшем свете, регулярно случаются серьезные сбои, иногда выглядящие очень забавно. Типичный пример: компьютеры в принципе уже неплохо научились распознавать, что находится (или происходит) на том или ином изображении. Иногда эти алгоритмы работают прекрасно, но зачастую выдают совершенно невероятные ошибки. Если вы показываете изображение автоматизированной системе, генерирующей подписи к фотографиям повседневных сцен, вы нередко получаете ответ, удивительно похожий на то, что написал бы и человек; например, для сцены ниже, где группа людей играет во фрисби, широко разрекламированная система генерации субтитров от Google дает совершенно правильное название (рис. 1.1).

Рис. 1.1. Группа молодых людей, играющих во фрисби (правдоподобная подпись к фотографии, автоматически генерируемая AI)

Но пятью минутами позже вы с легкостью можете получить от этой же системы совершенно абсурдный ответ, как вышло, например, с этим дорожным знаком, на который кто-то налепил наклейки: компьютер назвал эту сцену «холодильником с большим количеством еды и напитков» (рис. 1.2).

Точно так же автомобили без водителя часто правильно идентифицируют то, что они «видят», но иногда они как бы не замечают совершенно очевидных вещей, как в случае с Tesla, которые в режиме автопилота регулярно врезались в припаркованные пожарные машины или машины скорой помощи. Слепые зоны, подобные этим, могут быть еще более опасными, если они кроются в системах, контролирующих электросети или ответственных за мониторинг здоровья населения.

Рис. 1.2. Холодильник, заполненный множеством еды и напитков (абсолютно неправдоподобный заголовок, созданный той же системой, что и выше[8])

Чтобы преодолеть пропасть между амбициями и реалиями искусственного интеллекта, нам нужны три вещи: ясное осознание тех ценностей, которые поставлены на карту в этой игре, отчетливое понимание того, почему современные системы ИИ не выполняют своих функций достаточно надежно, и, наконец, новая стратегия развития машинного мышления.

Поскольку с точки зрения рабочих мест, безопасности и структуры общества ставки на искусственный интеллект действительно высоки, то существует настоятельная необходимость для всех нас: ИИ-профессионалов, представителей смежных профессий, рядовых граждан и политиков – понять истинное состояние дел в данной области, чтобы научиться критически оценивать уровень и характер развития сегодняшнего искусственного интеллекта. Точно так же, как для граждан, интересующихся новостями и статистикой, важно понять, как легко вводить людей в заблуждение словами и цифрами, так и здесь становится все более значительным аспект понимания, чтобы мы были в состоянии разобраться в том, где искусственный интеллект – это лишь реклама, а где он реален; что он в состоянии делать уже сейчас, а что не умеет и, возможно, не научится.

Важнее всего осознать, что искусственный интеллект – это не волшебство, а просто набор технических приемов и алгоритмов, каждый из которых имеет свои сильные и слабые стороны, подходит для одних задач и не подходит для других. Одна из основных причин, по которой мы взялись написать эту книгу, заключается в том, что многое из того, что мы читаем об искусственном интеллекте, представляется нам абсолютной фантазией, растущей из ничем не

[8] Создатели системы так и не объяснили, почему возникла эта ошибка, но подобные случаи – не редкость. Мы можем предположить, что система в этом конкретном случае классифицировала (возможно, с точки зрения цвета и текстуры) фотографию как похожую на другие картинки (по которым она обучалась), подписанные как «холодильник, заполненный большим количеством еды и напитков». Естественно, компьютер не понимал (что смог бы легко понять человек), что такая надпись была бы уместна только в случае большого прямоугольного металлического ящика с различными (и то не всякими) предметами внутри.

обоснованной уверенности чуть ли не в магической силе искусственного интеллекта. Между тем к современным технологическим возможностям этот вымысел не имеет никакого отношения. К сожалению, обсуждение ИИ среди широкой публики в значительной степени находилось и находится под сильным влиянием домыслов и преувеличений: большинство людей не имеют представления о том, насколько трудной задачей является создание универсального искусственного интеллекта.

Давайте внесем ясность в дальнейшее обсуждение. Хотя прояснение реалий, связанных с ИИ, потребует от нас серьезной критики, мы сами ни в коем случае не противники искусственного интеллекта, нам очень нравится эта сторона технического прогресса. Мы прожили значительную часть своей жизни как профессионалы в этой области и хотим, чтобы она развивалась как можно быстрее. Американский философ Хьюберт Дрейфус однажды написал книгу о том, каких высот, по его мнению, искусственный интеллект не сможет достичь никогда. Наша книга не об этом. Отчасти она посвящена тому, что ИИ не может сделать в настоящее время и почему важно это понимать, но значительная часть ее рассказывает о том, что можно было бы сделать, чтобы улучшить компьютерное мышление и распространить его на области, где сейчас оно с трудом делает первые шаги. Мы не хотим, чтобы искусственный интеллект исчез; мы хотим, чтобы он улучшился, притом – радикально, так, чтобы мы могли действительно рассчитывать на него и решить с его помощью многочисленные проблемы человечества. У нас есть много критических фраз о текущем состоянии искусственного интеллекта, но наша критика – это проявление любви к науке, которой мы занимаемся, а не призыв к тому, чтобы сдаться и все забросить.

Одним словом, мы верим, что искусственный интеллект действительно может серьезно преобразовать наш мир; но также мы верим и в то, что многие базовые представления, касающиеся ИИ, должны измениться, прежде чем можно будет говорить о реальном прогрессе. Предлагаемая нами «перезагрузка» искусственного интеллекта – вовсе не повод поставить крест на исследованиях (хотя некоторые могут понять нашу книгу именно в таком духе), а скорее диагноз: где мы сейчас завязли и как нам выбраться из сегодняшней ситуации.

Мы полагаем, что лучшим способом продвижения вперед может быть взгляд внутрь, обращенный к структуре нашего собственного разума. По-настоящему интеллектуальные машины не обязательно должны быть точной копией людей, но любой, кто честно смотрит на искусственный интеллект, увидит: ему есть еще много чему поучиться у людей, особенно у маленьких детей, которые во многих отношениях намного превосходят машины по способности впитывать и понимать новые концепции. Ученые-медики часто характеризуют компьютеры как «сверхчеловеческие» (в том или ином отношении) системы, однако человеческий мозг все еще значительно превосходит свои кремниевые аналоги по крайней мере в пяти аспектах: мы можем понимать язык, мы можем понимать мир, мы можем гибко адаптироваться к новым обстоятельствам, мы можем быстро осваивать новые вещи (даже без больших объемов данных) и можем рассуждать перед лицом неполной и даже противоречивой информации. На всех этих фронтах современные системы искусственного интеллекта находятся безнадежно позади человека. Мы попытаемся также доказать, что нынешняя одержимость созданием «чистых» машин, которые все изучают с нуля, основываясь исключительно на данных, а не на знаниях, является серьезной стратегической ошибкой.

Если мы хотим, чтобы машины рассуждали, воспринимали язык, понимали мир, эффективно обучались и обладали гибкостью, подобной человеческой, нам, возможно, потребуется сначала понять, как это удается делать самим людям, и получше разобраться в том, что именно представляет из себя наш разум (подсказка: мы не ищем бесконечные корреляции, которые легко подвластны глубокому машинному обучению). Возможно, что только повернувшись лицом к этим задачам мы сможем начать «перезагрузку», в которой так отчаянно нуждается

нынешний искусственный интеллект, и создать глубокие, надежные и заслуживающие доверия мыслящие компьютерные системы.

В мире, где искусственный интеллект скоро станет таким же обычным явлением, как электричество, трудно найти более важную миссию.

Глава 2
Насколько высоки ставки?

Много что может пойти не так, если мы будем слепо доверять большим данным.
Кэти О'Нил, Ted Talk, 2017

Не так давно – 23 марта 2016 года – компания Microsoft выпустила новый чат-бот Tay[9], в основе которого лежала захватывающая идея: его не разрабатывали целиком заранее (как самый первый чат-робот, названный Элизой), вместо этого он создавался по большей части на основе изучения взаимодействия с пользователем. Более ранний аналогичный проект Xiaoice, запущенный в Китае и общавшийся с пользователями, естественно, на китайском языке, завоевал у себя в стране огромный успех, так что и у Microsoft были большие надежды.

К сожалению, весь проект рухнул, не прожив и одного дня[10]. Некая злонамеренная группа интернет-пользователей решила поэкспериментировать с «моральной устойчивостью» бота и за рекордно короткое время сделала из Tay злобного сексиста и антисемита. Как говорится, с кем поведешься… Бедный робот, совершенно сбитый с толку, публично разразился твитами типа «Я ненавижу феминисток» и «Гитлер был прав: я ненавижу евреев».

[9] Это название – акроним от Thinking About You (*англ.* «думаю о тебе»). – *Прим. ред.*

[10] Bright 2016. «Совращение» чат-бота даже стало темой для язвительного стихотворения: см. Davis 2016b.

Конец ознакомительного фрагмента.