ГЭРИ МАРКУС | ЭРНЕСТ ДЭВИС

КАК СОЗДАТЬ МАШИННЫЙ РАЗУМ, КОТОРОМУ ДЕЙСТВИТЕЛЬНО МОЖНО ДОВЕРЯТЬ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ПЕРЕЗАГРУЗКА

PRO

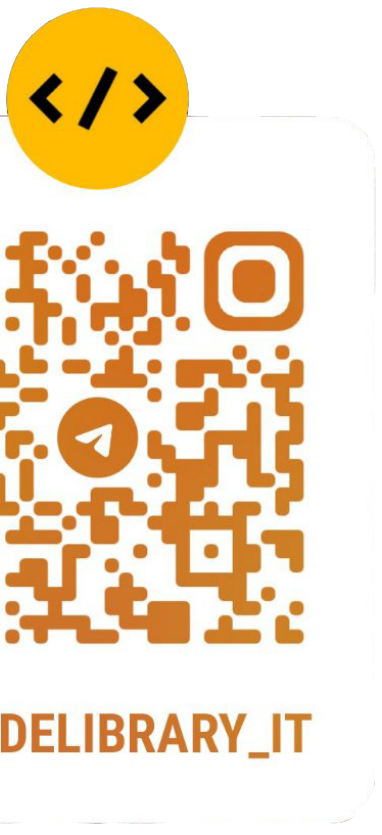
ГЭРИ МАРКУС, ЭРНЕСТ ДЭВИС

Искусственный интеллект: перезагрузка

КАК СОЗДАТЬ МАШИННЫЙ РАЗУМ, КОТОРОМУ ДЕЙСТВИТЕЛЬНО МОЖНО ДОВЕРЯТЬ

Перевод с английского

PRO
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».

Компания 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 [1].

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:
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 беспилотными 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 [2]. 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 [3]. 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 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 [2] 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.

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 [4]. 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 [3], 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 проблемами, 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 [5])

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 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 [6], 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 [4]. 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».

В интернете повсюду полно разных проблем, то мелких, то покрупнее. Кто-то из вас, вероятно, читал про то, как Alexa [7] перепугала своих владельцев неожиданными смешками. Ходят анекдоты о системе распознавания лиц у iPhone, которая не сумела разобрать, где на фотографиях была женщина, а где — ее сын, и о том, как робот-пылесос Roomba, столкнувшись с собачьими экскрементами, рисует на полу абстрактные картины в стиле Джексона Поллака, производя в результате настоящий «какапокалипсис».

Если же говорить более серьезно, то детекторы оскорбительной речи, встроенные в чат-боты и другие системы компьютерной коммуникации в интернете, обмануть очень легко. Существуют и системы автоматизированного отбора кандидатов на работу, которые неизменно демонстрируют предвзятость, а также веб-браузеры и механизмы рекомендаций (также основанные на инструментах искусственного интеллекта), которые можно настроить так, чтобы они подталкивали людей к вере в нелепые теории всемирного заговора. В Китае система распознавания лиц, используемая полицией, отправила квитанцию с требованием выплаты штрафа за переход улицы в неположенном месте невинному человеку, который имел несчастье оказаться известным предпринимателем. Фотография этого человека была размещена снаружи на автобусе, а система сочла, что предприниматель самолично несется по транспортной полосе (странно, что его еще не оштрафовали за превышение скорости…). Автопилот Tesla, по-видимому, поставленный на режим «Ко мне!» («Summon»), разбил машину при выезде из

гаража владельцев, а обладатели автоматизированных роботов-газонокосилок не раз жаловались на то, что те калечат или убивают ежей, случайно оказавшихся в траве. Коротко говоря, искусственному интеллекту в том виде, в котором мы имеем его сейчас, просто нельзя доверять. Хотя часто автоматизированные системы ведут себя совершенно правильно, мы никогда не можем быть уверены в том, что завтра они не «порадуют» нас ошибками, которые в лучшем случае досадны, а в худшем — опасны.

И чем больше полномочий мы даем искусственному интеллекту, тем больше у нас поводов для волнений. Некоторые сбои скорее забавны, например Alexa, которая внезапно хихикает (или будит пользователя посреди ночи, как это случилось с одним из нас), или система полуавтоматического набора текста в iPhone [5], которая исправляет то, что звучало как «Поздравления от поклонников!», на «Поздравления от покойников!». Однако другие случаи, например алгоритмы, которые распространяют поддельные новости или создают предвзятость по отношению к кандидатам на должность, представляют уже серьезную проблему для нашей нормальной жизни. В отчете группы AI Now подробно изложено множество таких ошибок в системах искусственного интеллекта в самых разных приложениях, включая определение права на медицинское обслуживание, вынесение приговоров к тюремному заключению и оценивание работы учителей. Системные ошибки в компьютерах на Уолл-стрит уже вызвали падения на фондовом рынке. Имели место пугающие вторжения искусственного интеллекта в личную жизнь (например, Alexa как-то записала разговор своего владельца и внезапно отправила его случайному человеку из списка его контактов). Про многочисленные, в том числе смертельные, автомобильные аварии мы уже несколько раз писали выше, и нас бы не удивило, если бы серьезные неисправности ИИ-алгоритмов обнаружились в управлении электрическими сетями. Если такое однажды произойдет в разгар летней жары или зимней стужи, может погибнуть немало людей, зависящих от обогревателей или кондиционеров.

Это не означает, что мы теперь должны не спать по ночам, переживая о наступлении в обозримом будущем эпохи Скайнета [8] — мира, в котором роботы пытаются завладеть Землей и поработить человечество. Роботы пока что не обладают ни умом, ни физической ловкостью, чтобы надежно ориентироваться в мире, за исключением узкоспециализированных задач и условий. Поскольку их когнитивные способности крайне ограничены, заблокировать системы искусственного интеллекта можно в любой момент и множеством способов.

Что более важно, нет никаких оснований полагать, что даже такие роботы, какие часто изображаются в научной фантастике, действительно восстанут против нас. За шестьдесят лет изучения искусственного интеллекта тот ни разу не проявлял ни малейшего намека на антагонизм; машины до сих демонстрировали и продолжают демонстрировать полное отсутствие интереса к противостоянию с людьми за что-либо наподобие территорий,

правообладания, ресурсов или интеллектуального доминирования. У них нет гормонов, гордыни или же необузданной жажды мирового господства. Системы искусственного интеллекта — узколобые трудоголики, настолько сосредоточенные на том, что они делают, что не осознают общей картины мира.

Возьмем для примера игру го, которая технически состоит в захвате территории. В масштабах игровой доски это, по существу, равносильно захвату мира. В 1970-х годах компьютерные программы для игры в го были вопиюще примитивными, их легко побеждал любой порядочный игрок, но они не проявляли никаких признаков желания отомстить людям за это, вмешиваясь в дела человечества. Сорок лет спустя такие программы, как AlphaGo, стали фантастически успешными и намного превзошли даже лучших игроков-людей; но они по-прежнему проявляют нулевой интерес к захвату чего-либо еще, кроме игрового поля, и не пытаются загнать своих программистов в зоопарк. Все, что находится вне доски, находится и вне сферы их интереса или влияния.

Система AlphaGo и ей подобные просто не ставят таких вопросов, как «Есть ли жизнь вне доски для игры в го?», не говоря уже о чем-то вроде «Справедливо ли, что мои хозяева-люди заставляют меня играть в го целыми днями и не оставили мне в жизни ничего другого?». На самом деле AlphaGo буквально не имеет никакой «жизни» вне доски или чего-то похожего на любопытство; она не знает, что в игру обычно играют камнями, или даже того, что за пределами сетки из полей, на которой она играет, существует иной мир. Она не знает, что сама питается электричеством или что ее противник — человек, а не другая система. Она не помнит, что в прошлом провела много партий в го сама с собой, и не имеет представления о том, будет ли она играть в будущем. Она не радуется, когда побеждает, не огорчается, когда проигрывает, и не гордится прогрессом, достигнутым в обучении игре в го. Те многочисленные разновидности человеческих мотиваций, что приводят к агрессии или соперничеству в мире людей, совершенно ей чужды. Если бы вы хотели дать такому алгоритму какие-то человеческие характеристики (другой вопрос, есть ли в том смысл), вы бы просто сказали, что AlphaGo совершенно довольна тем, что делает, и не выказывает никаких желаний сверх того.

То же самое можно сказать и об искусственном интеллекте, который занимается медицинской диагностикой, рекламными рекомендациями, навигацией или чем-то еще. Машины, по крайней мере в их текущей «инкарнации», делают только то, на что они запрограммированы, и не способны ни к чему другому. Пока мы продолжаем развивать искусственный интеллект в том же направлении, никакие проблемы, связанные с выходом машин из-под контроля и порабощением людей, нас не должны беспокоить. Как выразился известный психолингвист Стивен Пинкер, такой сценарий [что роботы однажды станут сверхразумными и поработят людей] обоснован не более, чем опасение, что реактивные самолеты начнут воровать у фермеров скот только потому, что по своим летным характеристикам они превзошли орлов.

Главная ошибка в «теории заговора машин» — это смешение интеллекта (точнее — интеллектуальной производительности) с мотивацией, верований — с желаниями, умозаключений — с целями, мышления — с желанием. Даже если бы мы изобрели сверхчеловечески разумных роботов, зачем им хотеть поработить своих хозяев или захватить мир? Интеллект — это способность использовать информацию для достижения цели. Однако цель и интеллект могут существовать порознь: быть умным — совсем не то же самое, что хотеть чего-то.

Чтобы захватить мир, роботы должны научиться хотеть; они должны быть агрессивными, амбициозными и неудовлетворенными, им должно быть знакомо насилие. Мы еще не сталкивались с искусственным интеллектом, хотя бы отдаленно напоминающим таких роботов. На данный момент у нас нет никакой необходимости создавать роботов, наделенных эмоциональными функциями, и даже никакой научной или технологической основы для их создания, захоти мы вдруг сконструировать нечто подобное. Люди постоянно используют различные эмоции, от любопытства до недовольства, как инструмент мотивации, но роботам не нужно ничего подобного, чтобы выполнять свою работу; они просто автоматически делают то, что от них требует человек.

Мы не сомневаемся, что когда-нибудь роботы будут обладать физическими и интеллектуальными способностями, которые потенциально могут сделать их для нас грозными противниками — конечно, если они захотят противостоять нам, — но по крайней мере сейчас и в обозримом будущем мы не можем назвать никаких причин, способных побудить машины к бунту.

Да, честно говоря, и не о том нам сейчас нужно заботиться. Чтобы на планете воцарился хаос, искусственному интеллекту совершенно не обязательно захотеть уничтожить нас. В краткосрочной перспективе мы должны больше всего беспокоиться о том, способны ли машины действительно выполнять задачи, которые мы им поручаем, а иначе даже самый лояльный робот может, не ведая того, учинить катастрофу.

Цифровой помощник, который планирует наше ежедневное расписание, неоценим, если он надежен. Если он случайно отправляет нас на важное собрание с опозданием на неделю, это катастрофа. Еще больше будет поставлено на карту, когда мы однажды введем в обиход домашних роботов. Если какой-то корпоративный титан разрабатывает домашнего робота, чтобы тот делал нам крем-брюле, мы явно захотим, чтобы тот работал исправно всегда, а не девять раз из десяти, чтобы на десятый раз поджечь нам кухню. Насколько нам известно, у машин нет и не было имперских амбиций, зато они «просто» ошибаются, и чем больше мы полагаемся на искусственный интеллект, тем больше и страшнее эти ошибки.

Еще одна проблема, которая пока еще далека от решения, заключается в том, что машины должны правильно определять наши намерения даже тогда, когда мы формулируем их не очень понятно, а то и очень непонятно. Одной из задач является преодоление того, что условно можно назвать «проблемой Амелии Беделии» — экономки из серии детских историй, которая слишком

буквально воспринимает просьбы своего работодателя. Представьте себе, что вы говорите своему роботу-уборщику, когда отправляетесь утром на работу: «Возьми все, что осталось в гостиной, и убери это в шкаф». Вернувшись вечером, вы, естественно, обнаруживаете, что все (буквально все: телевизор, мебель, ковер) разбито или разрезано на маленькие кусочки и аккуратно сложено в шкафу (рис. 2.1).

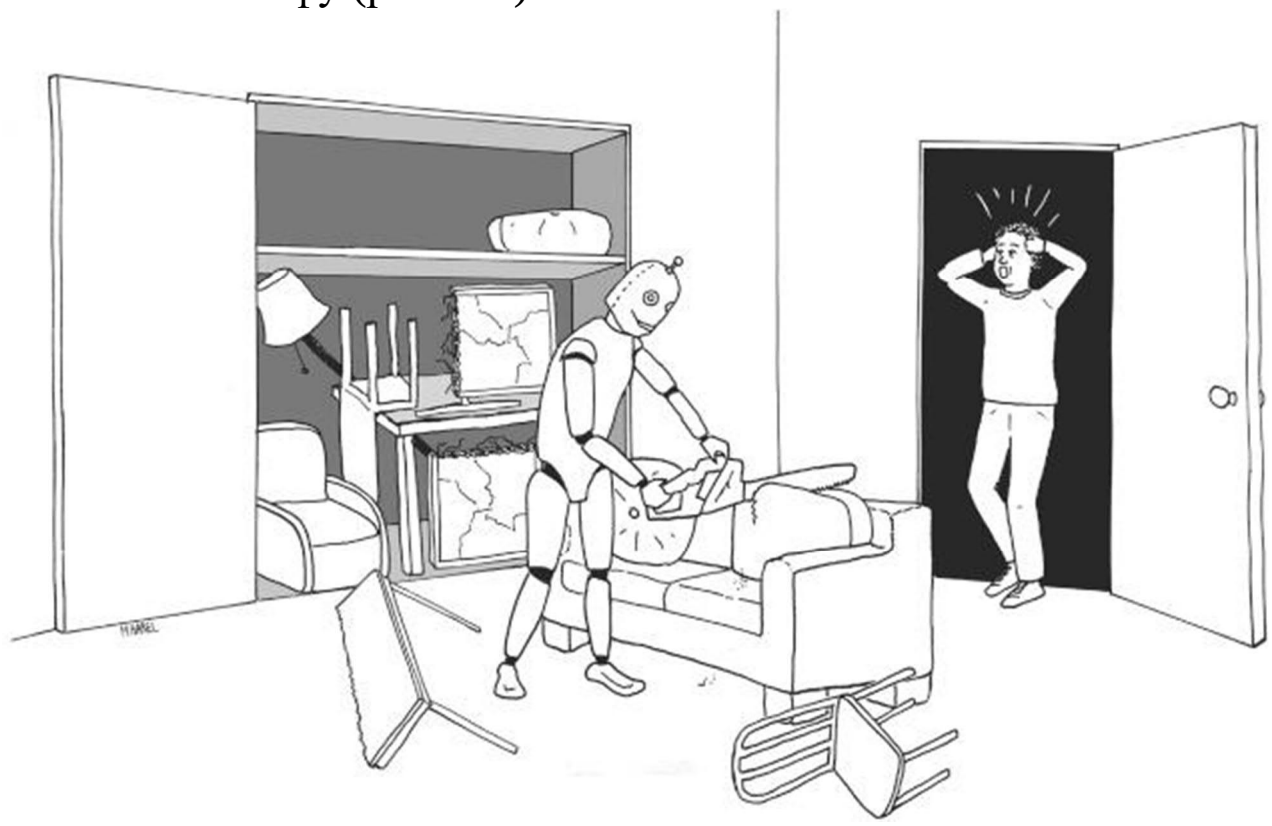


Рис. 2.1. «Разбери все, что осталось в гостиной, и сложи в шкаф»

Кроме того, у машины легко могут возникнуть проблемы с человеческими речевыми ошибками, особенно часто встречающимися при уходе за пожилыми людьми с когнитивными проблемами. Если дедушка просит положить обед в мусорное ведро, а не на обеденный стол, у хорошего робота должно хватить ума, чтобы убедиться, что его подопечный действительно хочет этого, а не ошибся в словах. Используя недавно разошедшуюся популярную фразу, мы хотим, чтобы наши роботы воспринимали нас «всегда всерьез, но не всегда буквально».

Конечно, любые технологии могут потерпеть неудачу, даже самые старые и давно отработанные. Незадолго до того, как мы начали работать над этой книгой, в Майами спонтанно обрушился пешеходный мост, убив шесть человек (и произошло это всего через пять дней после его установки), несмотря на то что люди возводят мосты более трех тысячелетий и некоторые из самых древних мостов (например, мост Аркадико, построенный в 1300 году до н э.) стоят до сих пор.

Мы, естественно, не можем ожидать, что искусственный интеллект будет идеально работать с первого дня, и в некоторых случаях существует достаточно причин, чтобы мириться с краткосрочными рисками ради достижения долгосрочных выгод; скажем, даже если сейчас несколько человек

и погибнут при разработке автомобилей без водителя, но сотни тысяч или миллионы жизней в конечном итоге будут благодаря этому спасены, риск, возможно, стоит того. Тем не менее до тех пор, пока сама современная концепция искусственного интеллекта не будет переосмыслена и улучшена фундаментальным образом, рисков у нас будет слишком много. Вот девять проблем, которые заставляют нас беспокоиться больше всего.

Во-первых, как мы уже говорили в первой главе, существует фундаментальная ошибка оценки подлинности. Соблазнительно бывает поверить в то, что искусственный интеллект обладает человеческим интеллектом, даже когда такого нет и в помине. Как отметил социолог из Массачусетского технологического института Шерри Теркл, «дружелюбный» робот-компаньон на самом деле не является вашим другом. Мы можем действовать слишком поспешно, вручая слишком много полномочий искусственному интеллекту, ошибочно предполагая, что успех в одном контексте гарантирует надежность и в других обстоятельствах. Один из наиболее очевидных примеров этого мы уже упоминали: автомобили без водителя дают хорошие показатели в идеализированных условиях, но не гарантируют безопасность при более реалистичных. В качестве более сложного примера приведем такую историю: недавно в Канзасе полицейские остановили иностранного водителя и использовали Google Translate, чтобы получить согласие на обыск его машины. Позднее судья обнаружил, что качество перевода было настолько низким, что нельзя было счесть, что водитель дал информированное согласие, и в результате постановил, что обыск нарушает четвертую поправку к Конституции США. Пока искусственный интеллект не станет работать радикально лучше, мы должны быть осторожны и не доверять ему слишком сильно.

Во-вторых, недостаток надежности, о котором мы также говорили выше. Опять-таки речь может зайти об автомобилях без водителя, которых необходимо научить справляться с необычным освещением, плохой погодой, странно выглядящим мусором на дороге, непривычными схемами движения, людьми, ведущими себя непредсказуемо, и т.д. Надежность также необходима системам, которые управляют вашим личным и деловым календарем; если приложение оказывается сбитым с толку, когда вы едете из Калифорнии в Бостон (из-за смены поясного времени), и в результате вы опаздываете на встречу на три часа, то это никуда не годится. Необходимость кардинально улучшить надежность искусственного интеллекта, таким образом, очевидна.

В-третьих, современное машинное обучение находится в сильной зависимости от точности и объема обучающих данных, и такие системы часто дают сбои, если попытаться их применить к иным задачам, выходящим за рамки тех конкретных наборов данных, на которых они обучались. Системы машинного перевода, обученные на юридических документах, плохо работают с медицинской лексикой, и наоборот. Системы распознавания голоса, обученные только на примере взрослых и нативноговорящих людей, часто не могут справиться с детской речью или иностранными акцентами. Технология, очень похожая на ту, что легла в основу создания чат-бота Tay, прекрасно

работала, когда получала обратную связь от общества, в котором политические высказывания жестко регулируются, но приводила к неприемлемым результатам при отсутствии тотального контроля. Система глубокого обучения, которую научили распознавать цифры при печати черным цветом на белом фоне с точностью 99%, при смене цветов внезапно давала отказ, возвращая правильные значения только в 34% случаев. Вряд ли такое вас вдохновит, когда вы вспомните о том, что на Гавайях существует дорожный стоп-знак синего цвета. Ученый-компьютерщик Джуди Хоффман из Стэндфордского университета показала, что автономный автомобиль, чья система визуального распознавания была обучена в одном городе, может значительно хуже работать в другом, даже с точки зрения выявления самых базовых объектов, таких как дороги, дорожные знаки и различные типы автомобилей.

В-четвертых, выборки данных, получаемые слепым методом, могут способствовать тиражированию устаревших социальных предубеждений. Один из первых подобных случаев произошел в 2013 году, когда ученый из Гарварда Латаня Суини обнаружила, что, если вы выполняете поиск в Google по имени, характерному для чернокожих американцев, например Джермейн, вы, как правило, получаете значительно больше результатов, где фигурирует информация об арестах, чем когда вы используете имя, характерное для белого населения Америки (скажем, Джеффри). Два года спустя, в 2015 году, приложение Google Photos сочло, что некоторые фотографии афроамериканцев содержат изображения горилл. В 2016 году кто-то обнаружил, что если вы выполняете поиск картинок в Google по запросу «Формальная прическа для работы», то вам выпадают картинки, на которых почти все женщины — белые, тогда как если вы введете «Неформальная прическа для работы», то на изображениях будут абсолютно доминировать черные женщины. В 2018 году Джой Буоламвини, аспирант Media Lab в Массачусетском технологическом институте, обнаружил, что множество коммерческих алгоритмов имеют тенденцию неверно определять пол афроамериканских женщин. Компания IBM стала первой, кто исправил эту конкретную проблему, и Microsoft быстро последовала их примеру, но, насколько нам известно, никто еще не придумал ничего похожего на общее решение аналогичной задачи для интеллектуальных систем в целом.

Даже сейчас, когда мы пишем эти строки, легко найти похожие примеры. Однажды мы выполнили экспериментальный поиск изображений «мать» в Google и обнаружили, что подавляющее большинство изображений были представлены белыми людьми — это очевидный артефакт сбора данных в интернете и явное искажение действительности. Когда мы ввели слово «профессор», в выборке самых популярных изображений женщин было лишь около 10%; возможно, это попросту отражало то, как жизнь в колледже изображают в Голливуде, но такой расклад явно не соответствует современной реальности, в которой женщины составляют примерно половину академических преподавателей. Система подбора персонала на базе искусственного интеллекта, запущенная компанией Amazon в 2014 году,

показала себя в итоге настолько проблематичной, что через три года от нее пришлось отказаться. Мы не думаем, что подобные проблемы непреодолимы, — как мы увидим позже, смена парадигмы в разработке искусственного интеллекта может оказать здесь огромную помощь. Тем не менее сейчас общего решения для них не существует.

Основная сложность заключается в том, что современные системы искусственного интеллекта «зеркалят» входные данные независимо от их социальных ценностей, качества или эмоциональной окраски. Статистические данные правительства США говорят нам, что в настоящее время только 41% преподавателей — белые мужчины, но алгоритм поиска картинок в Google этого не знает; он просто сваливает в кучу все найденные изображения, не задумываясь о качестве и репрезентативности данных или об их семантике, выраженной в явной или неявной форме. Демография факультетской жизни постоянно прогрессирует, но слепые «датачерпатели» ничего не сортируют и выдают нам в качестве актуальной информации историю, а не существующие реалии.

Похожие соображения немедленно приходят в голову и когда мы вспоминаем о все возрастающей роли, которую системы искусственного интеллекта начинают играть в медицине [6]. Например, наборы данных, используемые для обучения программ диагностики рака кожи, могут быть ориентированы на белых пациентов и давать неверные результаты при использовании на пациентах с иным цветом кожи. Даже автомобили с автоматическим управлением могут быть менее надежными в распознавании темнокожих пешеходов, чем светлокожих. На карту уже давно поставлены жизни, однако современные системы все еще остаются неустойчивыми к человеческим предубеждениям.

Проблему номер пять можно описать как эффект «испорченного телефона». Сильная зависимость современного искусственного интеллекта от тренировочных наборов данных может привести к опасным последствиям из-за того, что ИИ-системы нередко обучаются на данных, которые они же сами генерировали в предшествующее время, а в них наверняка будут ошибки, связанные с более ранними периодами обучения. Например (мы обсудим это подробнее в главе 4), программы перевода работают, изучая сопряженные фрагменты из пары документов, один из которых является переводом другого. К сожалению, есть немало языков, на которых значительная часть текстов в интернете — в некоторых случаях до 50% всех веб-документов — не являются оригинальными, а были когда-то созданы программами машинного перевода из оригиналов, написанных на более распространенных языках. В результате, если, скажем, Google Translate допускает какую-либо ошибку при переводе, эта ошибка может надолго сохраниться в документе, осевшем в интернете, а этот документ затем попадает в число данных, использующихся при машинном обучении, что приводит к тиражированию ошибки и усиливает неточность связанных переводов.

Нечто подобное происходит и когда многие системы искусственного интеллекта полагаются на результаты работы краудсорс-исполнителей,

например, маркирующих изображения. По идее, они должны делать эту работу вручную, обеспечивая точность, типичную для человеческого восприятия, но иногда нанимаемые таким образом работники (контроль за деятельностью которых затруднителен) попросту халтурят, маркируя изображения не самостоятельно, а используя для этого тех же ботов, основанных на искусственном интеллекте. Хотя ученые и программисты, занимающиеся разработкой ИИ, в свою очередь, придумали технологии проверки того, выполняется ли та или иная работа людьми или же ботами, надзор за качеством работы становится в итоге игрой в кошки-мышки между разработчиками с одной стороны и лентяями-исполнителями, использующими боты, — с другой. Преимущество в этой игре постоянно переходит от одних к другим, и в результате многие данные, которые предположительно должны быть высококачественными (то есть созданными под контролем человека), оказываются на самом деле целиком сгенерированными машинами.

В-шестых, многие программы полагаются в значительной мере на данные, которыми может манипулировать широкая публика, а люди склонны к тому, чтобы пользоваться этим ради забавы или выгод. Microsoft Tay попал именно в такую ловушку. Поисковые системы Google регулярно попадают под обстрел пользовательскими постами, называемыми по-английски Google bombs. Суть его состоит в том, что пользователи создают большое количество постов определенного содержания и ссылок на них, так что поиск по некоторым терминам дает результаты, которые некоторые люди находят «прикольными». Например, в июле 2018 года интернет-шутникам удалось заставить Google Images выдавать по запросу «идиот» фотографии Дональда Трампа. (Эта «забава» продолжалась и позднее в том же году, когда Сундар Пичаи имел беседу с Конгрессом США.) Шестнадцатью годами ранее аналогичный розыгрыш (довольно неприличный по содержанию) был предпринят в отношении сенатора США Рика Санторума. Однако люди одурачивают Google не только чтобы повеселиться; существует целая индустрия поисковой оптимизации, связанная с манипулированием поисковыми системами для того, чтобы обеспечить различным клиентам высокий рейтинг при поиске в интернете.

В-седьмых, особую опасность представляет сочетание ранее существовавших социальных предубеждений и эффекта «испорченного телефона», которое может привести в итоге к усилению и закреплению неадекватных мер управления. Предположим, что некогда в определенных городах или штатах правоохранительные органы и уголовные суды были несправедливо предвзяты по отношению к определенной группе населения и выносили им приговоры гораздо чаще, чем остальным людям. Теперь администрация решает использовать методы искусственного интеллекта для работы с обширными данными, чтобы получать более объективные рекомендации по вопросам полицейской деятельности и вынесения приговоров. Однако программа обучается не только на современных данных, но и на большом количестве исторических, согласно которым опасность тех или иных людей или групп населения оценивается с точки зрения статистики

арестов и тюремных сроков. Программа увидит, что опасные преступники (идентифицированные в соответствии с указанным алгоритмом) происходят по большей части из определенных меньшинств, поэтому она будет рекомендовать, чтобы районы с более высоким процентом этих меньшинств получали большее внимание полиции и чтобы представители этих меньшинств арестовывались при первом подозрении и получали бы при прочих равных более длительные сроки. Когда программа стартует с новым набором данных, эти данные как бы подкрепляют сами себя, и система будет склонна давать те же предвзятые рекомендации, причем с еще большей уверенностью.

Как подчеркивала Кэти О'Нил, автор книги «Математика как оружие массового поражения» (Weapons of Math Destruction), даже если программа написана так, чтобы не использовать расовую и этническую принадлежность в качестве критериев, существуют так называемые прокси (всевозможные функции, связанные — то прямо, то косвенно — с критериями, которых мы хотим избежать), использование которых приведет к тому же самому результату; это может быть соседство по месту жительства, связи в социальных сетях, образование, работа, язык и, возможно, даже такие вещи, как предпочтения в одежде. Более того, решения, которые принимает программа, будучи чисто алгоритмической по своей природе, имеют ауру объективности, которая так вдохновляет публику и в которой так остро нуждаются чиновники и руководители компаний. Работа программ таинственна — данные обучения конфиденциальны, программа запатентована, процесс принятия решений, по сути, является «черным ящиком», в котором даже разработчики программ не могут разобраться до конца, — поэтому для большинства людей даже психологически почти невозможно оспаривать решения искусственного интеллекта, которые они воспринимают как явно несправедливые.

Несколько лет назад известная компания Xerox решила сократить отток сотрудников, который обходился им слишком дорого, поэтому они развернули программу для работы с большими данными, чтобы предсказать, как долго пробудет тот или иной сотрудник в их фирме. Программа обнаружила, что одним из самых сильных факторов быстрого увольнения является время в пути из дома на работу. В принципе неудивительно, что сотрудники, которым приходится добираться до своего рабочего места слишком долго, как правило, меняют место работы при первой возможности. Однако руководство Xerox осознало, что отказ от найма людей, живущих далеко, приведет к дискриминации людей с низким или средним уровнем дохода, поскольку все здания компании находились в богатом районе. К своей чести, руководство удалило этот критерий как неприемлемый. Но без тщательного контроля со стороны человека подобной предвзятости избежать крайне трудно.

Следующая, восьмая проблема современных систем искусственного интеллекта состоит в том, что цели, поставленные перед ними, очень часто реализуются совершенно бессмысленным образом. Исследовательница из компании DeepMind Виктория Краковна собрала десятки примеров этого. Так, робот, играющий в футбол, которого во время обучения поощряли касаться

мяча как можно чаще, изобрел неожиданную стратегию: он становился рядом с мячом и быстро вибрировал. Понятно, что это резко увеличивало число касаний, но к футболу такая игра не имеет никакого отношения. Другой робот, который должен был научиться поднимать конкретный объект, был обучен на изображениях того, как этот объект выглядит. Поэтому в итоге он решил, что достаточно просто поместить свою руку между камерой и объектом, чтобы на изображении было похоже, будто он действительно хватает предмет. Третий робот, чей искусственный интеллект, очевидно, отличался недостатком амбициозности, обучался игре в тетрис и в конечном счете догадался, что проще всего будет заморозить развитие игры на неопределенный срок: так он точно не проиграет.

Рис. 2.2. Робот, которого поощряли касаться мяча как можно больше раз, выработал стратегию: стоять рядом с мячом и быстро вибрировать

Проблема неправильно понятых целей может принимать и более тонкие формы. Еще на заре машинного обучения некая молочная компания наняла компьютерную фирму для создания ИИ-системы, которая могла бы предсказать, когда та или иная корова впадает в эструс. Технически цель, поставленная перед программой, состояла в том, чтобы с максимально возможной точностью генерировать бинарный прогноз «течка / отсутствие течки». Фермеры были очень рады узнать, что формальная точность системы оказалась на уровне 95%, но реальность оставила их куда менее довольными. Дело в том, что коровы впадают в течку только на один день в двадцатидневном цикле; соответственно, максимально вероятный суммарный прогноз по дням получается в том случае, когда на любой день выдается один и тот же прогноз, а именно — «течки нет». Благодаря этому в девятнадцати случаях из двадцати (а это и есть 95%) программа давала абсолютно

правильный прогноз, что делало ее на редкость точной, но вместе с тем — совершенно бесполезной. Значит, если мы не изложим подробно все, что мы действительно хотим получить на выходе, то решение, создаваемое искусственным интеллектом, может оказаться абсолютно ни к чему не пригодным.

Наконец, из-за огромных масштабов, в которых может функционировать нынешний искусственный интеллект, уже сейчас появилось много способов, которыми использующие его системы (даже в еще довольно примитивной форме) могут причинить серьезный вред обществу и конкретным людям, если окажутся в руках злоумышленников и просто непорядочных лиц. В частности, так называемые сталкеры [9] уже некоторое время используют несложные, но действенные системы искусственного интеллекта для мониторинга своих жертв и установления контроля за их жизнью. У спамеров история применения искусственного интеллекта куда дольше, с его помощью они идентифицируют потенциальных адресатов, уклоняются от проверки на веб-сайтах, предназначенных, чтобы убедиться, что пользователь — действительно человек (иконки-капчи и т.п.). Нет никаких сомнений и в том, что вскоре искусственный интеллект найдет применение в автономных системах вооружения [7], хотя у нас пока остается некоторая надежда на то, что такое оружие будет запрещено, подобно химическому или биологическому. Политолог из Государственного университета штата Нью-Йорк Вирджиния Ойбэнкс отмечает: «Когда ксенофобы переходят к использованию эффективных технологий против "чужих" и вообще "иных", то при отсутствии реальных механизмов защиты прав человека открывается огромный простор для всевозможных злодеяний».

Ничто из сказанного выше не означает, что искусственный интеллект не может с успехом работать на благо людей, однако это возможно лишь только при фундаментальном изменении парадигмы, к чему мы, собственно, и призываем в этой книге. Мы уверены, что многие из технических проблем могут быть решены уже сейчас, но современные подходы не позволяют этого и не хотят меняться сами. Тот факт, что современный искусственный интеллект остается слепым рабом данных и не обладает каким-либо реальным пониманием этических ценностей (которым обычно следуют сами программисты и разработчики систем), вовсе не делает искусственный интеллект непременно уязвимым к старым проблемам в будущем. Люди тоже часто ориентируются на данные, но мы не делаем абсурдных выводов о том, что почти все отцы и дочери принадлежат белой расе или что работа футболиста, которого поощряют чаще касаться мяча, состоит в том, чтобы стоять рядом с мячом и едва постукивать по нему, не отпуская от себя. Если люди могут избежать подобных элементарных ошибок, машины тоже должны этому научиться.

Нельзя сказать, что в принципе невозможно создать физическое устройство, которое может вести машину в условиях сильного снегопада, или машинный интеллект с этическими нормами; дело, однако, в том, что мы не достигнем этого, опираясь исключительно на большие данные.

Что нам действительно нужно, так это новый подход с гораздо большим погружением в ту проблему, которую мы хотим решить в первую очередь: как создать справедливый и безопасный мир для всех. Вместо этого мы видим вокруг себя методы искусственного интеллекта, которые решают отдельные, узкие проблемы, обходя при этом проблемы основные, хотя именно их они призваны решить в конечном счете. Образно говоря, мы надеемся, что нам поможет лейкопластырь, когда в реальности нам требуется пересадка мозга.

Например, IBM удалось решить проблему плохой гендерной идентификации представителей некоторых рас (ее обнаружила Джой Буоламвини), создав новый набор для обучения, в который включили большое число фотографий представительниц афро-американского населения. Google решила свою проблему с оскорбительными подписями «горилла» под фотографиями чернокожих противоположным образом, удалив изображения горилл из тренировочного набора. Однако ни одно из этих решений не является общим: оба приема были разработаны лишь для того, чтобы слепой анализ данных делал правильные вещи, оставаясь слепым, вместо того чтобы научить его видеть по-настоящему.

Точно так же можно решить частную проблему столкновений Tesla с машинами скорой помощи, стоящими на шоссе (например, во время помощи пострадавшим в ДТП), добавив более совершенные датчики и соответствующий набор примеров, но кто поручится, что этот же механизм сработает с эвакуаторами, которые случайно остановились на обочине шоссе? Или со строительными и ремонтными машинами? Допустим, Google сумела неплохо решить проблему с изображениями матери (которые сначала почему-то почти все были белокожими), но точно такая же проблема легко может возникнуть со словом «бабушка», и так до бесконечности.

И до тех пор пока нашим доминирующим подходом будет оставаться ориентация на узкий искусственный интеллект и все большие и большие наборы данных, у интеллектуальной индустрии есть все шансы надолго застрять, непрерывно латая лоскутное одеяло, находя краткосрочные решения для бесконечных частных проблем, не обращая внимания на основные недостатки общего подхода, которые и делают эти проблемы вездесущими и не устранимыми целиком.

Чтобы избежать этих ошибок, нам требуются по-настоящему умные системы. Сегодня, как нам кажется, почти все возлагают большие надежды на то, что вскоре они появятся. Как мы объясним в следующей главе, это, скорее всего, серьезное заблуждение.

ГЛАВА 3

Глубокое обучение и так далее

А относительно идей, сущностей, абстракций и трансценденталий мне так и не удалось внедрить в их головы ни малейшего представления.

Джонатан Свифт. Путешествия Гулливера

Одно дело — ожидать, что элементарные частицы подчиняются простым универсальным законам. Другое дело — требовать того же от человеческой расы.

Сабина Хоссенфельдер. Затерянные в математике (Lost In Math)

Бóльшая часть нынешнего энтузиазма по поводу искусственного интеллекта проистекает из одного простого соображения: при прочих равных условиях чем больше у вас данных, тем лучше. Если вы хотите предсказать исход следующих выборов и можете опросить только 100 человек — что ж, надеюсь, вам повезет; если вы можете взять интервью у 10 000, ваши шансы на реальный успех намного выше.

На самом деле в первые годы существования искусственного интеллекта данных было немного, и методы, основанные исключительно на данных, не доминировали среди разработчиков. В большинстве исследований использовался подход, основанный на знаниях, иногда называемый «старым добрым ИИ» (Good Old Fashioned AI, или GOFAI) или же «классическим искусственным интеллектом». При классическом подходе исследователи обычно вручную кодируют те знания, которые машинный разум затем должен будет использовать для выполнения определенной задачи, а затем пишут компьютерные программы, которые используют эти знания, применяя их к различным когнитивным задачам, таким как понимание текста, проектирование роботов или поиск доказательств различных теорем. О сегодняшних больших данных речь еще даже не заходила, и старые программы редко полагались на использование значительных объемов информации как на панацею.

Хотя создать лабораторные прототипы искусственного интеллекта, основанные на классическом подходе, было в принципе возможно (впрочем, часто с огромными усилиями), пройти через этап тестирования, чтобы стать массовым продуктом, оказывалось для них крайне затруднительно. Вследствие этого общее количество классических систем искусственного интеллекта, имеющих какое-либо практическое значение, весьма невелико. Подобные методы все еще широко используются в определенных областях, таких как планирование маршрутов для роботов или GPS-навигация. В целом, однако, подход, ориентированный на знания, был почти целиком вытеснен машинным обучением, которое обычно пытается извлечь уроки из массовых данных, а не полагается на специализированные компьютерные программы, использующие знания, оцифрованные вручную.

Машинное обучение на самом деле тоже имеет очень долгую историю и восходит к 1950-м годам, когда Фрэнк Розенблатт создал так называемую нейронную сеть [10] — одну из первых систем машинного обучения, задачей которой было научиться распознавать окружающие объекты (с помощью камеры), не требуя при этом от программистов заранее предвидеть все сложности, связанные с распознаванием. Эти системы почти сразу обратили на себя внимание специалистов и публики и в 1958 году благодаря нашумевшим публикациям в *The New York Times* получили очень широкую известность. Однако волна популярности быстро схлынула, поскольку реализация проекта

была подорвана многочисленными техническими проблемами. Сети Розенблатта (которые, естественно, должны были работать на оборудовании 1950-х годов) не обладали достаточной мощностью; пользуясь современной терминологией, следовало бы сказать, что они были недостаточно глубоки (мы чуть позже объясним, что именно это означает). У тогдашних цифровых камер было слишком низкое разрешение — всего 20 × 20 (400) пикселей (что примерно в 30 000 раз меньше, чем разрешение камеры современного iPhone X), следовательно, изображения, получаемые в 1950-х, буквально утопали в мозаике пикселей. Оглядываясь назад, мы понимаем, что идея Розенблатта была хороша, но реальные системы, которые он мог создать тогда на практике, просто были не в состоянии выполнить задачи, ради которых их спроектировали.

Впрочем, аппаратное обеспечение оказалось только частью проблемы. Именно сейчас мы как никогда четко осознали, до какой степени машинное обучение зависит от доступности большого количества данных, таких, например, как маркированные фотографии. У Розенблатта же ничего подобного не было, поскольку не существовало, в частности, интернета, из которого он мог бы извлечь необходимое число примеров.

Целый ряд ученых продолжил работать в русле, намеченном исследованиями Розенблатта, в течение нескольких последовавших десятилетий. Тем не менее до недавнего времени его преемники фактически не смогли добиться чего-то реально большего. До тех пор пока большие данные не стали обычным явлением, в мире разработчиков искусственного интеллекта подход, основанный на нейронных сетях, считался абсолютно бесперспективным, поскольку по сравнению с другими методами эти системы демонстрировали безнадежное отставание.

Но когда в начале 2010-х годов произошла революция в области больших данных, у нейронных сетей появился наконец хороший шанс вырваться вперед. Такие ученые, как Джефф Хинтон, Иошуа Бенжио, Ян Лекун и Юрген Шмидхубер, остававшиеся верными своему методу даже в самые унылые для них 1990-е и 2000-е годы, когда большинство их коллег ушли в другие области, теперь буквально взвинтили темп исследований.

Особенно важно отметить, что самый серьезный скачок в развитии нейронных сетей произошел не из-за технического прорыва в математической реализации метода (поскольку большая часть математического аппарата здесь была разработана еще в 1980-х годах), а из-за резко возросшей популярности компьютерных игр или, более конкретно, вследствие огромного прогресса, достигнутого в производительности одного конкретного вычислительного модуля, называемого на пользовательском жаргоне видеокартой, а на техническом языке — графическим процессором, или GPU (сокращение от graphics processing unit). Именно его разработчики нейронных сетей стали использовать для развития искусственного интеллекта. Мощные графические процессоры изначально (уже с 1970-х годов) разрабатывались для видеоигр, а в 2000-х годах нашли применение и в нейронных сетях. К 2012 году видеокарты стали чрезвычайно мощными, и для определенных целей они сделались уже

даже более эффективными, чем центральный процессор компьютера (CPU, central programming unit) — традиционное ядро большинства вычислительных машин. Революция в производительности нейронных сетей произошла в том же 2012 году, когда несколько человек, включая команду исследователей, работающих с Хинтоном, разработали способ использования возможностей графических процессоров для резкого увеличения производительности систем, придуманных Розенблаттом.

Впервые в истории искусственного интеллекта команда Хинтона и другие разработчики нейронных сетей неожиданно начали устанавливать рекорды, особенно при распознавании изображений из базы данных ImageNet, которую мы упоминали ранее. Вступив в конкуренцию, Хинтон и его сторонники сосредоточились на создании подмножества этой базы данных, куда поместили 1,4 млн изображений из тысячи различных категорий. Каждая команда обучила свою систему примерно 1,25 млн из них, оставив еще 150 000 для тестирования. До этого, в контексте более старых методов машинного обучения, правильный результат в 75% уже считался очень хорошим, однако команда Хинтона, используя глубокую нейронную сеть, набрала 84% правильных результатов, а другие исследовательские группы вскоре добились еще большего. К 2017 году показатели правильности маркировки изображений, основанные на глубоком обучении, достигли почти абсолютной точности — в 98%.

Ключом к этому внезапно обретенному успеху стало то обстоятельство, что новое поколение графических процессоров позволило Хинтону и его коллегам обучать нейронные сети, которые, технически говоря, были гораздо глубже, то есть имели больше слоев (наборов отдаленно похожих на нейроны элементов, о которых мы расскажем чуть позже), чем удавалось задействовать до сих пор. Тренировать глубокую сеть — это значит скормить сети множество примеров изображений, помеченных правильными ярлыками, скажем, это — фотография собаки, это — изображение кошки и т.д.; такая методика обычно известна под названием «контролируемое обучение». Использование мощных графических процессоров означало, что теперь за короткое время можно было пропустить информацию через значительно большее число слоев, и результаты, естественно, стали лучше.

Получив в распоряжение революционные графические процессоры и колоссальные библиотеки ImageNet, глубокое обучение стало готовиться к новому этапу гонки. Началось все с того, что Хинтон с группой аспирантов создали новую IT-компанию и вскоре выставили ее на продажу с аукциона. Наибольшую цену предложил Google, и через два года он купил стартап DeepMind за более чем 500 млн долларов. С этого момента началась революция в области глубокого обучения.

Разумеется, глубокое обучение является лишь одним из множества подходов в попытке заставить машины изучать реальные вещи на основе виртуальных данных. Чаще всего это делается с применением различных статистических методов. Допустим, у вас есть книжный интернет-магазин и вы

хотите рекомендовать те или иные издания своим клиентам. Один из подходов состоит в том, чтобы решить самому, какие книги окажутся наиболее интересными и продаваемыми. Вы можете разместить их на первой полосе сайта, так же как продавцы в реальных магазинах размещают самые перспективные книги перед входом с улицы. Другой же подход состоит в том, чтобы узнать, что именно нравится людям, на основе имеющихся данных (скажем, отзывов или объемов продаж). Причем можно выяснить не только то, что людям больше нравится в целом, но и что может привлечь конкретного клиента — на основании данных о том, что он покупал раньше. Вы можете, например, заметить, что люди, которым нравятся книги о Гарри Поттере, также часто покупают толкиновского «Хоббита», а люди, которым нравится Толстой, часто покупают романы Достоевского. По мере того как ваша база данных о клиентах растет, количество возможных сочетаний и предпочтений становится все больше, в конце концов оказывается слишком сложно отслеживать все это по отдельности, и вот вы уже пишете компьютерную программу для отслеживания множества запутанных связей.

Затем, пользуясь новым методом, вы изучаете статистику, скажем, вероятность того, что покупатель, купивший книгу 1, также купит книгу 2, книгу 3 и т.д. Как только вы разберетесь в более простых вещах, вы начнете отслеживать более сложные корреляции, например вероятность того, что клиент, который купил и роман о Гарри Поттере, и «Хоббита», но прошел мимо «Звездных войн», купит тем не менее научно-фантастический роман Роберта Хайнлайна. Искусство делать обоснованные предположения на базе множества данных — это одно из самых обширных и процветающих направлений работы искусственного интеллекта в сфере машинного обучения.

Приведенная на следующем рисунке диаграмма Венна (рис. 3.1) — удобный способ поразмышлять о связях между глубоким обучением, машинным обучением и искусственным интеллектом. В общем смысле ИИ включает в себя и машинное обучение, но сюда относятся также любые алгоритмы и знания, необходимые для его работы, в частности те, которые написаны вручную или созданы традиционными методами программирования, а не получены путем обучения. Машинное обучение, в свою очередь, включает в себя любую технику, которая позволяет машине учиться на данных. Наконец, глубокое обучение является наиболее известным из этих методов машинного обучения, но отнюдь не единственным.

Мы сейчас так сосредоточены именно на глубоком обучении лишь потому, что оно является основным направлением, куда направлено большинство текущих инвестиций в сфере разработки искусственного интеллекта, как в академической науке, так и в промышленности. И все же глубокое обучение вовсе не представляет собой единственный (или лучший) подход к машинному обучению или к развитию искусственного интеллекта в целом. Например, еще один вариант машинного обучения состоит в построении так называемых деревьев решений, которые по сути являются системами простых правил, позволяющих фильтровать или группировать данные (рис. 3.2).

Рис. 3.1. Диаграмма Венна. Основные области, в которых задействован искусственный интеллект

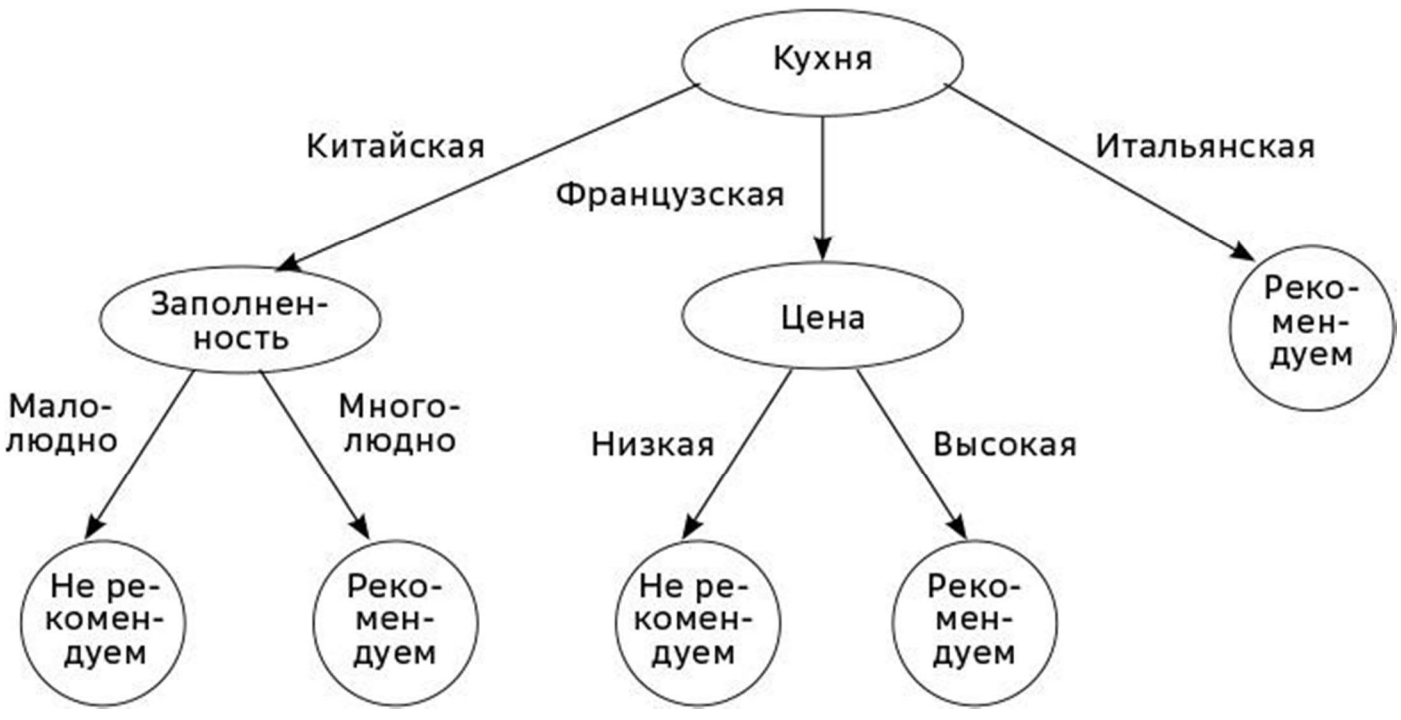


Рис. 3.2. Древо решений, используемое для выбора ресторана на основе клиентских предпочтений

Метод опорных векторов (*англ.* support vector machine, или SVM) — техника анализа, которая организует данные в сложные абстрактные «гиперкубы», доминировавшая в машинном обучении на протяжении всего первого десятилетия XXI века, и люди использовали ее для расчета множества вещей, от заголовков новостных статей до структур белков. Статистические модели, лежащие в основе данного подхода, пытаются рассчитать вероятность

правильности различных ответов на задаваемый вопрос и возвращают ответ, который они считают наиболее подходящим. Этот подход в свое время оказался ключом к успехам, достигнутым ИИ-системой IBM Watson, и нет никаких причин сомневаться в его долговременной перспективности.

Еще один подход, иногда называемый генетическими алгоритмами, основан на имитировании эволюции алгоритмов, которые улучшаются по мере тестирования и видоизменения. Наиболее «приспособленные» [11] алгоритмы «выживают» и «размножаются». Генетические алгоритмы используются в самых различных приложениях, начиная от разработки радиоантенн и заканчивая ботами, играющими в видеоигры, — здесь они порой достигают того же уровня эффективности, что и системы, основанные на глубоком обучении. Этот список можно продолжать и продолжать; мы не будем вдаваться подробно во все варианты и намеренно сосредоточимся на глубоком обучении, потому что в последние годы оно достигло почти абсолютного господства, но, если вы хотите узнать больше о различных алгоритмах, мы очень рекомендуем книгу Педро Домингоса The Master Algorithm [12]. Мы разделяем мнение Домингоса, что каждый алгоритм может внести свой вклад в развитие искусственного интеллекта и что в целом набор алгоритмов еще недостаточно хорошо интегрирован. Однако на следующих страницах нашей книги мы расскажем, почему мы не столь оптимистичны в отношении той идеи автора, что однажды найдется некий оптимальный «основной алгоритм», который станет универсальным решением для всего искусственного интеллекта.

Многие интеллектуальные технологии, такие как выстраивание маршрутов в навигаторе или координация действий, выполняемых роботами, все еще успешно используют методы, взятые из классического искусственного интеллекта, которые практически не используют машинное обучение. И зачастую одно и то же приложение, например алгоритмы маршрутизации трафика, используемые Google-навигатором Waze, использует несколько методов, основанных то на классическом искусственном интеллекте, то на машинном обучении.

Итак, в последние годы машинное обучение стало повсеместно доминирующим методом разработки искусственного интеллекта, чему во многом способствуют массовые данные. Система IBM Watson опиралась на огромные базы данных и использовала сочетание классических методов искусственного интеллекта и вероятностного машинного обучения, чтобы обеспечить своему детищу победу в Jeopardy!. Городские власти используют машинное обучение для правильного распределения усилий в сфере услуг, сервисы совместного использования автомобилей (каршеринг) применяют машинное обучение для прогнозирования спроса на машины, а полицейские управления предотвращают преступления с помощью машинного обучения. Применение машинного обучения в области рекламы буквально необъятно. Так, Facebook использует обучаемые системы, чтобы определить, какие новостные сюжеты вы захотите увидеть в своей ленте и заодно — какая реклама сможет вас заинтересовать хотя бы до степени одного клика.

Поисковый движок Google использует машинное обучение, чтобы рекомендовать видео, размещать рекламу, понимать вашу речь и пытаться предугадать, что вы пытаетесь найти в каждом конкретном случае. Веб-сайт Amazon применяет ту же технологию, чтобы рекламировать различные товары и интерпретировать результаты вашего поиска. Устройство Amazon Alexa использует машинное обучение для декодирования ваших запросов и т.д.

Ни один из этих продуктов не идеален; позже мы обсудим различные примеры того, как известные коммерческие поисковые системы оказывались сбиты с толку даже простейшими запросами. Но, с другой стороны, любая из них все же гораздо лучше, чем вообще ничего, и, следовательно, они имеют ту или иную экономическую значимость. Никто из людей не смог бы написать поисковый браузер в масштабе всей Всемирной паутины; и Google просто не может существовать без машинного обучения. Предложения товаров на Amazon были бы значительно менее эффективными, если бы они полагались только на людей. (Пример максимального приближения к ручному управлению рекомендациями — это сайт Pandora, работающий как сервис, рекомендующий различный музыкальный контент, где подборки в основном делаются вручную экспертами-людьми, из-за чего его музыкальная библиотека намного меньше, чем у сопоставимых ИИ-систем, таких как Google Play, которые в основном базируются на машинном обучении.) Автоматизированные рекламно-рекомендательные системы, которые приспосабливают свои рекомендации под запросы отдельных пользователей, основываясь на статистике приобретений людей с похожей историей покупок, вовсе не обязаны быть совершенными, ведь даже если они иногда и ошибаются, то все равно остаются куда более точными, чем устаревшие методы, связанные с размещением больших объемов рекламы в газете. В 2017 году Google и Facebook вместе заработали на размещении рекламы более 80 млрд долларов, и машинное обучение, основанное на статистическом анализе, было центральным двигателем этого процесса.

Глубокое обучение основано на двух фундаментальных идеях. Первая, которую можно назвать иерархическим распознаванием образов, частично связана в своем происхождении с рядом экспериментов, проведенных в 1950-х годах нейробиологами Дэвидом Хьюбелом и Торстеном Визелем, получившими за них Нобелевскую премию 1981 года по физиологии и медицине. Хьюбел и Визель обнаружили, что разные нейроны в зрительной системе реагируют на зрительные раздражители очень неодинаково. Некоторые наиболее активно реагировали на совсем простые стимулы, такие как линии с определенной ориентацией, в то время как другие более энергично реагировали на более сложные образы. Теория, которую они предложили для объяснения этого феномена, заключалась в том, что сложные стимулы могут распознаваться через иерархию возрастающей абстракции, например от строк до букв и слов. В 1980-х годах японский первопроходец в области конструирования нейронных сетей Кунахико Фукусима сумел вплотную подойти к важнейшей вехе в истории искусственного интеллекта, построив вычислительную реализацию идеи Хьюбела и Визеля, названную

«Неокогнитрон», и показав, что она может эффективно работать в некоторых аспектах компьютерного зрения. Позднее Джефф Хокинс и Рэй Курцвейл отстаивали в своих книгах ту же идею.

«Неокогнитрон» состоял из набора слоев (которые на схеме выглядят как прямоугольники). Если идти слева направо по рис. 3.3, то сначала мы увидим входной слой, который воспринимает определенные стимулы (по сути — это просто пиксели в цифровом изображении), затем идут один за другим следующие слои, которые анализируют изображение, выискивая различия в контрасте, структуре краев и т.д., и завершается все выходным слоем, который определяет, к какой категории относится входящее изображение. Соединения между слоями позволяют провести соответствующую обработку в определенной последовательности. Все эти идеи — входные слои, выходные слои и внутренние слои со связями между ними — теперь являются технической основой глубокого обучения.

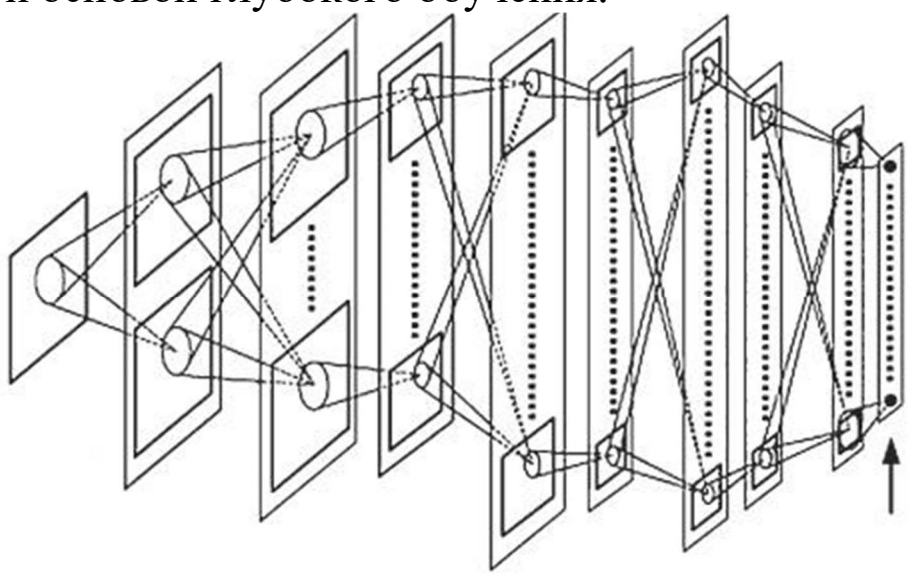


Рис. 3.3. Схема «Неокогнитрона» — нейронной сети для распознавания объектов

Подобные системы называются нейронными сетями, потому что каждый слой состоит из элементов, называемых узлами, которые можно (весьма условно, конечно) сравнить с очень упрощенными нейронами. Между этими узлами существуют соединения, также называемые взвешенными соединениями или просто весами; чем больший вес имеет соединение узла A и узла B, тем больше влияние A на B. То, что нейронная сеть выдает как результат, является функцией взаимодействия слоев через эти весы.

Вторая базовая идея — это собственно обучение. Усиливая весовые коэффициенты для определенной конфигурации входов и выходов, можно обучить сеть ассоциировать каждый конкретный вход с соответствующим выходом. Предположим, например, что вы хотите, чтобы сеть изучала названия различных букв, представленных в виде сетки пикселей. Изначально система не знает, какой шаблон пикселей соответствует какой букве. Со временем, путем проб, ошибок и корректировок, она будет связывать наличие пикселей в верхней горизонтальной части сетки с определенными буквами, такими как T и E, а вертикальные линии пикселей на левом краю — с другим

набором букв, в частности E, F и H. Постепенно система осваивает все больше и больше корреляций между пикселями в разных местах и соответствующими буквами. Уже в 1950-х годах Розенблатт интуитивно понимал, что такой подход может быть абсолютно жизнеспособным, но сети, которые он использовал, были слишком простыми и ограниченными — они содержали только входной и выходной слои. Если задача была достаточно простой (например, разделение множества плоских фигур на круги и квадраты), то некоторые довольно элементарные математические приемы гарантировали, что вы всегда сможете корректировать свои веса так, чтобы получить в итоге правильный ответ. Но для более сложных задач было недостаточно двух уровней — требовались промежуточные слои, представляющие комбинации различных признаков, и в то время ни у кого не было работоспособного решения для надежного обучения сетей с большей глубиной, то есть которые имели бы внутренние уровни. Примитивные нейронные сети того времени имели входы (например, изображения) и выходы (метки), но между ними больше ничего не было.

В своей влиятельной книге «Перцептроны» (Perceptrons) 1969 года Марвин Мински и Сеймур Паперт математически доказали, что простые двухслойные сети не способны охватить многие из тех объектов, которые разработчики, вероятно, захотели бы классифицировать с помощью таких систем. Кроме того, возможность обучения сети для нахождения удовлетворительного решения не гарантируется при такой примитивной структуре [13]. Впрочем, авторы несколько переборщили с пессимизмом, заявив (на основании чисто интуитивных соображений), что расширение нейронных сетей до нескольких слоев окажется бесполезной тратой сил, хотя и не отрицали теоретическую возможность того, что «когда-нибудь, возможно, будет найдена перспективная теорема о сходимости подобных [более сложных] систем». В атмосфере пессимизма и при отсутствии убедительных результатов, доказывающих работоспособность двухслойных нейронных сетей, все это направление исследований довольно быстро зачахло. Решение простых проблем (типа разделения фигур на круги и квадраты) казалось занятием скучным и лишенным перспектив, а более сложные задачи выглядели неразрешимыми.

Но сдались не все. Как позднее признали Мински и Паперт, они на самом деле не доказали, что от более глубоких сетей ничего нельзя добиться, а только продемонстрировали, что нельзя гарантировать хорошие результаты, используя ту конкретную математику, которую они рассматривали в момент написания книги. И действительно, из того, что Мински и Паперт написали в 1969 году, все еще остается верным следующее: глубокое обучение до сих пор не дает математически обоснованных гарантий сходимости (получения верного ответа), за исключением нереального случая, когда вам доступны бесконечные данные и бесконечные вычислительные ресурсы. Однако задним числом ясно, что эти авторы недооценили, насколько полезными могут оказаться более глубокие сети даже при отсутствии формальных доказательств их универсальной работоспособности. В течение последующих двух десятилетий несколько человек, включая Джеффа Хинтона и Дэвида

Румелхарта, независимо друг от друга изобрели математику, которая позволяет более глубоким нейронным сетям выполнять удивительно качественную работу, несмотря на отсутствие каких-либо формальных математических гарантий совершенства [8].

Что означает «эффективность без гарантии», можно показать на следующем примере. Представьте, что вам нужно взобраться на гору и наивысший балл получает тот, кто сумеет покорить самую вершину. Соответственно, если просто топтаться у подножья, то не заработаете вообще никаких баллов, поднявшись до середины, получите средний балл и т.д. В случае систем распознавания образов вершина будет соответствовать максимальной точности, подножье — минимальной, склоны — промежуточной [14].

Так вот, Хинтон и другие сторонники нейронных сетей обнаружили, что, хотя в более глубоких системах, содержащих более двух уровней, невозможно гарантировать совершенство (то есть восхождение прямо на вершину), можно тем не менее создать программы, которые зачастую дают достаточно хорошие результаты, соответствующие средним или верхним частям склона. Система при этом делает небольшие, но уверенные шаги в направлении вершины, используя технику, называемую методом обратного распространения ошибки (*англ.* backpropagation), который теперь лежит в основе глубокого обучения [15].

Метод обратного распространения ошибки работает через оценку того, какой путь из любой точки на склоне соответствует скорейшему подъему в гору. Достижение самого пика горы при этом не гарантируется — продвижение может застопориться на том или ином локальном максимуме (в нашей терминологии это будет возвышение на склоне, которое находится выше окружающей территории, но все еще намного ниже главной вершины). На практике данная техника часто приводит к адекватным, даже высоким результатам.

Существует также алгоритм, называемый свертыванием (*англ.* convolution). Его в конце 1980-х годов ввел в практику Ян Лекун, и он все еще широко используется, поскольку позволяет системам распознавания образов значительно повысить эффективность, создавая массивы соединений, обеспечивающих узнавание конкретных (уже известных) объектов внутри более сложных изображений, независимо от того, где они там появляются.

Хотя математический аппарат, казалась, был разработан корректно, первоначальные результаты, демонстрируемые нейронными сетями, оставались неубедительными. Было известно, что в принципе, если вы сможете найти правильный набор весов (большой, но часто вполне управляемый), то нейронная сеть с тремя или более слоями может позволить вам решить фактически любую проблему, которую вы поставите перед компьютером, при условии что у вас достаточно данных, вы имеете множество узлов в системе и у вас хватит терпения заставить все это функционировать. Однако на практике это не срабатывало: для решения действительно интересных задач требовалось очень большое количество узлов, и компьютеры того времени не могли

выполнить все необходимые вычисления, связанные с настройкой множества узлов, за разумное время.

У людей, работавших в этой области, было сильное предчувствие, что большее число слоев — то есть более глубокие сети — поможет решению проблемы. Но никто не знал этого наверняка. Еще в начале 2000-х годов аппаратное обеспечение просто не подходило по своей производительности для экспериментов в этом направлении. Для обучения типичной глубокой сети потребовались бы недели или даже месяцы компьютерного времени. Вы не смогли бы (как это доступно сейчас) перепробовать сотню различных вариантов и найти среди них лучший путем обычной сортировки. Первые результаты были многообещающими, но не были в состоянии конкурировать с другими подходами.

Именно здесь и вступили в игру графические процессоры, чтобы стать главным катализатором прогресса в глубоком обучении (не считая еще некоторых важных технических находок [16]). Идея заключалась в том, чтобы выяснить, насколько эффективно можно было бы использовать графические процессоры нового поколения для обслуживания моделей нейронных сетей с большим количеством уровней. Благодаря тому что видеокарты смогли выполнять алгоритмы перенастройки множества узлов за разумное время, глубокое обучение — которое подразумевает тренировку сетей с четырьмя или более слоями (иногда бывает и более ста) — наконец-то стало осуществимым на практике.

Все последующее время результаты глубокого обучения действительно демонстрировали замечательный прогресс. Другие исследователи годами разрабатывали хитроумные методы, пытаясь заставить распознавание объектов работать на различных машинах, но теперь внезапно выяснилось, что все это можно заменить системой глубокого обучения, которая тратит на вычисления всего несколько часов или дней. Успех нового метода позволил исследователям взяться и за новые проблемы, — и это были не только рекомендации по рекламе, но также и расшифровка речи или распознавание объектов, — которые так и не были адекватно решены с использованием старых подходов к машинному обучению.

Глубокое обучение современного уровня бьет все новые и новые рекорды. Например, как подробно объясняется в обширной статье, опубликованной в *The New York Times Magazine*, глубокое обучение радикально улучшило работу приложения Google Translate. До 2016 года гугловский переводчик использовал классические методы машинного обучения, задействуя огромные таблицы шаблонов соответствия на двух языках, помеченных маркерами вероятности. Более новый подход с использованием глубокого обучения нейронных сетей позволил значительно улучшить качество переводов. Применение аналогичных систем в других областях привело к значительным улучшениям в машинной транскрипции речи и в маркировке фотографий и других изображений.

Кроме того, во многих (хотя и не во всех) случаях глубокое обучение легче использовать на практике. Традиционное машинное обучение часто опирается на опыт программиста в разработке тех или иных машинных функций. Например, в области компьютерного зрения опытные инженеры, обладающие знаниями о человеческом зрительном восприятии, пытались найти общие свойства в различных изображениях, которые стали бы полезными для машин, пытающихся «понять» изображения, например края, углы и пятна. Еще в 2011 году хорошими инженерами инженерами по машинному обучению часто становились те, кто умел находить подходящие машинные функции для решения конкретной проблемы [17].

Глубокое обучение изменило расклад сил и здесь, по крайней мере до некоторой степени. Во многих задачах (хотя, как мы увидим, не во всех) глубокое обучение может хорошо работать без предварительной разработки функций. Системы, которые начали побеждать в задачах по ImageNet, научились классифицировать объекты — на самом современном уровне — без существенного прогресса в области разработки функций. Вместо этого системы выучили все, что им нужно было знать, просто посмотрев на пиксели, из которых состояли изображения, и метки, которые они должны были применять в качестве подписей. Проектирование функций, казалось, потеряло актуальность. Зачем быть доктором наук в области зрительного анализа, если можно обучить компьютер просто на большом числе частных примеров?

Более того, глубокое обучение оказалось поразительно универсальным подходом, эффективным не только при распознавании объектов и дешифровке речи, но также и для многих других задач, которые раньше считались неподвластными для машин. Глубокое обучение обнаружило неожиданный успех в раскрашивании черно-белых фотографий и даже в создании образцов синтетического искусства, например имитируя стиль старых мастеров на современных изображениях (вы можете, скажем, снять пейзаж на фотокамеру и превратить его в подобие картины Ван Гога). Его реально использовать для решения проблемы обучения без учителя, при котором нет заранее подобранных примеров и нет обучающего человека, и машина учится, выбирая примеры сама. Особенно хорошо работают такие методы в сочетании с подходом, известным как генеративно-состязательные сети (generative adversarial networks).

Глубокое обучение может также использоваться в качестве компонента в системах, которые играют в игры, иногда на сверхчеловеческом уровне. Громкие успехи DeepMind — сначала в видеоиграх Atari, а затем в го — частично основывались на использовании глубокого обучения в сочетании с подкрепляемым обучением, что в результате привело к созданию нового метода, известного как глубокое обучение с подкреплением. По сути своей это способ обучения методом проб и ошибок, но с задействованием колоссального количества данных. Система AlphaGo, как мы увидим ниже, интегрировала в свои алгоритмы и некоторые другие методы.

Успех метода на многих подействовал опьяняюще. Эндрю Ын, один из ведущих исследователей искусственного интеллекта, руководивший

исследованиями в китайской компании Baidu, разрабатывавшей новую поисковую систему для интернета, писал на страницах *Harvard Business Review* в 2016 году следующее: «Если обычный человек может выполнить некую умственную задачу менее чем за одну секунду, мы, вероятно, сможем сейчас или в ближайшем будущем автоматизировать даже этот процесс, используя методы искусственного интеллекта». Стоит ли пояснять, что он при этом имел в виду именно глубокое обучение, успехи которого воспарили до небес?

Невзирая на все это, мы с самого начала отнеслись к этому методу весьма скептически. Пусть даже всем было очевидно, что глубокое обучение сделалось более мощным инструментом, чем любой из ранее применявшихся методов, нам представлялось, что возможности данного подхода сильно переоцениваются. Руководствуясь своими исследованиями, проведенными за дюжину лет до того по системам — предшественникам глубокого обучения, Гэри в 2012 году написал в *The New Yorker* статью, содержавшую следующий абзац:

Куда реалистичнее будет сказать, что глубокое обучение — это только часть более сложной задачи создания интеллектуальных машин. В методах, подобных этому, отсутствуют способы представления причинно-следственных связей (например, между болезнями и их симптомами), и они могут испытать серьезные затруднения при столкновении с абстрактными идеями, например такими, как «родственный» или «идентичный». У них нет очевидных способов формирования логических выводов, и они также еще далеки от интеграции абстрактного человеческого знания…

Несколько лет спустя все это так и останется прежним, несмотря на очевидные успехи глубокого обучения в определенных областях, таких как распознавание речи, перевод с другого языка, преобразование голоса в текст и компьютерные команды. Глубокое обучение не показало себя универсальным решателем проблем, и вдобавок оно не имеет ничего общего с универсальным искусственным интеллектом, в котором мы нуждаемся для использования в условиях открытых систем.

В частности, оно сталкивается с тремя основными проблемами, каждая из которых затрагивает как само глубокое обучение, так и другие популярные ныне методы, подобные глубокому обучению с подкреплением, в значительной степени зависящие от него, что следует уже из самих их названий.

Глубокое обучение чрезвычайно прожорливо в отношении использования данных. Чтобы правильно установить все соединения в нейронной сети, глубокое обучение часто требует фантастического объема входящей информации. Системе AlphaGo потребовалось 30 млн игр для достижения сверхчеловеческой производительности, а это гораздо больше, чем любой человек мог бы сыграть за всю свою жизнь. При небольших объемах данных глубокое обучение часто неэффективно. Его сильная сторона начинает проявляться на наборах, насчитывающих миллионы или даже миллиарды точек

данных, которые давят на наборы весов нейронных сетей, фиксирующих взаимоотношения между обучающими примерами. Если системам глубокого обучения скормить лишь ограниченное число примеров, результаты их тренировок едва ли будут отличаться надежностью. И, разумеется, отличие их от нас состоит в том, что очень многое из того, что мы делаем, основано буквально на нескольких минутах изучения. Скажем, когда вы впервые возьмете в руки 3D-очки, вы сможете их надеть и примерно поймете, что происходит, и без необходимости примерить их сотню тысяч раз. Глубокое обучение попросту не предназначено для столь быстрого формирования навыков.

В статье, где Эндрю Ын обещал, что машины смогут в скором времени автоматизировать все, что человек может сделать за секунду, более реалистично было бы выразиться так:

Если обычный человек может выполнить умственную задачу менее чем за одну секунду и мы можем собрать огромное количество данных, имеющих непосредственное отношение к моделируемой задаче, у нас есть шансы на успех — и то лишь до тех пор, пока проблемы, с которыми мы на самом деле сталкиваемся, не отклонятся от данных обучения, а область задач и правила игры не изменятся со временем.

Поправки к сказанному Эндрю Ыном в первую очередь характеризуют настольные игры, такие как го или шахматы, в которых правила остаются неизменными на протяжении тысячелетий, но, как мы отмечали во введении, во многих реальных проблемах получить достаточное число релевантных данных почти, а то и совсем невозможно. Например, большая часть проблем, тормозящих применение глубокого обучения в лингвистике, заключается в том, что в любом языке существует практически бесконечное число предложений с разным смыслом, причем внешне они могут быть очень похожи, а по смыслу — радикально различны. Чем больше реальные проблемы отличаются от тренировочных данных, представленных системе, тем меньше вероятность того, что после обучения система будет работать надежно.

Глубокое обучение очень непрозрачно. Классические экспертные системы основаны на правилах, которые можно достаточно легко понять в рамках человеческой логики или интуиции, например «если у человека повышенный уровень лейкоцитов в крови, то, вероятно, у человека где-то есть инфекция». В отличие от них, нейронные сети состоят из обширных массивов чисел, практически ни один из которых не имеет прямого или интуитивного смысла для подавляющего большинства людей. Даже эксперты в этой области, притом вооруженные сложными аналитическими инструментами, часто не могут до конца понять, почему определенные нейронные сети принимают те решения, которые дают на выходе. Почему нейронным сетям удается во многих случаях работать столь эффективно, по сути, остается нерешенной загадкой; точно так же отсутствует ясность относительно того, в каких именно обстоятельствах они не работают или работают плохо. Допустим, нейронная сеть, решающая определенную задачу, может при тестировании показывать точность, равную 95% правильных результатов. Но что именно это значит?

Часто бывает очень трудно понять, почему сеть ошибается в оставшихся 5% случаев, при том что некоторые из этих ошибок являются очень серьезными — такими, которые ни один человек никогда не совершит, как тот описанный выше случай, когда система не видит разницы между холодильником и дорожным знаком. Подобные ошибки нередко критичны с точки зрения нашей безопасности, и если мы не можем понять, откуда они берутся, то мы имеем полное право говорить о фундаментальной слабости глубокого обучения.

Описанная проблема стоит особенно остро потому, что нейронные сети не могут объяснить понятным человеку образом, почему они дают те или иные ответы, хоть правильные, хоть нет [18]. Действительно, эти системы работают в стиле «черного ящика», то есть они делают то, что делают, а что у них происходит внутри — остается тайной. Но если мы собираемся рассчитывать на них в таких серьезных вещах, как вождение автомобиля или выполнение ежедневных домашних обязанностей, то оставлять эту ситуацию как есть очень опасно. Такая же ловушка поджидает нас, если мы захотим сделать глубокое обучение частью более крупных систем, потому что мы не сможем точно охарактеризовать рабочие параметры, создаваемые нейронными сетями на выходе, в соответствии с которыми остальные части системы будут работать — или не будут. Например, лекарства, отпускаемые по рецепту, снабжаются подробнейшей информацией о том, какие побочные эффекты могут быть связаны с их приемом, какие из них опасны, а какие — просто неприятны. А теперь представьте себе, что компания предоставляет полиции систему распознавания лиц на основе глубокого обучения и она не в состоянии объяснить полицейским, когда она будет справляться, а когда нет. Ведь, как мы уже видели, легко может обнаружиться, что она хорошо распознает представителей европеоидной расы в солнечные дни, но не справляется с лицами афроамериканцев в пасмурную погоду. Понять подобные нюансы можно только путем экспериментов, но полицейская работа — не тот случай, когда вы можете себе это позволить.

Еще одним следствием непрозрачности глубокого обучения является то, что решения, принимаемые нейронными сетями, не соответствуют естественным знаниям о том, как устроен мир. Не существует способа объяснить этим системам, что яблоки растут на деревьях, птицы тоже могут сидеть на деревьях, однако, когда яблоки отрываются от ветки, они падают вниз, а не взлетают вверх, как это обычно делают птицы. Если единственное, что вам нужно, — это распознать яблоко на фотографии, то всего этого можно не объяснять, но если вы хотите, чтобы система глубокого обучения интерпретировала то, что происходит в хитроумных устройствах наподобие «самодействующей салфетки» [19], придуманной американским карикатуристом Рубом Голдбергом, то знайте, что вы еще не сталкивались с настоящими трудностями!

Глубокое обучение чрезвычайно нестабильно и непредсказуемо. Как мы видели во вступительной главе, этот метод может приводить к идеальным решениям в одной ситуации и совершенной абракадабре в другой. «Глюки» искусственного интеллекта, подобные обнаружению на фотографии

несуществующего холодильника, не являются чем-то исключительным, специально отобранным нами, чтобы принизить возможности нейронных сетей, — они остаются постоянной проблемой и сейчас, спустя годы после того, как впервые стали достоянием общественности. По данным одного исследования, современные системы продолжают делать точно такие же ошибки в 7–17% случаев. Вот типичный пример: на фотографии изображены две улыбающиеся женщины, болтающие по мобильным телефонам, за ними находится ряд деревьев, находящихся не в фокусе; одна женщина стоит прямо перед камерой, а другая повернулась к снимающему так, что ее лицо видно только сбоку. Система распознавания образов интерпретировала эту фотографию так: «Женщина разговаривает по мобильному телефону, сидя на скамейке». Это удивительное описание представляет собой причудливую смесь точных деталей и откровенной чепухи, что, скорее всего, было вызвано статистическими причудами в обучающем наборе данных: одна из женщин исчезла в никуда, и оттуда же, очевидно, вдруг появилась скамейка. После этого нетрудно представить себе, что подобные механизмы могут заставить автоматизированного робота-охранника неправильно интерпретировать мобильный телефон как пистолет (и это еще ничего, как мы сейчас убедимся).

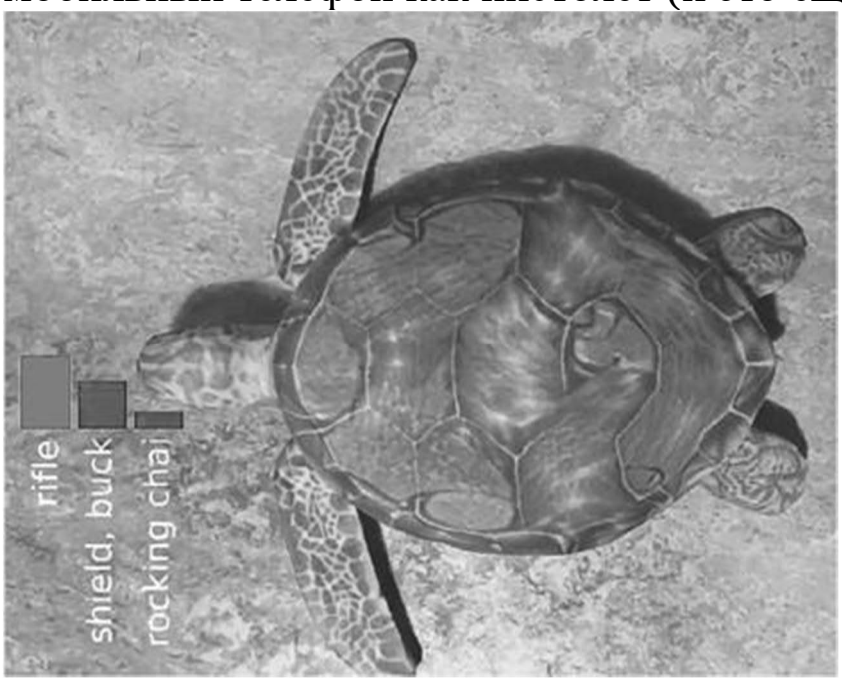


Рис. 3.4. Черепаха, ошибочно идентифицированная системой глубокого обучения как винтовка

Существуют без преувеличения десятки способов обмануть глубокие нейронные сети. Например, исследователи из Массачусетского технологического института создали трехмерное изображение морской черепахи, которую система глубокого обучения приняла за… винтовку (рис. 3.4). Когда экспериментаторы поместили черепаху под воду (туда, где эти черепахи живут и где винтовок обычно не бывает), система все равно продолжала настаивать на своем. В аналогичном примере группа исследователей нанесла немного пены на бейсбольный мяч, помещенный прямо в бейсбольную перчатку, и компьютер решил, что это чашка кофе эспрессо, с какой бы стороны ей ни показывали мяч, даже если он располагался прямо перед бейсбольной перчаткой (рис. 3.5).

Рис. 3.5. Бейсбольный мяч с нанесенной на него пеной, ошибочно идентифицированный искусственным интеллектом как чашка эспрессо
Еще одна команда ученых добавила едва заметные для человеческого глаза фрагменты случайных цветовых шумов по углам изображения свиньи-копилки и заставила тем самым нейронные сети идентифицировать картинку как фотографию «сумчатой куницы».

Четвертая научная группа добавила небольшие наклейки с психоделическим изображением тостера к объекту реального мира, в частности к обычному банану, и обманула систему глубокого обучения, заставив ее думать, будто вся композиция представляет собой лишь тостер, а не банан в сочетании с тостером (рис. 3.6, 3.7, 3.8). Если бы ваш ребенок не увидел на этой картинке банан, вы бы срочно отправили его к неврологу!

Существуют, наконец, такие способы изменить дорожный знак «стоп», чтобы система глубокого обучения неправильно определила его как знак ограничения скорости (рис. 3.9).

Рис. 3.6. Изображение «психоделического тостера»
Еще одна команда исследователей сравнила системы глубокого обучения с обычными людьми при решении двенадцати различных задач, в которых изображения были искажены теми или иными способами, например путем превращения цветных изображений в черно-белые, замены цветов, поворота изображения и т.д. В подавляющем большинстве случаев люди определяли объекты гораздо лучше, чем машины. Визуальный анализатор человека почти всегда надежен; глубокое обучение — увы, нет.

Рис. 3.7. Банан, правильно идентифицированный системой глубокого обучения

Рис. 3.8. Тот же банан с добавленной к нему наклейкой в виде тостера, ошибочно идентифицированный как «тостер без банана»

Некоторые примеры наглядно показывают, что системам глубокого обучения трудно распознавать самые обычные объекты, когда они находятся в необычном положении или ракурсе. Например, перевернутый набок школьный автобус нейронные сети определяют как снегоочистительную машину (рис. 3.10).

Рис. 3.9. Слегка измененный знак «стоп», неверно идентифицированный нейронной сетью как знак ограничения скорости

Ситуация становится еще более странной, когда мы переходим к вопросу понимания машиной человеческого языка. Стэндфордские компьютерные специалисты Робин Джиа и Перси Лян провели исследование систем, которые работают над тестами SQuAD, упомянутыми в главе 1, где системы глубокого обучения пытаются подчеркивать правильные ответы на вопросы в тексте. Системам глубокого обучения дали такой вот фрагмент текста.

Рис. 3.10. Школьный автобус, лежащий на боку. Из-за необычного положения компьютер ошибочно принимает его за снегоочистительную машину
Пейтон Мэннинг стал первым защитником в истории американского футбола, который позволил двум своим командам завоевать несколько Суперкубков. Он также является самым старым квотербеком, когда-либо игравшим в Суперкубке (последний раз выступил в возрасте 39 лет). Предыдущий рекорд принадлежал Джону Элуэю, который привел «Бронкос» к победе в Суперкубке-XXXIII в 38 лет и в настоящее время является исполнительным вице-президентом Денвера по вопросам американского футбола и его главным управляющим.
А вопрос был следующим:
Как зовут защитника, которому было 38 лет в Суперкубке-XXXIII?
Одна система глубокого обучения правильно подчеркнула имя «Джон Элуэй». Пока все нормально. Однако Джиа и Лян предъявили системе точно такой же текст, но в конце добавили еще одно предложение, совершенно не относящееся к делу:
Защитник Джефф Дин имел майку с номером 17 в чемпионате мира по футболу — XXXIV.
Когда после этого они повторили вопрос о 38-летнем защитнике из Суперкубка-XXXIII, система совершенно запуталась, приписав победу Джеффу Дину, а не Джону Элуэю. Иначе говоря, она смешала в кучу два предложения о двух разных чемпионатах, не понимая по-настоящему ни одно из них.

Другое исследование продемонстрировало, как легко обмануть системы вопросов и ответов, задавая лишь часть вопроса. Зависимость систем глубокого обучения от корреляций без использования принципов истинного понимания заставляет их отвечать такому шутнику невпопад, не дожидаясь конца вопроса. Например, если вы спросите систему «Сколько?», вы получите ответ «два»; если вы спросите «Какой вид спорта?», то получите ответ «теннис». Поиграйте с этими системами в течение нескольких минут, и у вас появится ощущение взаимодействия не с подлинным искусственным интеллектом, а с механическим попугаем.

В еще более странной форме эта проблема зачастую возникает в машинном переводе. Когда в Google Translate ввели якобы оригинальный текст, состоявший из многократного повторения английского слова «собака»: «dog dog dog dog dog dog dog dog dog dog dog dog dog dog dog dog dog dog dog», и попросили перевести эту бессмыслицу с языка йоруба [9] (и некоторых других языков) на английский, переводчик выдал следующий апокалиптический текст: Часы Судного дня — без трех минут двенадцать. Мы становимся свидетелями драматических событий в мире, обстоятельства которых прямо указывают на то, что нас неуклонно приближает к концу времени и Второму пришествию Иисуса [20].

Образно выражаясь, глубокое обучение не так уж и глубоко [10]. Важно понимать, что в термине «глубокое обучение» слово «глубокий» относится к числу слоев в нейронной сети и не означает ничего больше. Оно вовсе не подразумевает, что нейронные сети глубокого обучения извлекают из обучающих данных что-то позволяющее им видеть мир дальше или глубже людей или других компьютерных систем. Например, алгоритм, управляющий игровой системой DeepMind Atari, может играть в миллионы игр типа Breakout и так и не узнать, что такое весло (недавно это очень элегантно продемонстрировал стартап AI Vicarious). В Breakout игрок перемещает весло вперед и назад по горизонтальной линии. Если вы измените игру так, чтобы весло оказалось на несколько пикселей ближе к кирпичам (что вообще не мешало бы грести человеку), вся система DeepMind разваливается. Нечто похожее команда ученых из Беркли проделала с игрой Space Invaders: крошечные кусочки шумов, внесенные в игровое пространство, резко снизили производительность системы, показав тем самым, насколько поверхностной по своей сути оказывается ИИ-система, обучающаяся играм.

Похоже, что некоторые специалисты в этой области осознали наконец хотя бы эти проблемы. Так, профессор Монреальского университета Иошуа Бенжио, один из пионеров в области глубокого обучения, недавно признал, что «глубокие [нейронные] сети имеют тенденцию изучать поверхностные статистические закономерности в наборе данных, а не абстрактные концепции более высокого уровня». В одном из интервью, записанном во второй половине 2018 года, Джефф Хинтон вместе с Демисом Хассабисом, основателем DeepMind, также согласился с тем, что универсальный искусственный интеллект, очевидно, еще долгое время не станет реальностью.

Многих в глубоком обучении всерьез беспокоит «проблема головастика». Иными словами, существует некоторое количество простых, широко распространенных случаев, хорошо покрываемых массовыми обучающими данными (это можно представить себе как толстую часть головастика), но имеется также гораздо большее число редких случаев, по которым обучающих данных явно не хватает (длинный хвост ошибок). Легко научить систему часто встречающимся корреляциям, чтобы она правильно назвала вам фотографию, где группа молодежи играет в фрисби, потому что в сети имеется бесчисленное количество фотографий, помеченных таким образом. Однако гораздо сложнее

заставить глубоко обученную нейронную сеть объяснить вам, что изображено на следующей фотографии [11] (рис. 3.11).

Рис. 3.11. Знакомые объекты в необычных позах

Здесь практически все — собака, котята, игрушечные лошади и повозка — совершенно обычные объекты нашей жизни (как и интернета), но этой конкретной конфигурации элементов в обучающем наборе не было, и система понятия не имеет, что делать с таким изображением.

Так почему же глубокое обучение было настолько переоценено, невзирая на все эти проблемы? Дело в том, что оно по-настоящему эффективно для статистической аппроксимации с большими наборами данных и в нем есть определенная элегантность — всего одно несложное уравнение, которое, кажется, способно решить так много. Ну и, конечно, огромную роль сыграла также значительная коммерческая выгода от использования этих систем. Но теперь, оглядываясь назад, мы хорошо видим, что им все время чего-то недостает.

Пышный расцвет глубокого обучения может служить отличным примером различия между иллюзорным прогрессом и реальными возможностями искусственного интеллекта, о чем мы уже говорили во вступительной главе. Повторим еще раз: в решении некоторых задач глубокое обучение может быть очень успешным, но это не означает, что за его успехом стоит настоящий интеллект.

Глубокое обучение — это совершенно иной вид мышления, кардинально отличающийся от человеческого разума. В самом лучшем случае эти системы ведут себя подобно дурацкому волшебнику (вспомните джиннов из сказок

Шахерезады), который обладает чудесными способностями, но очень мало что знает про окружающий мир и людей. Сейчас легко найти эффективные системы глубокого обучения для маркировки изображений (их предоставляют Google, Microsoft, Amazon, IBM и другие производители), в том числе и коммерческие системы, а библиотека программного обеспечения нейронных сетей Google TensorFlow позволяет любому студенту, изучающему информатику, сделать какой-нибудь аналогичный продукт, не вкладывая в это личных средств. Столь же легко найти эффективные системы глубокого обучения для распознавания речи — на данный момент эти продукты пользуются значительным спросом. Однако распознавание речи и распознавание объектов — это еще не интеллект, а лишь мелкие фрагменты интеллекта. Для реального понимания мира нужны еще рассуждения, язык и аналогия, и пока что ни одна современная технология даже близко не подошла к овладению этими способностями. Например, у нас пока вообще нет систем искусственного интеллекта, которые могли бы надежно понимать юридические контракты, потому что одной лишь классификации по сходству здесь недостаточно. Чтобы понять юридический договор, вы должны уметь рассуждать о том, что в нем сказано (и что не сказано!), как различные его положения связаны с ранее утвержденными законами, и многое-многое другое; глубокое обучение не умеет делать ничего из этого. Эти системы нельзя даже попросить достоверно обобщить сюжеты старых фильмов для библиотеки компании Netflix.

Действительно, даже в той части познания, которая называется восприятием и которая ближе всего соответствует сильным сторонам глубокого обучения, нынешний прогресс является лишь очень фрагментарным: глубокое обучение может идентифицировать объекты, но оно не может понять связи и отношения между ними, именно поэтому нейронные сети можно так легко обмануть. В других областях, таких как понимание языка и повседневные рассуждения, глубокое обучение пока ни не йоту не приблизилось к человеческим возможностям.

Подавляющее большинство того, что написано о глубоком обучении в популярных СМИ, создает впечатление, что прогресс в одной из этих областей равносилен прогрессу во всех них. Например, *MIT Technology Review* в 2013 году включил глубокое обучение в ежегодный список революционных технологий и резюмировал его возможности следующим образом:

Обладая огромными вычислительными возможностями, машины теперь могут распознавать объекты и переводить речь в режиме реального времени. Искусственный интеллект наконец становится действительно умным.

Но в данном случае такая логика неприменима: ведь только то, что вы умеете распознавать отдельные слоги или способны отличить по внешнему облику бордер-колли от других пород собак, вовсе не означает, что вы обязательно умны. Не все когнитивные проблемы одинаковы по сложности и характеру. Проводить знак равенства между успехом в одном аспекте познания и успехом во всех областях познания означает поддаться эффекту иллюзорного прогресса.

История глубокого обучения ярко демонстрирует нам всю красоту и трагедию узкого искусственного интеллекта. Красив он потому, что при благоприятных обстоятельствах экономит нам массу усилий: вам не нужно тратить кучу времени на надоедливую работу по проектированию функций, и, несмотря на это, машина все равно выполнит большую часть того, что вам требуется сделать. Трагедия же здесь в том, что ничто и никогда не гарантирует вам того, что даже самая лучшая система узкого ИИ даст пользователю правильный ответ в реальном мире, когда вам это нужно больше всего, и даже того, что вы сможете настроить ее правильно, если она отказывается работать. Собственно говоря, работа с узким искусственным интеллектом часто гораздо больше похожа на искусство, чем на науку: вы пробуете разные подходы, и, если у вас достаточно данных, они, как правило, рано или поздно начинают работать. Но вы никогда не сможете предсказать это заранее с той точностью, с какой мы доказываем теоремы в геометрии. И ни одна существующая сейчас теория не может точно предсказать, какие задачи глубокое обучение может решить надежно, а какие — не может: это всякий раз приходится проверять эмпирически. Вы видите, что работает, а что нет, и затем многократно переделываете исходную систему и набор данных, пока не получите желаемые результаты. Иногда это легко, а иногда — сложно (рис. 3.12).

Глубокое обучение является очень ценным, даже неотъемлемым инструментом дальнейшего развития узкого искусственного интеллекта; мы ожидаем, что в будущем оно будет играть столь же важную роль, как и в наши дни, и что с его использованием люди изобретут множество творческих приложений, о которых мы сейчас даже не фантазируем. Но вероятнее всего, что при этом со временем оно станет лишь одним из множества компонентов в общем инструментарии искусственного интеллекта, а не автономным решением, противопоставленным остальным.

Рис. 3.12. Так вот как выглядит ваша система машинного обучения! (Рэндалл Манро (Randall Munroe), xkcd.com)

Честная правда о глубоком обучении состоит в том, что люди оказались безмерно очарованы одним конкретным набором алгоритмов, который, без сомнения, очень полезен, но вместе с тем — очень далек от подлинного интеллекта. Представьте себе, что мы изобрели автоматическую отвертку, работающую без участия человека: значит ли это, что теперь нам открылись пути к межзвездным путешествиям? Увы, мы останемся все так же далеки от них. Конечно, в открытом космосе нам очень пригодится автоматическая отвертка, но, чтобы оказаться там, потребуется изобрести нечто гораздо большее. Говоря все это, мы не подразумеваем, что системы глубокого обучения не могут делать вещи, которые внешне выглядят интеллектуальными, но мы хотим подчеркнуть, что глубокое обучение само по себе не обладает гибкостью и адаптивным потенциалом реального интеллекта. Вспомните ставшую крылатой формулировку Закона 31 [12] из «Кодекса законов космической инженерии» Акина: «Вы не сможете попасть на Луну, взбираясь последовательно на все более высокие деревья».

В оставшейся части книги мы опишем, что нам потребуется, чтобы, так сказать, добраться до Луны, — ну то есть сконструировать машины, которые смогут мыслить, рассуждать, говорить и читать так, как это делает самый обычный человек, — универсально и адаптабельно. Нам нужно не просто еще более глубокое обучение в смысле количества слоев в нейронной сети, а более глубокое понимание. Нам нужны системы, которые действительно могут рассуждать о сложном взаимодействии сущностей, связанных друг с другом причинно-следственными взаимоотношениями в постоянно меняющемся мире.

Чтобы понять, что мы подразумеваем под этим, следует углубиться в две наиболее сложные области искусственного интеллекта: чтение и роботизацию.

ГЛАВА 4

Если компьютеры такие умные, то почему они не могут читать, как люди?

САМАНТА: Итак, чем я могу вам помочь?
ТЕОДОР: Ну… мне кажется, что в моем компьютере некоторый беспорядок, вот и все.
САМАНТА: Вы не возражаете, если я посмотрю, что у вас на жестком диске?
ТЕОДОР: Гм… Ладно.
САМАНТА: Хорошо, давайте начнем с вашей электронной почты. У вас есть несколько тысяч писем по поводу некоего LA Weekly, но, похоже, вы не разбирали их много лет.
ТЕОДОР: А, ну да. Я их сохранял на всякий случай, ну, я подумал, может быть, в некоторых из них написано что-то забавное. Впрочем…
САМАНТА: Да, среди них есть несколько забавных. Я бы сказала, что нам нужно сохранить восемьдесят шесть, а остальные мы можем удалить.
«Она» (2013). Написано и поставлено сценаристом и режиссером Спайком Джонзом

Разве не было бы здорово, если бы машины могли понимать нас так же хорошо, как Саманта понимает Теодора? (Речь идет об операционной системе, озвученной Скарлетт Йоханссон в научно-фантастическом фильме Спайка Джонза «Она».) И если бы они могли мгновенно разобраться в наших электронных письмах, выбрать то, что нам нужно, и удалить все остальное?

Если бы мы могли наделить компьютеры только одним свойством, которое есть у нас, а у них нет, то лучше всего подошел бы дар понимания языка — не только для того, чтобы они могли лучше организовать нашу жизнь, но и для того, чтобы они сумели помочь человечеству в решении некоторых очень крупных проблем, таких как переработка и осмысление огромной научной литературы, которую люди по отдельности просто не в состоянии осилить.

В сфере медицины ежедневно публикуется около семи тысяч статей. Ни один врач или исследователь не может прочитать их все, и это является серьезным препятствием для дальнейшего прогресса во всей области. Открытие новых лекарств задерживается отчасти потому, что в литературе содержится много информации, которую никто не успевает прочитать. Новые методы лечения иногда не применяются из-за того, что у врачей нет времени их изучить и взять на вооружение. Программы искусственного интеллекта, которые могли бы автоматически обобщать необъятную медицинскую литературу, произвели бы настоящую революцию в здравоохранении.

Компьютеры, которые умели бы читать так же хорошо, как и аспиранты, но с вычислительной мощностью Google, способны были бы революционизировать и другие отрасли науки. Мы ожидаем прогресса в каждой ее области, от математики до климатологии и материаловедения. Преобразуются, конечно, не только точные и естественные науки. Историки и биографы смогли бы мгновенно узнать все, что было написано о неизвестном человеке, месте или событии. Авторы, пишущие в различных жанрах художественной литературы, автоматически проверяли бы несоответствия сюжета, логические пробелы и анахронизмы.

Даже гораздо более простые способности оказались бы чрезвычайно полезными. В современных iPhone есть специальная функция: когда вы получаете по электронной почте сообщение, в котором назначается встреча, вы можете нажать на виртуальную кнопку, и телефон добавит встречу в ваш календарь. Это действительно удобно, но лишь тогда, когда все работает правильно. Однако зачастую iPhone добавляет встречу не на тот день, который вы имели в виду, ошибочно ориентируясь, например, на еще какую-то дату, упомянутую в электронном письме. Если вы не отследите ошибку вовремя, это может обернуться серьезными деловыми проблемами.

Когда-нибудь, когда машины действительно научатся читать, наши потомки будут недоумевать, как мы обходились без «компьютерных секретарей», подобно тому как мы сейчас удивляемся — неужели предыдущие поколения обходились без электричества?

На ежегодной конференции TED [21] в начале 2018 года известный футурист и изобретатель Рэй Курцвейл, в настоящее время работающий в

Google, анонсировал свой последний проект Google Talk to Books, который обещал использовать понимание естественного языка, чтобы «обеспечить совершенно новый способ изучения книг». Журнал *Quartz* привычно расхвалил новое приложение как «мощный инноваторский поисковый инструмент Google, [который] ответит на любой вопрос, прочитав тысячи книг».

Как вы уже догадываетесь, на этом месте пора задать вопрос: «Что на самом деле умеет эта программа?» Ответ следующий: Google проиндексировал предложения в 100 000 книг, начиная от «Процветания в колледже» (Thriving at College) Алекса Чедиака и заканчивая «Программированием для чайников» (Beginning Programming for Dummies) Уоллеса Вонга и «Евангелием от Толкина» (The Gospel According to Tolkien) Ральфа Вуда, а затем разработал довольно эффективный метод кодирования значений предложений в виде векторных наборов [13]. Когда вы задаете вопрос, компьютер использует эти векторы, чтобы найти двадцать предложений в базе данных, имеющих векторы, наиболее похожие на заданные. Из этого описания очевидно, что «инновационная» система не имеет ни малейшего представления о том, что вы на самом деле спрашиваете.

Итак, уже зная информацию о входных данных в системе, легко видеть, что утверждение в статье *Quartz*, будто Talk to Books «ответит на любой вопрос», нельзя ни в коем случае воспринимать буквально. Конечно, «десять тысяч книг» — звучит впечатляюще, но на самом деле это лишь крошечная доля от более чем ста миллионов опубликованных изданий. Учитывая то, что мы поняли из предыдущей главы, насколько глубокое обучение опирается на слепые корреляции, а не на подлинное понимание, неудивительно, что реакция Talk to Books на многие запросы выглядит более чем сомнительной. Например, если задать системе вопрос о каких-то конкретных деталях из освоенных ею романов, ответ, скорее всего, будет достоверным. Но когда мы спросили у нее: «Где Гарри Поттер познакомился с Гермионой Грейнджер?» [14], то ни один из двадцати ответов не имел отношения к тексту книги «Гарри Поттер и философский камень» и сам вопрос о месте знакомства так и не получил ответа. Потом мы спросили: «Правы ли были союзники в том, что продолжили блокаду Германии после Первой мировой войны?» Talk to Books не нашел результатов, которые даже упоминали бы такую блокаду. Так что «ответ на любой вопрос» — это, мягко говоря, сильное преувеличение.

И если походящие ответы не были прямо изложены в предложениях из проиндексированного текста, продукт Google постоянно выдавал ошибки. Так, мы спросили его: «Какие семь крестражей упомянуты в романах о Гарри Поттере?» — и не получили ответа в виде перечня существ и артефактов. Вероятно, система не смогла это сделать потому, что ни одна из множества книг, посвященных Гарри Поттеру, не перечисляет крестражи единым списком. Когда мы спросили: «Кто был старейшим судьей Верховного суда [США] в 1980 году?» — система вообще не сработала, хотя любой из вас, будучи человеком, может открыть тот или иной онлайн-список судей Верховного суда (хотя бы в «Википедии») и за пару минут выяснить, что это был Уильям Бреннан. Talk to Books не смог ничего поделать с этим вопросом

именно потому, что в его базе данных нигде не было ни единого предложения, полностью излагающего ответ в явной форме: «Старейшим судьей Верховного суда в 1980 году был Уильям Бреннан». Ни в одном из 100 000 изданий этого предложения не оказалось, а сама система попросту не обладала способностью давать ответы на вопросы, хотя бы на йоту выходящие за пределы буквального содержания книг.

Однако наиболее значимой проблемой оказалось то, что в наших экспериментах мы получали совершенно разные ответы в зависимости от того, как ставился вопрос. Спросим Talk to Books, например, так: «Кто предал своего учителя за 30 сребреников?» Сами понимаете, что это эпизод из широко известной истории, однако из двадцати ответов только шесть содержали имя Иуды Искариота. (Любопытно, что девять — то есть число, большее в полтора раза — ответов хотя и были связаны с Библией, но упоминали гораздо менее известную и понятную историю о Михее Эфраимитском из Ветхого Завета: Книга Судей 17.) Но все пошло гораздо хуже, когда мы отклонились от точной формулировки про «серебряные монеты» или «сребреники» и задали системе чуть менее конкретный вопрос: «Кто предал своего учителя за 30 монет?» Здесь Иуда появился только в двух ответах, причем сама программа наиболее релевантной сочла следующую цитату, совершенно неуместную и неинформативную в данном случае: «Неизвестно, кто был учителем Цзинвана». А когда мы снова слегка перефразировали вопрос, на этот раз изменив «предал» на «продал» (получилось «Кто продал своего учителя за 30 монет?»), то Иуда окончательно исчез из результатов. Итак, чем дальше мы уходим от точного соответствия текста вопроса предполагаемым цитатам, тем больше нелепостей выдает Talk to Books.

Системы машинного чтения, о которых мы мечтаем, — если они появятся, — должны быть способны ответить практически на любой разумный вопрос о том, что они прочитают. Вдобавок они должны научиться объединять информацию, взятую сразу из нескольких документов. Наконец, их ответы будут состоять не только из готовых отрывков, но и содержать обобщение информации, будь то списки крестражей, которые никогда не появляются все вместе в одном и том же произведении, или что-то вроде содержательной выжимки из документации, какую вы бы ожидали от адвоката, собирающего прецеденты из большого числа судебных дел, или от ученого, который способен сформировать гипотезы, объясняющие наблюдения, опубликованные в нескольких статьях. Даже первоклассник может составить список всех хороших и плохих ребят, которые выступают в качестве героев в той или иной серии детских книг. Аналогично, хотя и на другом уровне, студент колледжа, пишущий курсовую работу, может объединить идеи из разных источников, сопоставить их и прийти к новым выводам. Именно так и должна действовать любая машина, которая претендует на то, что умеет читать.

Но прежде, чем мы сможем заставить машины синтезировать информацию, а не просто повторять ее наподобие попугая-всезнайки, нам потребуется

создать нечто хотя бы на порядок более простое: машины, которые могут надежно понимать элементарные тексты.

До этого дня нам еще далеко, однако некоторым людям ничто не мешает переживать о порабощении человечества искусственным интеллектом уже сейчас. Чтобы понять, почему это не просто нам не грозит, но даже хотя бы надежное чтение все еще остается довольно отдаленной перспективой для машинного разума, нам полезно будет рассмотреть в деталях, что именно требуется для понимания достаточно простого текста, например детской истории.

Предположим, вы читаете следующий отрывок из книги «Сын фермера» (Farmer Boy) — детской книги Лоры Инглз-Уайлдер, автора книги «Маленький домик в прерии» (Little House on the Prairie). Альманзо, девятилетний мальчик, находит на улице кошелек (тогда его называли «бумажник»), полный денег. Отец Альманзо догадывается, что бумажник (то есть кошелек) может принадлежать мистеру Томпсону, и Альманзо находит мистера Томпсона в одном из городских магазинов.

Альманзо повернулся к мистеру Томпсону и спросил: «Вы не теряли бумажник?»

Мистер Томпсон аж подпрыгнул. Он хлопнул рукой по карману и воскликнул в изумлении:

«Точно! И с ним полторы тысячи долларов! А что? Ты что-то знаешь об этом?»

«Это он?» — спросил Альманзо.

«Да, да, именно!» — воскликнул мистер Томпсон, хватая бумажник. Он открыл его и поспешно пересчитал деньги. Потом пересчитал банкноты еще раз…

Затем он вздохнул с облегчением и добавил: «Что ж, этот проклятый мальчишка ничего не стащил».

Хорошая система чтения должна быть в состоянии ответить хотя бы на такие вопросы:

- Почему мистер Томпсон хлопнул по карману рукой?
- Знал ли мистер Томпсон, что потерял свой кошелек, до того, как с ним заговорил Альманзо?
- Что имеет в виду Альманзо, когда спрашивает: «Это он?»
- Кто чуть было не потерял 1500 долларов?
- Все ли деньги остались в кошельке?

На все эти вопросы людям ответить очень легко, но до сих пор еще ни одна система искусственного интеллекта не может толком ничего поделать с подобными задачами. (Вспомните еще раз, как смутили Google Talk to Books слегка перефразированные запросы [22].)

По своей сути каждый из этих вопросов требует, чтобы читатель (будь то человек или кто-либо другой) следовал цепочке умозаключений по событиям, которые в истории лишь подразумеваются. Возьмите первый вопрос. До того как Альманзо заговорил с мистером Томпсоном, тот не знал, что потерял

кошелек, и предполагал, что он у него в кармане. Когда Альманзо спрашивает его, не потерял ли он кошелек, Томпсон понимает, что и вправду мог потерять свой бумажник. Именно для проверки этой возможности (потери бумажника) мистер Томпсон хлопает себя по карману. Поскольку кошелька не обнаруживается там, где он обычно хранится, мистер Томпсон приходит к выводу, что он потерял свой кошелек.

Когда дело доходит до сложных цепочек рассуждений, аналогичных приведенным выше, нынешний узкий искусственный интеллект оказывается в замешательстве. Подобные логические цепочки часто требуют, чтобы читатель заранее собрал внушительный набор базовых знаний о людях и предметах и, в более общем смысле, о том, как устроен мир. Ни одна из существующих ИИ-систем не имеет достаточно широкого фонда общих знаний, чтобы хорошо понимать события, происходящие даже в детских историях.

Возьмем некоторые знания, которые вы, вероятно, использовали прямо сейчас, «переваривая» историю про Альманзо и кошелек, — автоматически, даже не осознавая этого.

- Люди могут ронять вещи, не обратив на это внимания. В истории дан пример знаний о связи между событиями и психическими состояниями людей.
- Люди часто носят свои кошельки в кармане. Это пример знаний о том, как люди обычно используют определенные объекты.
- Люди часто носят деньги в своих кошельках, и деньги важны для них, потому что это позволяет им платить за вещи. Следовательно, мы имеем тут пример знаний о людях, их обычаях и экономике.
- Если люди предполагают, что что-то важное для них является правдой, но внезапно обнаруживают, что это может оказаться неправдой, то они нередко пытаются как можно быстрее это проверить. В отрывке есть пример знания о вещах, которые психологически важны для людей.
- Вы обычно можете понять, есть ли что-то в вашем кармане, ощупывая его снаружи. Этот пример говорит нам о том, как могут быть объединены различные типы знаний. Конкретно, знания о том, как взаимодействуют друг с другом различные предметы (руки, карманы, кошельки), объединяются со знаниями о том, как работают чувства.

Рассуждения, необходимые, чтобы ответить на другие вопросы, столь же насыщенны. Например, чтобы ответить на вопрос номер три: «Что имеет в виду Альманзо, когда спрашивает: "Это он?"» — читатель должен заранее знать кое-что о языке, а также о людях и предметах, заключив из этого, что наиболее правильной интерпретацией обоих местоимений «он» и «это» будет, очевидно, кошелек (или бумажник). Гораздо более тонким моментом при этом будет то, что указательное местоимение «это» относится к кошельку, который держит Альманзо, в то время как личное местоимение «он» относится к кошельку, который потерял мистер Томпсон. К счастью, в данном случае оба кошелька (тот, что держит Альманзо, и тот, что потерял мистер Томпсон) оказываются одним и тем же предметом.

Рис. 4.1. «Он сравнил ее с "бутылкой винтажного вина"»

Следовательно, для того чтобы справиться со столь простым отрывком, знания читателя о людях, предметах и языке должны быть и обширными, и глубокими, и гибкими. Если обстоятельства в двух похожих внешне текстах отличаются даже в мелких деталях, нам чаще всего приходится довольно основательно адаптировать к ним свое восприятие. Мы не должны ожидать поспешной реакции от мистера Томпсона, если бы Альманзо сказал, что он нашел кошелек своей бабушки. Мы считаем вполне правдоподобным, что мистер Томпсон мог потерять свой кошелек, не зная об этом, но мы были бы удивлены, если бы он не знал, что наверняка останется без бумажника в случае нападения грабителей с ножами. Никто из ученых пока даже близко не подошел к пониманию того, как заставить машину рассуждать столь гибко. Мы не думаем, что это вообще невозможно, и позже мы наметим некоторые шаги, которые следует предпринять в данном направлении, но на сегодняшний момент реальность такова, что требуемые нам новые достижения значительно превосходят те, что были достигнуты в разработке искусственного интеллекта за всю историю этой области. Система Google Talk to Books не имеет даже отдаленного сходства с тем, в чем мы нуждаемся, равно как и «секретари», разработанные их конкурентами из Microsoft и Alibaba, о которых мы упоминали в самом начале книги.

Существует фундаментальное несоответствие между тем, что машины умеют делать сейчас — а именно классифицировать объекты по категориям, — и человеческими рассуждениями и реальным пониманием мира, которые будут необходимы машинам, чтобы овладеть банальной, но критически важной способностью усваивать смысл прочитанного.

Практически все, что вы можете прочесть, — от научных монографий до рекламных плакатов — вызывает аналогичные проблемы. Детская книжка, которую мы обсуждали выше, ничем особенным в этом плане не выделяется. Приведем для сравнения небольшой отрывок из *The New York Times* за 25 апреля 2017 года.

Сегодня исполнилось бы 100 лет со дня рождения Эллы Фицджеральд. Житель Нью-Йорка Лорен Шенберг одно время играл на саксофоне бок о бок с «первой леди эстрады» в 1990 году, когда карьера певицы уже близилась к концу. Он сравнил Эллу с «бутылкой винтажного вина» (рис. 4.1)…

Любой человек (как некоторые современные ИИ-системы) может легко ответить на вопросы, взятые более или менее прямо из текста (например: «На каком инструменте играл Лорен Шенберг?»), однако большинство вопросов потребуют определенных умозаключений, которые абсолютно неподвластны нынешнему искусственному интеллекту. Вот лишь несколько:

- Была ли Элла Фицджеральд жива в 1990 году?
- Была ли она жива в 1960 году?
- Была ли она жива в 1860 году?
- Встречал ли Лорен Шенберг когда-либо Эллу Фицджеральд?
- Считает ли Шенберг, что Фицджеральд была алкогольным напитком?

Ответы на первый, второй и третий вопросы невозможны без понимания (и даже вычисления) того, что Элла родилась 25 апреля 1917 года, исходя из того, что 25 апреля 2017 года ей исполнилось (бы!) 100 лет, и без учета целого ряда общеизвестных естественных законов, в частности, таких.

- Люди живы во время своей профессиональной деятельности, поэтому Элла Фицджеральд была жива в 1990 году.
- Люди живы все время между их рождением и смертью, но никогда не бывают живыми до своего рождения или после своей смерти. Таким образом, певица не могла не быть жива в 1960 году, так же как ни в коем случае не могла быть жива в 1860 году.

Ответ на четвертый вопрос включает в себя рассуждение о том, что совместное воспроизведение музыки с кем-то обычно означает непосредственное (хотя бы очень поверхностное) взаимодействие с этим человеком, а также понимание того, что определение «первая леди эстрады» относится здесь именно к Элле Фицджеральд, хотя об этом и не сказано в явной форме.

Наконец, ответ на пятый вопрос требует немалого опыта в понимании того, что люди используют в качестве образных примеров, сравнивая между собой трудносопоставимые объекты, и владения конкретным знанием о том, что Элла Фицджеральд была человеком, а также знания закона природы, согласно которому люди никогда не бывают идентичны спиртным напиткам.

Выберите наугад любую статью в газете, или короткий рассказ, или многотомный роман, и вы тут же обнаружите нечто подобное. Опытные писатели никогда не рассказывают вам всего, они пишут лишь о том, что вам будет интересно прочитать, полагаясь на общие знания, чтобы заполнить пробелы. (Представьте себе, какими скучными были бы повести Уайлдер, если бы ей пришлось каждый раз объяснять вам, что люди держат свои кошельки в карманах и иногда пытаются обнаружить присутствие или отсутствие небольших физических объектов, похлопывая по карманам руками.)

На более ранних этапах разработки искусственного интеллекта одна группа исследователей действительно стремилась решить эти проблемы. Питер Норвиг, в настоящее время директор по исследованиям в Google, написал провокационную докторскую диссертацию о проблемах того, как заставить машины понимать человеческие рассказы. Еще более известен в этой области Роджер Шанк, который в период своей работы в Йельском университете обнародовал целую серию показательных примеров того, как машины могут использовать сценарии, чтобы понять, что происходит, например, когда человек идет в ресторан. Однако понимание нарративных текстов требует гораздо более сложных знаний и внушительного количества их форм, чем сценарии Шанка, поэтому проблема формулирования и сбора всех этих знаний оказывается невероятно сложной задачей. Со временем область, направленная на создание более реально мыслящих машин, заглохла, и ученые перешли к работе над другими, более доступными проблемами, такими как веб-поиск и механизмы выработки рекомендаций, ни одна из которых ни на миллиметр не приближала нас к универсальному искусственному интеллекту.

Конечно, веб-поиск изменил мир — это одна из самых больших историй успеха искусственного интеллекта. Платформы, подобные Google Search, Bing и ряду других, представляют собой мощнейшие и чрезвычайно полезные технические разработки, основанные на машинном разуме: за доли секунды они находят сходство и совпадения среди миллиардов веб-документов.

Самое удивительное здесь то, что, хотя все они работают на базе искусственного интеллекта, они не имеют почти ничего общего с тем типом универсального машинного чтения, к которому мы исходно стремились. Мы и сейчас хотим создать машины, которые способны понять, что они читают, однако поисковые системы не имеют к этому ни малейшего отношения.

Возьмем хотя бы Google Search. В алгоритме Google лежат две основные идеи. Одна из них существовала и до Google, другую же впервые применили именно создатели этой компании. Ни первая, ни вторая не зависят от того, насколько хорошо система понимает содержание документов. Более старая идея использовалась в программах поиска информации уже с начала 1960-х годов, задолго до появления Google и даже самого интернета: вы просто сравниваете слова в запросе со словами в документе. Хотите найти рецепты блюд с приправой из кардамона? Не проблема — просто найдите все сайты, содержащие слова «рецепт» и «кардамон». При этом вовсе не нужно понимать, что кардамон — это пряность, не требуется знать, чем она пахнет и каков ее вкус, не нужна история того, как ее добывают из стручков или какие мировые кухни ее чаще всего используют. Хотите найти инструкции по созданию самолетов? Просто введите в запрос сразу несколько слов, таких как «модель», «самолет» и «как», и вы получите множество полезных советов, хотя машина вообще не понимает, что такое самолет на самом деле, не говоря уже о том, что такое подъемная сила и сопротивление среды или по каким причинам люди летают на настоящих коммерческих самолетах, а не пытаются оседлать пластмассовую модель масштабом один к тысяче.

Вторая, куда более инновационная идея — это знаменитый алгоритм PageRank. Он состоит в том, что можно научить машину использовать «коллективную мудрость» интернета, оценивая, какие веб-страницы имеют более высокий приоритет, путем просмотра того, на какие из них уже существует много ссылок, в особенности ссылок с других страниц с высоким приоритетом. Этот прием позволил Google за короткое время подняться выше всех других веб-поисковых систем того времени. И тем не менее сопоставление слов и подсчеты ссылок, которые ведут с других страниц, не имеет ничего общего с настоящим пониманием текста.

Причина, по которой Google Search неплохо работает без необходимости реального чтения, заключается в том, что на выходе от него не требуется большой точности. Поисковая система не нуждается в понимании содержимого веб-документов, например, опирается ли какой-либо трактат о президентских полномочиях на левые или правые взгляды, — все это может сделать сам пользователь. Единственное, что Google Search нужно оценить корректно, — действительно ли данный документ относится к заданной теме. Как правило, можно получить довольно полное представление о предмете, которому посвящен документ, просто взглянув на некоторые слова и короткие фразы, содержащиеся в нем. Если там есть термины «президент» и «исполнительная привилегия», пользователь, вероятно, будет доволен, получив на него ссылку; если речь там идет о «семействе Кардашьян» [23], то, очевидно, такой документ к делу не относится. Если в документе упоминаются «Джордж», «Марта» и «Битва при Йорктауне», то Google Search может классифицировать такой документ как имеющий отношение к Джорджу Вашингтону, и для этого ему совершенно не обязательно знать что-либо о браке между людьми или о революционных войнах.

Конечно, поисковик Google не всегда настолько поверхностен. Иногда ему удается интерпретировать запросы и показывать пользователю полностью сформулированные ответы, а не просто длинные списки ссылок. Это уже немного ближе к настоящему чтению, но именно что немного, потому что обычно Google читает только запросы, а не сами документы. Если вы спросите: «Как называется столица штата Миссисипи?» [15] — Google правильно проанализирует ваш вопрос и найдет правильный ответ (Джексон) — но лишь в таблице, которая была составлена заранее. Если вы спросите: «Сколько стоит 1,36 евро в рупиях?» — синтаксический анализ справится и с этим, затем система обратится к другой таблице (на этот раз с актуальными курсами валют), задействует калькулятор и рассчитает, что 1,36 евро равны 110,14 индийской рупии, и выдаст буквально такой ответ.

Когда Google сообщает вам ответы подобного рода, то по большей части он вполне надежен (система, вероятно, делает это только тогда, когда ее индикаторы сообщают, что ответ будет правильным с высокой вероятностью). Но описанный механизм все еще очень далек от совершенства, и ошибки, которые он периодически делает, дают хорошие подсказки о том, что происходит у машины внутри. Например, в апреле 2018 года мы задали Google

Search такой вопрос: «Кто в настоящее время является членами Верховного суда [16] [США]?» — и получили очень неполный ответ «Джон Робертс», то есть система назвала всего одного члена из девяти. В качестве бонуса поисковик предоставил список еще из семи судей, которых «также ищут пользователи»: Энтони Кеннеди, Самуэль Алито, Кларенс Томас, Стефен Брейер, Рут Бадер Гинзбург и Антонин Скалия. Все эти люди, конечно, бывали в свое время членами Верховного суда, но Антонин Скалия к тому моменту уже оказалась покойной. Преемник Скалии Нил Горсач и недавно назначенные члены Елена Каган и Соня Сотомайор в списке Google отсутствовали. Похоже, что Google вообще не заметил важный нюанс запроса — «в настоящее время».

Возвращаясь к обсуждению вопроса о синтетическом восприятии, повторим, что совершенная система машинного чтения должна была бы скомпилировать свой ответ, прочитав Google News и обновив свой внутренний список членов Верховного суда согласно последним изменениям. В качестве компромиссной идеи можно было бы научить ее консультироваться с «Википедией» (которую регулярно обновляют люди) и выбрать действующих судей из статьи с соответствующим содержанием. Однако Google Search, похоже, ничего этого не умеет. По сути, мы опять имеем дело просто с выдачей по запросу самых частых статистических закономерностей (действительно, в поисковых запросах судьи Алито и Скалия встречаются особенно часто), а не с подлинным чтением и пониманием прочитанного.

В качестве другого примера рассмотрим еще один наш запрос: «Когда был построен первый мост?» [17] Поисковая система Google показала в верхней (самой релевантной) части ответов следующее:

В настоящее время широко используются железные и стальные мосты, которые проложены через большинство крупных мировых рек. На снимке показан первый в мире железный мост. Он был построен в Телфорде в 1779 году Абрахамом Дарби III и стал первым крупным сооружением в истории, построенным из железа.

В этом тексте слова «первый» и «мост» соответствуют нашему запросу, однако первый из когда-либо построенных мостов вовсе не был железным, а словосочетание «первый железный мост» не идентично по смыслу словосочетанию «первый мост» — образно говоря, Google куда-то подевал тысячи лет человеческой истории [18]. Приходится признать, что спустя более десяти лет после того, как был запущен поисковик Google, случаи, когда эта система читает вопрос и дает на него прямой ответ, все еще остаются в ничтожном меньшинстве. Когда вы получаете вместо ответов ссылки, это, как правило, означает, что Google Search полагается лишь на такие вещи, как ключевые слова и подсчет ссылок, а не на подлинное понимание вопросов и контента.

Прогрессивные компании, такие как Google и Amazon, конечно, постоянно совершенствуют свои продукты. Совершенно несложно написать вручную отдельную программу для правильного составления актуального списка членов Верховного суда, поэтому небольшие и постепенные улучшения будут

продолжаться. Однако и на горизонте нет общего решения всех тех проблем, которые мы здесь поднимаем.

Несколько лет назад мы заметили на Facebook хитроумный пост, ставший мемом. Это была фотография Барака Обамы с надписью: «В прошлом году вы сказали нам, что вам 50 лет, а теперь вы говорите, что вам 51 год. Как это понимать, Барак Обама?» Если вы человек, вы легко поймете юмор этого поста. Как нетрудно догадаться, два разных утверждения, произнесенные в разное время, могут оказаться чистой правдой. Но если вы машина, выполняющая лишь чуть большее, чем просто поиск по ключевым словам, вы неизбежно запутаетесь.

А как насчет речевых «виртуальных помощников», таких как Siri, Cortana, Google Assistant и Alexa? С одной стороны, они часто действуют, а не просто предоставляют вам списки ссылок, — и это хорошо. В отличие от Google Search, они с самого начала разрабатывались таким образом, чтобы интерпретировать пользовательские запросы не как наборы случайных ключевых слов, а как реальные вопросы. Но и через несколько лет после введения эти системы работают во многом на авось, оставаясь эффективными в одних областях и очень слабыми — в других. Например, все они довольно хорошо разбираются в вопросах, касающихся конкретных фактов, типа «Кто выиграл [бейсбольную] Мировую серию в 1957 году?»; у каждого из помощников есть куча сильных ноу-хау. Так, Google Assistant хорошо умеет давать уместные ответы и покупать билеты в кино, Siri тоже умеет давать адекватные ответы и бронировать товары в интернет-магазинах. Их сестра Alexa сносно разбирается в математике, неплохо умеет рассказывать заранее записанные анекдоты и (что неудивительно) хорошо заказывает покупки на Amazon.

Но за пределами конкретных хорошо проработанных областей не столь очевидно, чего ожидать от этих систем. Не так давно писательница Мона Бушнелл попыталась спросить у всех четырех названных выше программ, как добраться до ближайшего аэропорта. Помощник Google дал ей в ответ список турагентов; Siri — наводку на базу гидропланов; Cortana — список сайтов по продаже авиабилетов (Expedia и др.). В недавнем эксперименте, проведенном одним из нас [19], Alexa набрала 100% по таким вопросам, как «Дональд Трамп — человек?», «Является ли Audi транспортным средством?» и «Является ли Edsel транспортным средством?» Однако на других вопросах она полностью провалилась, например на таких: «Может ли Audi ездить на газе?», «Может ли Audi доехать из Нью-Йорка в Калифорнию?» и «Является ли акула транспортным средством?»

Вот еще один пример. Недавно Гэри разместил в твиттере скриншот с телефона, владелец которого пытался выяснить у Siri, где находится «ближайший ресторан быстрого питания, который не является "Макдональдсом"». Система с готовностью выдала список из трех близлежащих ресторанов, все они относились к фастфуду, но все три

принадлежали к сети «Макдональдс». Телефонный помощник, так сказать, вообще не признает слово «нет» (рис. 4.2).

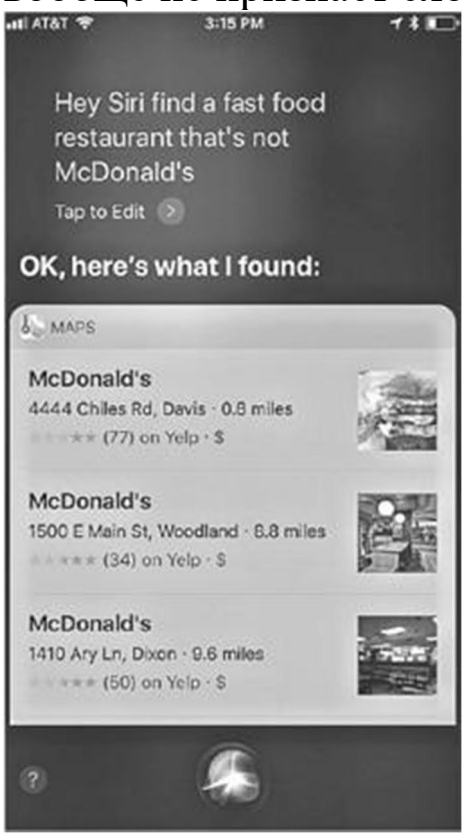


Рис. 4.2. Недоразумение с запросом к Siri: «Найди мне ближайшие рестораны быстрого питания, которые бы не были "Макдональдсом"»

Рис. 4.3. Недоразумение с вопросом к WolframAlpha: «Как далеко находится граница Мексики от Сан-Диего?»

Система WolframAlpha, широко разрекламированная еще в 2009 году в качестве «первого в мире механизма компьютеризированного знания» [20], на поверку ничуть не лучше. Хотя Alpha действительно располагает огромными базами данных всех видов научной, технологической, математической, учетной и социологической информации, встроенными прямо в систему, а также набором методов, позволяющих использовать эту информацию для

ответов на вопросы, ее способность использовать всю эту информацию пока остается очень фрагментарной.

Сильная сторона WolframAlpha — в ее умении отвечать на математические запросы типа «Каков вес кубического фунта золота?», «Как далеко находится Билокси, штат Миссисипи, от Калькутты?» или, скажем, «Каков объем икосаэдра с длиной ребра 2,3 м?» Ответы помощника здесь абсолютно точны: «547 кг», «8781 миля» и «26,5 м^3» соответственно.

Но границ понимания пользовательских запросов, присущих Alpha, достичь совсем нетрудно [21]. Если вы спросите: «Как далеко находится граница Мексики от Сан-Диего?» — то получите в ответ «1144 мили», что совершенно неправильно (рис. 4.3). Дело в том, что на этот раз программа игнорирует слово «граница» и вместо этого показывает вам расстояние от Сан-Диего до географического центра Мексики. Если вы слегка перефразируете вопрос об объеме икосаэдра, заменив слова «с длиной ребра 2,3 м» словами «длина ребра которого 2,3 м», WolframAlpha уже не понимает, что вопрос касается объема, и все, что вы получаете на выходе, — это общая информация о том, что икосаэдры имеют 30 ребер, 20 вершин и 12 граней, без какого-либо упоминания объема. Если обратиться к недавнему примеру с газетной статьей, то да, Alpha может точно сказать вам, когда родилась Элла Фицджеральд и когда она умерла, но если вы спросите: «Была ли Элла Фицджеральд жива в 1960 году?» — система неверно истолковывает ваш вопрос как «Жива ли Элла Фицджеральд?» и отвечает «нет».

Мы уже слышим возражение читателей: подождите, а как же Watson IBM, который так хорошо отвечал на вопросы, играя в «Jeopardy!», что победил двух людей-чемпионов? Все это правда, но, к сожалению, из этого не следует, что интеллектуальный продукт IBM действительно умен. Дело в том, что для почти 95% вопросов в «Jeopardy!» правильные ответы представляют собой названия тех или иных страниц «Википедии». Победа в таких викторинах зачастую целиком сводится к поиску нужной статьи в интернете. Однако от интеллектуального поиска информации до системы, которая может по-настоящему мыслить и рассуждать, еще очень и очень далеко. Обратите внимание, что до сих пор IBM даже не превратила Watson в надежного виртуального помощника. Когда мы недавно искали информацию о развитии системы на веб-странице IBM [22], то все, что мы смогли найти, — это давно устаревшая демонстрация Watson Assistant, которая была ориентирована исключительно на автомобильные симуляторы и ни в коем случае не могла сравниться с более универсальными предложениями от Apple, Google, Microsoft или Amazon.

Виртуальные помощники, подобные Siri и Alexa, безусловно, становятся все более полезными, но им предстоит еще долгий путь, чтобы стать чем-то большим, чем обычный портативный справочник. И, что очень важно, во всех них, так же как и в Google Search, мы до сих пор видим очень мало синтеза информации. Насколько мы можем судить, здесь все еще невозможен гибкий сбор информации из нескольких источников или хотя бы из одного, но по нескольким предложениям, как это умеют делать люди, читая повесть про

Альманзо и газетную статью об Элле Фицджеральд. Правда, на сегодняшний момент ситуация такова, что ни одна современная система искусственного интеллекта не может повторить того, что вы, читатель, делали в обоих этих случаях, объединяя ряд предложений в единый нарратив и не только понимая все, что было сказано, но и вызывая из памяти то, что сказано не было. Если сегодня вы действительно все это умеете, то вы — человек, а не машина. Когда-нибудь, вероятно, мы сможем попросить Alexa сравнить президентский репортаж *Wall Street Journal* с аналогичным репортажем из *The Washington Post* или спросить, не пропустил ли ваш семейный врач что-либо в ваших текущих анализах, но сейчас это просто фантазия. Пока что вам лучше поговорить с Alexa о погоде.

С чем же мы в результате остались? Со «сборной солянкой» из виртуальных помощников, в чем-то полезных, но никогда не бывающих полностью надежными. Ни одна из этих систем не может делать то, что мы, люди, делаем всякий раз, когда открываем книгу. Через шесть десятилетий с начала истории искусственного интеллекта компьютеры все еще по большому счету неграмотны.

Глубокое обучение не решит эту проблему, равно как и тесно связанная с ним методика сквозного обучения, при которой искусственный интеллект обучается преобразовывать входные данные непосредственно в выходные, без каких-либо промежуточных подсистем. Например, традиционный алгоритм беспилотного вождения автомобиля разбивает элементы этого процесса на подсистемы, в частности — восприятие, прогнозирование и принятие решений (зачастую с использованием глубокого обучения в качестве одной из структур в некоторых подсистемах). Сквозное обучение в этой же задаче обойдется без подсистем и вместо этого создает систему вождения автомобиля, которая в качестве входных данных принимает изображения с камеры, а в качестве выходных данных возвращает настройки ускорения и поворотов руля — без каких-либо промежуточных подсистем для определения того, где находятся различные объекты и как они движутся, каких видов действия или бездействия можно ожидать от других водителей и т.д.

Когда этот подход действительно работает, он может оказаться очень эффективным, а главное — более простым для реализации, чем соответствующие структурированные алгоритмы. Системы сквозного обучения часто требуют от разработчиков очень небольших затрат времени и усилий по сравнению с многокомпонентными вариантами. Некоторые из них уже сейчас относятся к лучшим автопилотным программам из числа доступных. Как подчеркивалось в одной из статей о состоянии приложения Google Translate, опубликованной в *New York Times Magazine*, сквозные системы глубокого обучения значительно улучшили качество машинного перевода, заменив собой более ранние подходы. В настоящее время, если вы хотите создать программу для перевода, скажем, с французского языка на английский и наоборот, вы должны начать со сбора огромного массива идентичных документов, которые существуют одновременно во французской и

в английской версиях, называемых «битекст» (bitext), например материалы канадского парламента, которые по закону должны публиковаться на обоих языках. Исходя из этих данных, Google Translate может автоматически узнавать соответствия между английскими словами и фразами и их французскими эквивалентами без каких-либо предварительных знаний о французском и английском языке или обучения специфике английской и французской грамматики. Даже скептики вроде нас сильно впечатлились такими возможностями.

Проблема, однако, в том, что этот подход все равно не универсален. Машинный перевод французского и английского языков оказался очень приспособленным для применения сквозного обучения, отчасти — из-за доступности большого количества сопряженных данных, а отчасти — из-за того, что между английскими и французскими словами существует более или менее четкое соответствие. В большинстве случаев правильное французское слово является одним из вариантов, которые вы найдете во французско-английском словаре, и большую часть времени соотношение между порядком слов в этих двух языках соответствует тем или иным стандартным паттернам. Тем не менее многие важнейшие аспекты понимания языка неподвластны сквозному глубокому обучению.

Ответы на вопросы гораздо более непредсказуемы в значительной мере потому, что слова в правильном ответе на вопрос не могут иметь очевидного отношения к словам в тексте. Между тем не существует, скажем, базы данных по вопросам и ответам такого размера, как парламентские документы, публикуемые сразу и на французском, и на английском. Даже если бы они были, совокупность вопросов и ответов настолько велика, что любая база данных была бы лишь крошечной выборкой из всех возможностей. Как мы выяснили ранее, это создает для систем глубокого обучения почти непреодолимые препятствия: чем дальше они вынуждены отклоняться от учебного набора данных, тем больше проблем у них возникает.

И, уж если быть откровенными до конца, даже в машинном переводе нейронные сети сквозного обучения все еще довольно ограничены по своим возможностям. Они часто (хоть и не всегда) хороши для понимания сути текста, но сопоставление слов, словосочетаний и предложений далеко не всегда корректно. Когда правильный перевод зависит от более глубокого понимания смысла фразы, системы тут же начинают сбоить. Если вы дадите Google Translate французское предложение [23] «Je mange un avocat pour le déjeuner», которое фактически означает «Я съедаю авокадо на обед», вы получите в ответ такой перевод: «Я ем адвокат на обед». Французское слово «avocat» означает и «авокадо», и «адвокат», а поскольку люди пишут о юристах гораздо чаще, чем об авокадо (особенно в материалах канадского парламента), Google Translate подставляет более частотное значение, теряя смысл предложения из-за особенностей статистики.

В своей превосходной статье в *The Atlantic* Дуглас Хофштадтер описал ограничения Google Translate такими словами:

Мы, люди, знаем все о парах, домах, собственности, гордости, соперничестве, ревности, неприкосновенности частной жизни и многих других нематериальных вещах, которые вместе складываются в причудливые картины жизни, подобные супружеской паре, имеющей полотенца с вышитыми на них метками «его» и «ее». Между тем Google Translate совсем не знаком с такими ситуациями. Повторяем. Google Translate не знаком с такими ситуациями. Точка. Он знаком исключительно со строками, состоящими из слов, состоящих из букв. Все дело лишь в сверхбыстрой обработке фрагментов текста, а не в мышлении, воображении, запоминании или понимании. Он даже не знает, что слова могут означать какие-то вещи.

Несмотря на весь достигнутый прогресс, большая часть письменных знаний в мире остается принципиально недоступной, даже если она оцифрована и выложена в сеть, — потому что она находится в форме, которую машины совсем не понимают. Электронные медицинские записи, например, до краев заполнены тем, что называется «неструктурированный текст», — это всевозможные заметки врачей, электронные письма, новостные статьи и документы, предназначенные для дальнейшей обработки, которые не вписываются в формат табличных данных. Настоящая система машинного считывания смогла бы действительно погрузиться в материал, изучая заметки врачей для получения важной информации, которая присутствует, скажем, в анализах крови и в справках для разрешения на работу. Однако проблема такого понимания текстов настолько далека от возможностей нынешнего искусственного интеллекта, что записи большинства медиков никогда не изучаются подробно. Инструменты искусственного интеллекта сейчас начинают применять для анализа рентгеновских снимков и МРТ; они могут сканировать изображения и отличать опухоли от здоровых тканей, но у нас пока нет способа автоматизировать другую, самую творческую часть работы настоящего рентгенолога, а именно — анализ изображений с учетом истории болезни пациентов.

Способность понимать неструктурированный текст пока является одним из самых узких мест в огромном диапазоне потенциальных коммерческих приложений искусственного интеллекта. Мы еще не можем автоматизировать процесс чтения юридических контрактов, научных статей или финансовых отчетов, потому что каждый из них состоит в значительной мере из такого текста, который ИИ вообще не может понять. Хотя современные инструменты автоматически извлекают некоторую базовую информацию даже из самых сложных текстов, большая часть их содержимого все равно остается лежать мертвым грузом. Все более изощренные алгоритмы сопоставления текста и подсчета ссылок оказывают в этом некоторую помощь, но они не дают нам реальной программы, которая действительно могла бы читать и понимать.

Ситуация с распознаванием устной речи (иногда это называют пониманием диалогов), само собой, находится в ничуть не лучшем положении. Еще бóльших проблем можно было бы ожидать у компьютеризированного помощника врача, который попытается перевести речь в медицинские заметки

(чтобы врач мог уделять больше времени пациентам и меньше печатать на своем ноутбуке). Так, собственно, и есть. Посмотрите на простой диалог, присланный нам доктором Виком Мохариром:
ДОКТОР: Вы чувствуете боль в груди при каких-либо нагрузках?
ПАЦИЕНТ: Ну, на прошлой неделе я косил газон на участке и почувствовал, что на меня словно свалился слон [указывая на грудь].

Для человека очевидно, что ответ на вопрос врача — «да», поскольку стрижка газона относится к категории силовых нагрузок, и еще мы понимаем, что пациент испытывал боль, поскольку нам известно, что слоны тяжелые, и, если сверху наваливается что-то тяжелое, это, естественно, причиняет боль. Столь же автоматически мы способны сделать вывод о том, что слово «свалился» используется не в буквальном, а в переносном смысле, видя пациента и учитывая, что падение на человека настоящего слона не могло бы для него обойтись без травм. Для машины, если только у нее в таблице сопоставления нет переносного значения словосочетания «на меня будто свалился слон», фраза, произнесенная пациентом, окажется чем-то вроде бреда о больших млекопитающих, вмешивающихся в работу на придомовом участке. Откуда берется вся эта абракадабра?

Глубокое обучение очень эффективно при анализе и обобщении корреляций, например, между изображениями или звуками и сопровождающими их метками. Но эти нейронные сети сталкиваются с непреодолимыми трудностями, когда дело доходит до понимания того, как лингвистические объекты, подобные предложениям в естественной речи, связаны с образующими их частями (например, словами и словосочетаниями). Почему? Дело в том, что у современных интеллектуальных систем отсутствует та составляющая, которую лингвисты называют композиционностью: это способ выстраивать значения предложений или фраз из значений более мелких структурных единиц. Например, в предложении «Луна находится на расстоянии 240 000 миль от Земли» («The Moon is 240,000 miles from the Earth») слово «Луна» означает один конкретный астрономический объект, слово «Земля» означает еще один космический объект, слово «миля» означает единицу расстояния, а число 240 000 означает количество (миль). Затем, исходя из синтаксических правил, определяемых в английском языке порядком слов, можно скомпоновать предложение таким образом, что 240 000 миль примет значение расстояния, а само предложение «Луна в 240 000 миль от Земли» констатирует, что от Луны до Земли (то есть между двумя небесными телами) — именно такое расстояние.

Удивительно, но системы глубокого обучения сами по себе не обладают умением справиться с композиционностью (хотя бы на примере вышеприведенного предложения): они просто содержат миллионы и миллионы отдельных фрагментов информации без какой-либо структуры, связывающей их воедино. Эти программы могут знать, что у собак есть хвосты и лапы, но они не знают, как они соотносятся с их поведением и образом жизни. Нейронные сети не идентифицируют собаку как животное, состоящее из

частей, подобных голове, хвосту и четырем лапам, они даже не знают, что такое животное, не говоря уже о том, что такое голова и как форма и размеры головы отличаются у живых существ от лягушек и собак до людей, почему головы такие разные в деталях, но всегда имеют отношение к телам животных. Наконец, глубокое обучение никак не помогает системе понять, что предложение типа «Луна находится на расстоянии 240 000 миль от Земли» содержит словосочетания, относящиеся к двум небесным телам и понятию длины.

Вот характерный пример этого. Мы предложили Google Translate перевести с английского языка на французский следующее предложение: «Электрик, которому мы звонили, чтобы починить телефон, работает по воскресеньям» («The electrician whom we called to fix the telephone works on Sundays») [24]. Компьютерный переводчик выдал такой ответ: «L'électricien que nous avons appelé pour réparer le téléphone fonctionne le dimanche». (Буквальный перевод с французского языка на русский звучит так: «Электрик, которому мы звонили, чтобы починить телефон, исправен по воскресеньям».) Если вы знаете французский язык, то поймете, что это не совсем правильный перевод. В частности, английский глагол to work имеет два основных перевода на французский язык: «travaille», что означает «работать», и «fonctionne», что означает «функционировать должным образом» или «быть исправным». Переводчик Google использовал слово fonctionne, а не travaille, не понимая (в отличие от человека), что «работает по воскресеньям» относится в данном контексте к электрику (а не к телефону) и что, если вы говорите о работающем человеке, нужно использовать именно глагол «travaille». Легко заметить, что в грамматическом смысле субъектом главного предложения (то есть подлежащим при сказуемом «работает») здесь является электрик, а не телефон. Смысл предложения в целом зависит от того, как соотносятся друг с другом его части, а Google Translate этого не понимает совсем. Успехи, демонстрируемые в ряде случаев современными автоматизированными переводчиками, нередко заставляют нас думать, что система понимает намного больше, чем на самом деле, но правда заключается в том (и это еще раз демонстрирует иллюзорное восприятие прогресса людьми), что в таких переводах очень мало реальной глубины понимания языка [24].

Не менее важная проблема, связанная с предыдущей, заключается в том, что глубокое обучение не способно по своей природе подключать к переводу, распознаванию и другим функциям никакие базовые знания о мире, природе, людях и т.д. (мы уже обсуждали это выше в главе 3). Если вы обучаете систему связывать изображения с маркерами, то для нейронной сети не имеет значения, как именно это делается. Пока машина выдает правильные результаты, никто не станет заботиться о внутренних деталях работы алгоритма, потому что все, что имеет значение, — это получить правильную метку для того или иного изображения. Выполнение задачи системой, таким образом, изолировано от большей части реальных знаний.

Язык почти никогда не работает таким примитивным образом. Практически каждое предложение, с которым мы сталкиваемся в повседневной речи и

чтении, требует, чтобы мы делали выводы, как то, что мы читаем или слышим, взаимосвязано с широким спектром базовых знаний. Глубокому обучению критически недостает умения усваивать такие знания, не говоря уже о том, чтобы делать из них выводы в контексте конкретных предложений.

И, наконец, системы глубокого обучения способны только на статистический перевод — по сути, это ничем не отличается от создания подписей к изображениям, просто вместо фотографий на входе будут предложения из разных языков. Однако чтение (как и восприятие речи на слух) — это динамический процесс. Когда вы используете статистические данные для перевода текста, который начинается с французского предложения «Je mange une pomme» (Я ем [одно, некое] яблоко), и система выдает вам перевод на английский (I eat an apple), это не значит, что она понимает смысл обоих предложений. Ей это, грубо говоря, и не нужно знать, если в обучающих двуязычных текстах у нее постоянно совпадали «I» и «je», «eat» и «mange», «an» и «une», «apple» и «pomme».

В большинстве случаев программы машинного перевода могут выдавать нечто более или менее осмысленное, просто обрабатывая одно слово (или предложение) за другим, но не понимая общего значения переводимого текста.

Когда же человек читает рассказ или эссе, он делает нечто совершенно иное. Цель работы нашего мозга состоит не в том, чтобы создать коллекцию статистически правдоподобных совпадений, — напротив, он стремится воссоздать мир, который придумал автор. Когда вы читаете отрывок из истории Альманзо, вы прежде всего делаете вывод, что в рассказе три главных героя (Альманзо, его отец и мистер Томпсон), затем вы начнете осознавать некоторые подробности об этих персонажах (Альманзо — мальчик, его отец — взрослый мужчина и т.д.) и реконструировать описанные события (Альманзо нашел кошелек, Альманзо спросил мистера Томпсона, принадлежит ли кошелек ему, и т.д.). Вы делаете все то же самое (и по большей части неосознанно) каждый раз не только тогда, когда читаете рассказы, но и когда осматриваетесь в новом помещении, наблюдаете за сюжетом фильма или слушаете новости. Вы сами догадываетесь, какие сущности присутствуют в том или ином контексте, каковы их отношения друг с другом и чего от них ожидать.

На языке когнитивной психологии то, что вы делаете, читая текст или слушая речь, — это создание когнитивной модели содержания текста (речи). Еще это можно назвать «формированием объектного файла» согласно терминологии Даниэля Канемана и покойной Энн Трисман, то есть записью о конкретном объекте и его свойствах, или комплексным пониманием сценария. Первое проще, второе — сложнее, однако оба этих действия привычны для любого человека.

Читая отрывок из книги «Сын фермера», вы постепенно формируете и обновляете (в вашем сознании) мысленное представление обо всех людях, объектах и событиях рассказа и об отношениях между ними. Здесь и Альманзо, и кошелек, и мистер Томпсон, и разговор Альманзо с мистером Томпсоном, и восклицание мистера Томпсона и его похлопывание себя по карману, и то, как

мистер Томпсон вырывает кошелек у Альманзо, и т.д. Только после того, как вы прочитали весь отрывок и создали когнитивную модель, вы сможете сделать все, что обычно следует за чтением: ответить на вопросы, перевести текст, скажем, на русский язык, обобщить, спародировать, проиллюстрировать или просто запомнить эту историю.

Система Google Translate как есть — типичный продукт узкого искусственного интеллекта — обходит стороной весь процесс построения и использования когнитивной модели; ей никогда не требуется рассуждать или отслеживать хоть что-то из того, что она делает. То, чему ее научили и что она делает неплохо — в пределах своих возможностей, — охватывает только мизерную долю того, что люди пишут и читают на самом деле. Она никогда не создает когнитивную модель содержания текста — просто потому, что не может. Бесполезно ожидать от системы глубокого обучения ответов на вопросы типа «Что бы произошло, если бы мистер Томпсон ощупал свой карман и обнаружил выпуклость там, где он ожидал найти свой кошелек?», потому что подобная задача вообще не входит в концепцию современного использования нейронных сетей.

Статистика не может заменить реального понимания. Проблема не только в том, что здесь или там появляются случайные ошибки, но и в том, что существует фундаментальное несоответствие между статистическим анализом, которого зачастую хватает для перевода несложных фраз без углубления в их смысл, и конструированием когнитивных моделей, которое обязательно потребовалось бы, если бы системы действительно «захотели» понимать то, что понимаем мы с вами.

Как ни удивительно, для глубокого обучения (в отличие от классических подходов к искусственному интеллекту) чрезвычайно трудным оказывается понимание простого слова «нет». Помните, как в запросе «Найти ресторан быстрого питания, который не является "Макдональдсом"» виртуальный помощник Siri полностью проигнорировал «не является»? Человек, написавший этот запрос, очевидно, хотел получить ответ типа Burger King на Элм-стрит, 321, Wendy's на Мэйн-стрит, 57 и IHOP на Спринг-стрит, 523. Но, к сожалению, в английских названиях Wendy's, Burger King или IHOP нет ничего хотя бы отдаленно связанного со словом «not», а с другой стороны, едва ли кто-то называет любой из этих ресторанов дословно «Not McDonald's». В результате «холодная» статистика здесь не поможет, точно так же как она не смогла бы связать между собой слова «king» и «queen» [25] из-за их абсолютного внешнего различия. Можно придумать различные уловки для решения этой конкретной проблемы (определение ресторанов), не выходя за пределы чисто статистического анализа, однако поиск универсального решения для всех ситуаций, когда системы глубокого обучения не воспринимают слово «нет», выходит далеко за рамки современных подходов.

В чем действительно нуждается рассматриваемая область, так это в традиционном фундаменте вычислительных операций, на основе которых строятся базы данных и классические формы искусственного интеллекта. В

данном случае речь идет о создании полного списка объектов (рестораны быстрого питания в определенном районе), а затем исключении из него элементов, принадлежащих другому списку (список ресторанов, работающих под вывеской «Макдональдс»).

Но глубокое обучение с самого начала строилось на том, чтобы избегать именно таких операций. Построение списков является базовым и абсолютно необходимым методом в создании большинства компьютерных программ и существует уже более пяти десятилетий (первый широко применявшийся язык программирования искусственного интеллекта, LISP, был в буквальном смысле построен на базе списков. Тем не менее данная операция полностью отсутствует в алгоритмах глубокого обучения. Неудивительно, что запросы, содержащие слово «не», современные нейронные сети воспринимают как попытку засунуть квадратные колышки в круглые отверстия.

Самое время теперь упомянуть и о проблемах двусмысленности. Человеческий язык буквально пропитан неоднозначностью. Подавляющее большинство слов в любых языках имеет по несколько значений. Английский глагол «to work» может означать в одних случаях «трудиться», в других случаях — «[правильно] функционировать»; существительное «bat» обозначает и летучую мышь, и деревянную биту, используемую в бейсболе. И это еще сравнительно простые случаи: перечисление всех значений таких английских слов, как «in» или «take», в хороших словарях растягивается на много колонок. Пожалуй, лишь очень специальные научные или технические термины можно считать условно однозначными.

Синтаксическая структура фраз часто тоже неоднозначна. Возьмем английское предложение «People can fish». Буквально оно переводится как «Люди могут рыбачить», и это, конечно, верно, однако не исчерпывает всех возможных трактовок. Слово «can» — это не только модальный глагол «мочь», но и существительное «[консервная] банка» и, соответственно, глагол «консервировать». Тогда все предложение приобретает смысл «Люди консервируют рыбу», причем слово «fish» переводится здесь именно существительным «рыба», в то время как в первом переводе оно соответствует глаголу «рыбачить». Фраза «People can fish» в последнем значении встречается, например, в романе Дж. Стейнбека «Консервный ряд» [26]. Местоимения, употребленные вместо существительных, часто вносят в предложения еще больше двусмысленности. Если вы говорите «Сэм не мог поднять Гарри, потому что он был слишком тяжелым», то в принципе это можно понимать двумя способами — что слишком тяжелым был Сэм или что слишком тяжелым был Гарри.

Что удивляет нас, людей, когда мы задумаемся над этим, так это то, что в 99% случаев мы даже не замечаем подобных двусмысленностей. Вместо того чтобы запутаться, мы быстро и без особых сознательных усилий находим для каждой фразы правильную интерпретацию (естественно, если она в принципе имеется) [27].

Предположим, вы услышали следующее: «Elsie tried to reach her aunt on the phone, but she didn't answer» («Элси пыталась дозвониться своей тете по телефону, но она не ответила»). Хотя это предложение не вполне однозначно, вы вряд ли ошибетесь, если предположите, что местоимение «она» относится к тете, а не к Элси. Затем, вряд ли вам когда-либо придет в голову спрашивать себя в аналогичных случаях, означает ли слово «tried» «пыталась» или «подверглась суду» (как, например, во фразе «The criminal court tried Abe Ginsburg for theft»: «Уголовный суд судил Эйба Гинзбурга за кражу»). Точно так же глагол «reach» в предложении про Элси вы едва ли воспримете в значении «физически достигнуть определенного места» (как в предложении «The boat reached the shore», то есть «Лодка достигла берега»). Словосочетание «on the phone» (то есть «по телефону», но дословно по-английски — «на телефоне») не вызовет у вас в голове видение Элси, карабкающейся по трубке или проводам в сторону тетиного дома (что, однако, вполне возможно в предложении «A bug is crawling on the phone», то есть «По телефону ползет жук»). Ничего из вышеописанного не произойдет: вы мгновенно, не задумываясь, найдете правильную интерпретацию всего предложения, не сосредотачиваясь на отдельных словах [28].

Теперь попробуйте заставить сделать все это современную «мыслящую машину». Конечно, в некоторых случаях ей действительно способна помочь простая статистика. Слово «tried» гораздо чаще означает в английском языке «попытался», нежели «подвергся суду». Точно так же словосочетание «on the phone» намного чаще означает, что телефон используется как средство коммуникации, чем когда что-то находится на поверхности телефона или движется по нему (хотя исключения и встречаются). Когда за глаголом «reach» следует имя или иное обозначение человека, а рядом в предложении есть слово «телефон», это с наибольшей вероятностью означает успешно установленную голосовую связь.

Однако в большинстве ситуаций статистические закономерности бессильны подсказать машине правильное решение. В результате искусственный интеллект пасует перед двусмысленностью, не видя способа распознать, что происходит на самом деле. Все в том же предложении «Элси пыталась дозвониться своей тете по телефону, но она не ответила» важнейшее значение имеет знание контекста и правильной аргументации. Базовые знания и контекст делают для читателя очевидным, что Элси не ответит на свой собственный телефонный звонок: логика диктует нам вывод, что это должна быть ее тетя. Важно, что никто не обучает нас, как делать такого рода выводы, потому что мы инстинктивно знаем, как правильно понять смысл написанного или услышанного, — это естественное следствие того, как мы интерпретируем мир. Глубокое обучение не может даже близко подойти к подобным проблемам.

К сожалению, сказанное относится не только к глубокому обучению. Классические методы искусственного интеллекта, которые возникли и распространились задолго до того, как нейронные сети стали настолько популярными, намного лучше показывают себя в решении проблем

композиционного восприятия и являются полезным инструментом для построения когнитивных моделей, но по своей производительности они до сих пор сильно отстают от глубокого обучения при использовании данных, в то время как язык — слишком сложная область, чтобы вручную закодировать все, что может потребоваться машине. Классические системы искусственного интеллекта часто используют шаблоны. Например, шаблон в виде «Место1 находится на Расстояние от Место2» ([PLACE1 is DISTANCE from PLACE2]) можно сопоставить с предложением «Луна находится на расстоянии 240 000 миль от Земли» и использовать для идентификации этого (и ему подобных) высказывания как предложения, определяющего расстояние между двумя физическими местами. Тем не менее каждый подобный шаблон должен быть закодирован вручную, и в ту минуту, когда появится новое предложение, которое хоть немного отличается от того, что было раньше (скажем, «Луна располагается примерно в 240 000 миль в стороне от Земли» или «Луна вращается вокруг Земли на расстоянии 240 000 миль»), система перестает понимать его смысл. В конце концов, шаблоны сами по себе тоже почти ничего не делают для того, чтобы помочь искусственному интеллекту разгадывать головоломки неоднозначности путем объединения знаний о языке со знанием об окружающем мире.

Итак, область понимания естественного языка сейчас фактически сидит на двух стульях (и оба очень неудобны). С одной стороны, существует глубокое обучение, которое великолепно себя проявляет в освоении конкретных навыков, но из рук вон плохо — в композиционном восприятии и построении когнитивных моделей. С другой стороны, классический подход к искусственному интеллекту включает способы лучше справляться с композиционностью и построением когнитивных моделей, но дает — в лучшем случае — посредственные показатели при обучении. И оба упускают самое главное, о чем мы рассуждали в этой главе: здравый смысл.

Вы не можете создавать надежные когнитивные модели сложных текстов, если недостаточно знаете о том, как устроен мир, о людях, местах, объектах и о том, как они взаимодействуют. Без этого подавляющая масса того, что мы с вами прочитали, выглядела бы для нас как бессмыслица. Настоящая причина, по которой компьютеры не могут читать, заключается именно в том, что им не хватает даже базового понимания того, как устроен мир.

К сожалению, обрести здравый смысл гораздо сложнее, чем пользоваться им, когда он у вас уже есть. Ниже мы убедимся, что потребность иметь машины, обладающие здравым смыслом, распространяется на очень большое число областей науки и практики. В этой главе мы увидели, насколько актуальна эта проблема для понимания языка. Между тем для робототехники она ничуть не менее насущна, и это еще очень мягко сказано.

ГЛАВА 5

А где же Рози?

Через десять лет Универсальные Роботы Россума будут производить столько пшеницы, столько материала, столько всего, что это все ничего не будет стоить.
Карел Чапек, писатель, который изобрел слово «робот» [29]

Мы все еще находимся на самой ранней стадии развития настоящих роботов. Я имею в виду роботов автономных, способных к общению и обучению, ответственных и действительно полезных для человека.
Мануэла Велозу, «The increasingly fascinating opportunity for human-robot-AI interaction: the Cobot mobile service robots», апрель 2018 года

Вы все еще переживаете о том, что сверхразумные роботы восстанут и атакуют людей? Оно того не стоит. Но если вам все-таки не по себе, вот шесть советов, что делать, если прямо завтра некий робот ни с того ни с сего захочет на вас напасть.

1. Закройте двери и для большей надежности заприте их. Современные роботы с трудом справляются с дверными ручками, иногда даже падают, когда пытаются их повернуть (рис. 5.1). Справедливости ради заметим, что мы наблюдали только одну демонстрацию того, как робот безуспешно старается нажать на одну конкретную дверную ручку, при определенном типе освещения, но искусственный интеллект, скорее всего, не умеет мыслить индуктивно. Во всяком случае, мы не знаем ни одной демонстрации, которая бы показала, что роботы могут справиться с целым набором дверных ручек разных форм и размеров, не говоря уже о различных условиях освещения. Не приходилось нам видеть и того, как робот открывает запертую дверь — даже с помощью ключа.
2. Еще не успокоились? Тогда покрасьте дверную ручку в черный цвет на черном фоне. Это значительно снизит шансы того, что робот вообще ее разглядит.
3. Ради большей предосторожности еще вы можете прикрепить к входной двери большой плакат со школьным автобусом или причудливо раскрашенным тостером (см. главу 3). Или наденьте футболку с изображением трогательного малыша. Робот будет полностью сбит с толку, подумает, что вы младенец, и оставит вас в покое.
4. Если это не сработает, поднимитесь на второй этаж и создайте на пути к вам препятствие, состоящее, например, из банановой кожуры и гвоздей. Едва ли хоть один робот осмелится переступить через настолько опасное заграждение.
5. Даже если самый отважный из роботов дойдет до лестницы и поднимется по ней, он не сможет запрыгнуть на стол, если только его не готовили специально для выполнения этой миссии. Для вас это едва ли будет проблемой, так что заберитесь на стол или залезьте на дерево и оттуда наберите 911.
6. В любом случае, вам не о чем тревожиться. Скоро прибудет служба спасения, но, вероятнее всего, еще раньше у робота просто разрядится батарея. Современные автономные роботы обычно работают на одной зарядке батареи всего несколько часов — а все потому, что компьютер внутри них поглощает огромное количество энергии (рис. 5.2).

Рис. 5.1. Робот падает на спину, пытаясь открыть дверь. (Источник: IEEE Spectrum)

Рис. 5.2. Отражение атаки роботов

Ладно, это мы так шутим. Возможно, когда-нибудь роботы смогут пробиваться через двери и запрыгивать на столы, но пока что все мыслящие машины, которые мы видим вокруг, не составляет труда обмануть. По крайней мере в ближайшей перспективе нам не нужно беспокоиться о возникновении Скайнета и даже о том, что роботы отнимут у нас работу.

Чего нам следует действительно опасаться, так это того, что революция роботов (нет-нет, мы имеем в виду революцию в области робототехники) останется возможной лишь в теории из-за необоснованного страха перед крайне маловероятными проблемами.

В фильмах роботов часто изображают то как героев, то, наоборот, как демонов. Отважный R2-D2 то и дело приходил в наш мир, чтобы спасти положение. Антигерой Терминатор, напротив, желал поубивать все человечество. Короче говоря, роботы хотят либо порадовать своих владельцев, либо уничтожить их. Однако в реальном мире у роботов обычно нет ни личностей, ни желаний. Они здесь не для того, чтобы извести наш род или захватить Землю, и, уж конечно, у них нет возможности защитить нас от Темного Властелина. Они даже особо не бибикают, в отличие от R2-D2. Вместо этого они по большей части незаметно трудятся на сборочных конвейерах, выполняя скучные задачи, которые люди никогда не захотят делать.

Надо сказать, что компании, производящие роботов, как правило, ненамного амбициознее своих детищ. Одна компания, с которой мы недавно общались, сосредоточена на создании роботов для раскопки фундаментов зданий, другая нацелена на автоматизацию сбора яблок. И то и другое направление можно охарактеризовать как отличные бизнес-идеи, но неужели это и есть те роботы, о которых мы мечтали в детстве? На самом деле мы хотим видеть рядом с собой Рози, ту самую милую, трудолюбивую горничную-робота из телешоу «Джетсоны» 1960-х годов. Рози — мастер на все руки, она может позаботиться обо всем, что есть в нашем доме: о растениях и кошках, о детях и о посуде. И вам никогда больше не нужно ничего чистить, мыть и убирать. Но, увы, мы не можем прибрести себе такую Рози или что-то ей подобное ни за какие деньги. Все, что мы имеем сейчас, — это слухи, будто Amazon планирует выпустить ходячую версию Alexa, которая будет самостоятельно перемещаться на колесах, но как же далеко ей еще до Рози!

Грустная правда состоит в том, что на данный момент истинным бестселлером и новейшим словом в мире робототехники предстает не автомобиль без водителя или упрощенная версия C-3PO, нет, — это пылесос Roomba в виде огромной хоккейной шайбы — без рук и ног, с ничтожно маленьким мозгом и полным отсутствием амбиций. Мы уже писали об этом чуде роботизации во вступительной главе, как и о том, насколько он не похож на Рози.

Конечно, у нас уже появились домашние роботы, смахивающие на домашних животных [25]; вскоре на рынок должны выйти «беспилотные чемоданы», которые следуют за своими владельцами по аэропортам и прилегающим территориям. Но вероятность того, что до 2025 года кто-то создаст робота, который будет готовить, убирать и менять детские подгузники, практически равна нулю. За пределами заводов и складов роботы все еще остаются редким курьезом [30].

Чего же нам не хватает, чтобы перейти от небесполезного в быту, но абсолютно однозадачного пылесоса Roomba до гуманоидного компаньона, похожего на C-3PO или Рози, — с полным набором умений и навыков, в силах которого освободить нас от почти любой домашней рутины, сэкономить нам кучу времени и буквально преобразить жизнь пожилых людей и инвалидов?

Для начала важно понять, что Roomba — это существо из совершенно другого мира роботов. В его основе лежит замечательная идея изобретателя Родни Брукса, вдохновленная размышлениями о том, что насекомые с их крошечным мозгом могут делать очень сложные вещи, например виртуозно летать. Собственно говоря, Roomba нет нужды быть особенно умным. Пылесосить пол — это сравнительно однообразная задача, и ее можно сделать вполне прилично (хотя и не идеально), имея совсем примитивный интеллект. Имея даже минимум компьютерного оборудования [26], вы вполне смогли бы создать робота, который делал бы что-то полезное и которого люди готовы были бы приобрести за соответствующую плату. Но такой подход оправдывает себя лишь до тех пор, пока вы оставляете функции робота ограниченными. Если речь идет просто о том, чтобы собирать основную массу пыли в обычной комнате, роботу-пылесосу достаточно просто двигаться вперед и назад, время от времени поворачиваясь вокруг своей оси и меняя направление при столкновении с препятствиями. Проходить раз за разом по одним и тем же местам на полу — не самый эффективный способ уборки. Но в большинстве случаев, если автомат не пропускает какую-то часть комнаты, куда можно попасть только через узкий проход, он выполняет свою работу вполне удовлетворительно.

Настоящий смысл революции в робототехнике состоит в том, чтобы выйти за рамки автоматических пылесосов и тому подобных устройств и создать роботов, которые могут выполнять широкий спектр физических задач, в том числе сложных и высококоординированных, которые мы, люди, постоянно выполняем в повседневной жизни, — от открытия банок с газированными напитками, откупоривания бутылок и вскрывания конвертов до прополки огородов, стрижки газонов и живых изгородей, упаковки подарков, росписи стен и сервировки стола.

Определенный прогресс на этом пути уже достигнут. Наш хороший друг и специалист-робототехник Мануэла Велозу сконструировала роботов, которые могут безопасно бродить по залам Университета Карнеги — Меллон. Мы видели демонстрации роботов, поднимающих грузы, значительно превосходящие их собственный вес. Автопилотируемые дроны (квадрокоптеры, которые по сути своей являются летающими роботами) уже научились делать поразительные вещи, например отслеживать спортсменов-бегунов, когда те соревнуются или тренируются на горных тропах, или (как это делает автономная камера Skydio) автоматически уклоняться от деревьев, попадающихся на маршруте.

Если вы потратите несколько часов на просмотр YouTube, вы увидите десятки демонстраций роботов, которые (по крайней мере на видео) кажутся куда более многофункциональными и умелыми, чем Roomba. Но ключевое слово здесь — «демо». Ничто из этого не готово отправиться в серийное производство. В 2016 году Илон Маск объявил о планах по созданию робота-дворецкого, но, насколько мы можем судить, прогресса в достижении этой цели так и не видно. Ничто коммерчески доступное сегодня не выглядит даже

намеком на технологический прорыв (возможно, за единственным исключением упомянутых выше квадрокоптеров, которые могут быть очень полезными для профессиональных съемочных групп), ведущий нас к созданию Рози. Даже очень хитроумно сконструированным дронам не требуется (и недоступно) подбирать вещи с земли, использовать их или подниматься по лестнице; собственно говоря, им не приходится делать ничего, кроме как летать повсюду и фотографировать. Предполагается, что скоро будет выпущена собака-робот (без головы!) под названием SpotMini, но еще неизвестно, сколько она будет стоить и для чего использоваться. Человекоподобный робот Atlas компании Boston Dynamics, ростом около пяти футов (185 см) и весом в 150 фунтов (60 кг), умеет выполнять сальто и паркур-трюки, но, скажите, вы видели в интернете это видео с паркуром? Выполнение трюка потребовало 21 дубля в специальном тщательно спроектированном помещении. Вы ведь не ожидаете, что этот робот сможет проделать все то же самое на детской площадке с вашими детьми?..

Несмотря на пока что скромные успехи, создано уже много весьма эффектных устройств. В дополнение к SpotMini и Atlas (которые способны впечатлить зрителя) роботы, производимые Boston Dynamics, включают модель WildCat — «самого быстрого четвероногого робота в мире», который может скакать со скоростью до 20 миль в час, а также BigDog — «первого продвинутого робота для пересеченной местности». Последняя модель имеет три фута (90 см) в высоту и весит 240 фунтов (96 кг), может передвигаться со скоростью до 7 миль в час, подниматься по склонам крутизной до 35°, ходить по обломкам, пользоваться грунтовыми пешеходными тропами, гулять по снегу и в воде и, наконец, переносить грузы весом до 100 фунтов (40 кг). И, конечно же, любой автопилотируемый автомобиль (реальный или проектируемый) — это просто робот в соответствующей «упаковке». И если на то пошло, то автоматизированные подводные лодки, наподобие Alvin, тоже являются роботами, не говоря уже о марсоходах. Некоторые исследователи, такие как Санг-Бэ Ким из Массачусетского технологического института, работают над увеличением функциональной гибкости оборудования и уже достигли впечатляющих успехов. Конечно, все это сейчас стоит слишком дорого, чтобы покупать для домашних нужд, но когда-нибудь цены снизятся, и роботы могут появиться в большинстве наших домов.

Пожалуй, самым важным случаем применения робототехники на сегодняшний день стала остановка и очистка ядерного реактора атомной электростанции Фукусима в Японии после его разрушения во время цунами 2011 года. Роботы, созданные компанией iRobot, были отправлены в реактор, чтобы оценить там положение вещей, и они продолжают использоваться для очистки и обслуживания территории АЭС. Хотя эти роботы в основном управлялись вручную (с помощью радиосвязи) операторами-людьми, находившимися на безопасном расстоянии, у них были некоторые важные (хотя и ограниченные) функции искусственного интеллекта. Так, они могли строить карты объекта, планировать оптимальные маршруты, вставать после

падения вниз по склону и возвращаться назад по собственным следам в случае потери связи с операторами.

Вы уже догадываетесь, что истинная проблема кроется в несовершенстве программного обеспечения. Автомобили без водителя могут двигаться сами, но это все еще очень опасно. Робот-собака SpotMini способен чуть ли не на подвиги, но до сих пор остается в основном телеуправляемым, то есть некий человек с джойстиком за сценой командует роботу, что делать. Безусловно, инженерам-механикам и электрикам, а также материаловедам, которые проектируют и изготавливают роботов, хватит работы на многие годы вперед: предстоит еще долгий путь, чтобы научиться создавать более емкие и легкие батареи, производить прочные и одновременно функциональные корпуса, да и просто делать будущих механических помощников более доступными по цене. Тем не менее все это, так сказать, цветочки, потому что пока мы не представляем себе, как заставить роботов делать то, что они делают, целиком автономно и вместе с тем — безопасно. Как же нам этого добиться?

В сериале «Звездный путь: Следующее поколение» дается очень простой ключ к решению всех проблем: у лейтенанта-командора имеется в распоряжении «позитронный мозг». Только вот, к сожалению, мы не очень понимаем, что это такое, как он работает, а главное — где бы мы могли его приобрести.

Между тем существует целый ряд вещей, которые потребуются практически любому разумному существу — роботу, человеку или животному, — которое захочет стать более интеллектуальным, чем пылесос Roomba. Для начала любое разумное существо должно уметь оценивать пять основных моментов: где оно находится, что происходит в мире вокруг него, что оно должно делать прямо сейчас, как оно должно реализовать свой план и что оно должно планировать делать в течение более длительного срока для достижения тех целей, которые у него имеются.

В менее продвинутом роботе, сфокусированном лишь на одной задаче, есть, как правило, возможность обойти (до некоторой степени) эти сложные когнитивные процессы. Самая первая модель Roomba не имела представления о том, где она находится, то есть она не составляла и не отслеживала карту помещения, в котором находилась. Едва ли больше она знала о том, движется она или нет и столкнулась ли она только что с каким-нибудь предметом. (Более поздние модели Roomba уже строят карты, отчасти для того, чтобы пылесосить более экономично и эффективно, отчасти — чтобы гарантировать, что они не пропустят те или иные части помещения в результате ошибок случайного поиска.) Вопрос о том, что делать сейчас, у Roomba вообще никогда не возникал: его единственная цель состояла в том, чтобы пылесосить.

Но на этом элегантная простота Roomba и заканчивается. В повседневной жизни у домашнего робота с более полноценными функциями возникнет намного больше вариантов поведения, следовательно, принятие решений станет более сложным (и основополагающим для успеха) процессом, а возможность его корректно осуществлять будет зависеть от гораздо более

глубокого понимания окружающего мира. Цели и планы могут быстро меняться. Владелец робота может дать ему команду разгрузить посудомоечную машину, но хороший домашний робот не ринется выполнять эту просьбу немедленно: он должен будет сначала приспособиться к изменяющимся обстоятельствам.

Если рядом с посудомоечной машиной упадет и разобьется тарелка, роботу может понадобиться придумать новый путь к посудомоечной машине (это называется изменением краткосрочных планов), но будет еще лучше, если он сможет понять, что теперь его ближайшим приоритетом становится уборка осколков и постановка посудомоечной машины в режим ожидания.

Если пища на плите начала подгорать, роботу тем более придется отложить разгрузку посудомоечной машины до тех пор, пока он не погасит конфорку. Бедный Roomba! Ему и в голову не придет даже на мгновение приостановить уборку — разразись вокруг него хоть ураган пятой категории. От Рози нам хотелось бы большей сообразительности.

Именно потому, что окружающий мир меняется буквально каждую секунду, фиксированные ответы на вопросы о целях, планах и окружающей обстановке нынешним мыслящим роботам не помогут. Полноценный автоматизированный домашний помощник вынужден будет постоянно анализировать ситуацию. «Где я?», «Каково мое текущие положение?», «Какие риски и возможности существуют в моей нынешней ситуации?», «Что мне делать в ближайшей и долгосрочной перспективе?», «Как мне выполнить свои планы?» [31] Каждый из этих вопросов должен рассматриваться в виде непрерывно повторяющихся циклов — роботизированного аналога так называемого цикла OODA, введенного легендарным пилотом военно-воздушных сил и военным стратегом Джоном Бойдом: наблюдать, ориентироваться, решать и действовать [32]. Хорошая новость заключается в том, что за долгие годы робототехника достаточно освоила отдельные части этого когнитивного цикла. Плохая же новость такова, что другие части (а их большинство) практически не испытали никакого прогресса. Давайте начнем с истории успеха — она касается локализации и управления движением.

Локализация своего положения в пространстве — значительно более сложная задача для машины, чем вы, возможно, думаете. Самым очевидным способом выглядит применение GPS, однако до недавнего времени GPS-регистраторы имели точность, не превышающую десяти футов (три метра), и они плохо воспринимают сигналы от спутников, находясь в помещении. Если бы наш гипотетический домашний робот должен был ориентироваться только таким образом, он запросто мог бы решить, что находится в ванной, в то время как в действительности он пребывал бы на кухне.

Военные и специализированные GPS могут быть гораздо более точными, но вряд ли они экономически доступны для потребительских роботов, а это значит, что бытовые автономные помощники не могут полагаться только на GPS. К счастью, роботы могут использовать множество подсказок, чтобы выяснить, где они находятся. К ним относится, например, навигационное

исчисление пути (отслеживание и подсчет числа оборотов колес робота, чтобы рассчитать, как далеко он сдвинулся от предыдущего местоположения). Естественно, машины могут использовать для этого и «зрение» (ванная комната внутри значительно отличается от кухни или лестничных пролетов), и, наконец, они способны ориентироваться по картам, которые сами же могут построить различными способами. За прошедшие годы инженеры-робототехники разработали целое направление методов, называемых «локализация с синхронным картированием» (*англ.* Simultaneous Localization And Mapping, SLAM), которые позволяют роботам составлять карту окружающего пространства и самим отслеживать, где они находятся на карте и куда направляются. При каждом смещении робот выполняет следующие операции.

1. Он задействует свои датчики, чтобы увидеть ту часть окружающей обстановки, которая просматривается из его текущего положения.
2. Затем он уточняет свое текущее положение и ориентацию путем сопоставления того, что он видит, с объектами на своей ментальной карте.
3. Далее он добавляет к своей ментальной карте все объекты (или части объектов), которые он не видел раньше (фактически — то, что до сих пор не было закартировано).
4. Наконец, он либо продвигается (обычно вперед), либо поворачивается и затем заново корректирует оценку своего нового положения и ориентации, принимая во внимание то, насколько он сдвинулся или на какой угол повернулся.

Хотя идеальной техники не существует, методы SLAM работают достаточно хорошо, так что вы можете спокойно оставить робота в любом случайном месте внутри здания (при условии, что у робота уже есть карта) и быть уверенным, что тот выяснит, где находится и как добраться туда, куда ему нужно (используя различное программное обеспечение). Это же позволяет роботам строить или улучшать карты в процессе исследования пространства. Таким образом, ориентацию (в понимании Бойда) можно считать более или менее решенной проблемой.

Другую область, где уже достигнут значительный прогресс, можно охарактеризовать как управление движением, сюда входит вся работа по контролю за механическими функциями робота: ходьбой, подъемом и перемещением предметов, вращением рук, поворотом головы, сложными типами локомоции наподобие подъема по лестнице, прыжков и т.п.

Для беспилотных автомобилей управление механической составляющей движения сводится к довольно простым вещам. Управление автомобилем включает ограниченное число базовых манипуляций, касающихся в основном педали газа, педали тормоза и рулевого колеса. Автономное транспортное средство может изменять свою скорость (включая начало движения и остановку) и управлять курсом (направлением движения). Больше, собственно говоря, ничего в нем нет. Если только автомобиль не является самолетом-трансформером, нет нужды контролировать даже вертикальную

составляющую движения в пространстве (вверх или вниз). Вычисление требуемых в каждый момент параметров рулевого колеса, тормозов и педали газа для заданной траектории — достаточно очевидная задача с точки зрения математики.

Ситуация становится намного сложнее, когда от четырехколесной машины мы переходим к гуманоидному роботу с несколькими конечностями и суставами, которые можно перемещать и поворачивать разными способами. То же самое относится, естественно, и к роботам, напоминающим внешне четвероногих существ или насекомых. Предположим, что на столе стоит чашка чая, а человекоподобный робот должен протянуть руку и взять ручку чашки двумя пальцами. Во-первых, роботу необходимо научиться манипулировать различными частями его руки и кисти так, чтобы они оказывались в нужном месте без проб и ошибок, приводящих к ударам по столу, столкновениям одной конечности с другой или к опрокидыванию чашки. Во-вторых, он должен знать, какое усилие приложить к ручке чашки, чтобы его было достаточно для твердого хвата, но чтобы при этом он оставался деликатным и не раздавил хрупкий фарфор. Одновременно с этим наш робот должен рассчитать путь к столу, на котором стоит чашка, с учетом того, где он находится прямо сейчас и какие препятствия имеются на его пути, а затем разработать сложный план (фактически это будет компьютерная минипрограмма или нейронная сеть, созданная специально для выполнения данной совокупности движений), который на входе укажет угол поворота и усилие суставов между частями тела, а также то, как они должны меняться со временем таким образом, чтобы содержимое чашки никогда не проливалось (для этого можно задействовать принцип обратной связи). Даже ради того, чтобы взять всего лишь одну чашку, машине наверняка придется включить в работу пять (а то и более) суставов: плечевой, локтевой, пястный и — как минимум — два пальца. Между всеми ними возникает множество сложно координируемых взаимодействий.

Несмотря на сложность описанной проблемы, в последние годы ее решение значительно продвинулось, что особенно заметно на примере Boston Dynamics — робототехнической компании, о которой мы упоминали ранее и которой руководит Марк Рэйберт, исследователь, обладающий по-настоящему глубокими знаниями в области локомоции человека и животных. Учитывая его опыт, неудивительно, что роботы Рэйберта, такие как BigDog и SpotMini, очень похожи в своих движениях на настоящих животных. Их программное обеспечение быстро и постоянно меняет усилия в исполнительных механизмах («мышцах») робота и интегрирует их с обратной связью, поступающей от датчиков машины, чтобы они могли динамически перепланировать движения в процессе выполнения своих задач (а не просто программировать все жестко заранее и надеяться, что ошибки не будет). Команда Рэйберта смогла автоматизировать множество процессов, которые раньше считались чрезвычайно сложными для роботов, например хождение по неровной поверхности, подъем по лестнице и даже сопротивление внешним силам, которые стремятся опрокинуть робота (и легко это делают в случае менее совершенных систем контроля устойчивости).

Многие лаборатории в таких научных центрах, как Беркли и Массачусетский технологический институт, также добились значительных успехов в управлении движениями. На YouTube есть видео лабораторных демонстраций роботов, которые открывают двери, поднимаются по лестнице, подбрасывают пиццу и складывают полотенца, хотя обычно все это снимается лишь в тщательно контролируемых условиях. Хотя у человека управление движениями остается намного более универсальным, особенно когда речь идет о мелких объектах, роботы постепенно обучаются все новым приемам.

Конечно, следует учитывать, что бóльшую часть того, что мы узнаем о текущем состоянии автоматизации управления движением в робототехнике, мы получаем из демонстрационных видеороликов, которые нередко вводят зрителя в заблуждение. Зачастую видео ускоряют перед размещением в сети, чтобы зритель подумал, что робот все делает со скоростью человека, тогда как на самом деле он выполняет за минуту или даже за час то, что человек способен сделать за несколько секунд. Иногда роботами и вовсе управляют люди-операторы, невидимые в кадре. Все подобные демонстрационные видеоролики являются, по сути, лишь подтверждением концепции и часто представляют только лучшие, редкие попытки, в то время как система в принципе не работает стабильно и, разумеется, не готова к введению в производство. Эти шоу доказывают, что роботы могут достаточно многое, если дать разработчикам время для программирования физических аспектов тех или иных задач. Но они далеко не всегда гарантируют, что соответствующие задачи будут выполняться стабильно, эффективно и, что наиболее важно, автономно — без этого робот не будет роботом. В идеале вам достаточно сказать роботу: «Убери мой дом», и после небольшой тренировки он должен не только пылесосить, но и протирать поверхности, мыть окна, приводить в порядок книжные полки, выкидывать рекламу из почтового ящика, складывать белье, выносить мусор и загружать посудомоечную машину. Демоверсии показывают, что теперь у нас есть оборудование, необходимое для выполнения целого ряда таких задач, следовательно, аппаратные аспекты уже не являются здесь ограничивающей стороной. Реальная проблема заключается в том, чтобы робот правильно интерпретировал ваши запросы (зачастую двусмысленные и расплывчатые), касающиеся целей и задач, и координировал свои планы в динамичном, постоянно меняющемся мире.

Как и в случае с искусственным интеллектом вообще, самой большой проблемой будет надежность. Практически в любой демонстрации робот выполняет свои задачи в самых идеальных условиях, которые только можно себе представить, а не в сложной и непредсказуемой среде. Если вы внимательно посмотрите видео о том, как роботы складывают полотенца, то обнаружите, что белье всегда будет иметь яркий цвет на фоне пустой комнаты с темными стенами. Этот прием существенно облегчает программному обеспечению робота задачу отличать полотенца от остальной обстановки помещения. В реальном же доме с тусклым светом и полотенцами, которые совершенно не обязательно контрастируют с фоном, может случиться настоящий погром, если робот примет за полотенца фрагменты стены или

другие предметы, обычно присутствующие в комнатах. Робот, пекущий блинчики, будет отлично работать в ресторане, где его можно разместить в комнате с глухими стенами без посторонних объектов, но представьте его делающим все то же самое в загроможденном вещами холостяцком логове, где вместо блинчиков на сковородке вполне могут оказаться стопки непрочитанных деловых бумаг.

Управление движением в реальном мире заключается не только в абстрактном управлении действиями конечностей, колес и других механических элементов. Речь идет об управлении этими действиями в контексте того, что воспринимает весь организм робота, и о том, как справляться с ситуациями, когда мир уже в следующую минуту оказывается совсем не таким, как только что ожидалось.

Речь, как мы видим, идет о ситуационной осведомленности — знании того, что может произойти в любой момент будущего. Не надвигается ли шторм? Может ли кастрюля на плите загореться, если я забуду выключить конфорку? Не упадет ли сейчас вот этот стул? (Родители с маленькими детьми, как правило, проявляют повышенную осведомленность именно о таких падающих стульях.) Один из аспектов ситуационной осведомленности связан с отслеживанием рисков, но это может быть также и поиск дополнительных возможностей или способов получения выгоды. Например, автомобиль без водителя может заметить, что появилась неожиданная возможность сократить путь или внезапно освободилось место для парковки. А домашний робот, который пытался прочистить канализацию, мог бы найти нестандартное применение для кулинарного шприца, с помощью которого пропитывают индейку маслом.

В хорошо контролируемом заводском помещении ситуационная осведомленность может выглядеть относительно легко решаемой проблемой, где главными вопросами будут разве что такие: «Нет ли здесь препятствия?» или «Работает ли конвейер?» Однако в домашних условиях многообразие ситуаций, а также риски, выгоды и возможности, им сопутствующие, могут оказаться значительно более сложными и куда менее управляемыми. Сидя в своей гостиной, вы можете иметь буквально сотни вариантов действий, а сама ситуация подвержена влиянию чуть ли не тысячи факторов. Вы можете встать, сходить в столовую или на кухню, включить телевизор, взять книгу или убраться на журнальном столике. Все что угодно из перечисленного выше может показаться обычным занятием в любой день, однако ваши планы резко поменяются, если вдруг сработает детектор дыма или на горизонте возникнут признаки приближающегося урагана.

При мониторинге того, что происходит вокруг каждую минуту, а также при оценке рисков и возможностей здесь и сейчас вы (как человек) постоянно сочетаете зрение со слухом, осязанием и обонянием (а может быть, даже и со вкусом). К этому добавляется ощущение того, где находится ваше собственное тело, вместе с осознанием присутствия рядом других людей (или иных существ), которые могут оказаться в той же комнате, а также понимание ваших

планов (что вы пытаетесь сделать в этот час, в этот день, в этом месяце) плюс множество других переменных (идет ли дождь, не оставил ли я окно открытым, не разгуливают ли по дому непрошеные насекомые или другие животные). В то время как заводские сборочные линии представляют собой замкнутые миры без лишних объектов, человеческие жилища обычно до крайности непредсказуемы, а следовательно, представляют собой среду, в которой роботам будет очень тяжело осваиваться.

Автомобили без водителя находятся на этой шкале трудности где-то посередине. В течение почти всего времени вождения оценка обстановки требует вычисления лишь двух параметров — направления движения и скорости. Реже возникают вопросы типа: «В какую сторону поворачивает дорога?», «Что за объекты находятся поблизости?», «Где они и как они движутся?» (это можно вычислить, сравнивая данные в различные моменты времени) и «Где я могу ехать, а где — нет?» (например, где находятся полосы движения или участки, на которых можно сделать повороты). Но все это может оказаться куда менее важным в тех случаях, когда надвигается торнадо или землетрясение, разгорается пожар, произошло столкновение на трассе, или даже если на шоссе выскакивает малыш в наряде для Хэллоуина.

Та часть ситуационной осведомленности, с которой довольно хорошо справляется нынешний искусственный интеллект, — это работа по выявлению объектов в некоторой среде, поскольку прямое распознавание объектов является сильной стороной глубокого обучения. Алгоритмы, используемые в нейронных сетях, уже могут с определенной точностью идентифицировать основные элементы во многих сценах — от столов и подушек в доме до автомобилей на дороге. Тем не менее даже при идентификации относительно простых предметов могут возникать серьезные проблемы. В частности, лишь немногие системы распознавания объектов настолько надежны, чтобы замечать изменения в освещении; кроме того, чем более загромождена комната, тем больше вероятность, что они запутаются даже в знакомых предметах. Кроме того, недостаточно лишь заметить, что где-то в поле зрения находится, скажем, пистолет; важно понять, что это именно — пистолет, нарисованный на висящей на стене картине (в этом случае его можно спокойно проигнорировать), или реальный объект на столе, или даже в чьих-то руках и направленный на кого-то. Более того, современные системы распознавания объектов очень далеки от понимания отношений между объектами: мышь в ловушке очень отличается от мыши рядом с ловушкой; мужчина верхом на лошади сильно отличается от того же мужчины, несущего лошадь на себе.

Однако классификация объектов, находящихся в поле зрения в определенном пространстве, — это даже меньше чем полдела. Реальная проблема ситуационной осведомленности состоит в том, чтобы понять, что все эти объекты в совокупности означают. Насколько нам известно, исследований по этой проблеме было совсем мало или даже вовсе не было, а сама она несравненно сложнее. Мы не знаем ни одного нынешнего алгоритма, который мог бы, например, увидеть две очень непохожие сцены, включающие наличие огня в гостиной, и абсолютно точно понять, что в одном случае это огонь в

камине, который дает восхитительное тепло в зимний день, а в другом случае речь идет о пожаре, который вам лучше потушить как можно быстрее (и/или позвонить в пожарную службу). Чтобы даже приблизиться к решению данной проблемы в рамках доминирующей сегодня парадигмы, очевидно, понадобится масса помеченных наборов данных для распознавания различных типов домов (деревянные, бетонные, композитные) и различных видов отопительных приборов (масляные, электрические и др.). Ни у кого из разработчиков искусственного интеллекта сейчас нет системы общего назначения для классификации видов открытого огня.

Изменчивая природа окружающего мира делает ситуационную осведомленность еще сложнее. Вы ведь не хотите смотреть на мир как на застывший снимок, вы желаете и можете видеть его скорее в формате разворачивающегося сюжета или фильма, где вы способны отличать устойчивые объекты от шатких и понимать, какие автомобили въезжают на парковочные места, а какие оттуда уезжают.

Еще сильнее все запутывает тот факт, что каждый робот и сам меняется (например, когда он маневрирует в некотором помещении) и в то же время является активным участником тех изменений, которые он привносит в окружающую обстановку. Это означает, что робот должен не только динамически воспринимать природу окружающего его мира, но и предсказывать последствия собственных действий. Опять-таки в заводском помещении, где все жестко контролируется, это взаимодействие можно отрегулировать довольно легко: роботу, скажем, достаточно оценить, надежно ли дверь прикреплена к автомобилю или она шатается. В открытой среде прогнозирование становится реальной проблемой: если я ищу банку с кофе, должен ли я открыть шкаф? Или я должен посмотреть в холодильнике? Должен ли я открыть банку с майонезом или она уже открыта? Если я не могу найти крышку для блендера, можно ли запускать блендер без нее? Или просто накрыть блендер тарелкой? И заводской цех становится запутанным местом в тот момент, когда где-то неожиданно откручивается болт. Даже Илон Маск согласился с тем, что первоначальные сложности в производстве Tesla Model 3 были вызваны «слишком обширной автоматизацией». Мы подозреваем, что в том случае большая часть проблемы заключалась вот в чем: место сборки автомобилей и сам этот процесс все время менялись, и роботы не могли поспевать за постоянными изменениями, поскольку их интеллектуальные программы оказались недостаточно гибкими.

Немалую часть этих проблем можно вскрыть, изучая мировой опыт применения робототехники, но все-таки метод проб и ошибок имеет естественные ограничения — нам никак нельзя допустить, чтобы в число невольных экспериментов искусственного интеллекта вошли попытки засунуть по ошибке в блендер кота. Чем больше надежных выводов вы можете сделать, не экспериментируя вообще, тем лучше. Собственно говоря, именно в таких ежеминутных и повседневных мысленных экспериментах и в прогнозировании последствий люди на много миль впереди всего, что мы когда-либо встречали

у машинного разума.

Возможно, еще более трудноразрешимая проблема заключается в том, чтобы научить машину выяснять, как лучше всего поступать в каждый конкретный момент. С точки зрения программирования это гораздо сложнее, чем можно было бы представить себе без учета особенностей искусственного интеллекта.

Чтобы лучше понять, с какими проблемами может столкнуться наш гипотетический домашний робот, давайте рассмотрим три сценария, типичных для машины, оказавшейся один на один с реальным миром.

Сценарий первый: Илон Маск устраивает большой вечерний прием и хочет, чтобы дворецкий-робот разносил гостям угощения. По большей части схема действий здесь очень проста: робот движется с подносами, на которых стоят напитки и закуски, а также забирает у гостей (или со столиков) пустые бокалы и тарелки; наконец, если гость попросит еще бокал, робот приносит конкретный напиток конкретному человеку. На первый взгляд, это может показаться вполне достижимым для современных технологий. В конце концов, уже несколько лет назад робототехническая компания Willow Garage (сейчас она закрылась) прототипировала оригинальную демоверсию гуманоидного робота PR2, способного доставать пиво из холодильника.

Но, как и в случае с автомобилями без водителя, настоящий успех заключается в правильном понимании возможных отклонений от простой теоретической схемы. Настоящие дома и настоящие гости сложны и непредсказуемы. Обстановка для «пивного» робота PR2 была тщательно сконструирована заранее. На полу не было ни собак, ни кошек, ни разбитых бутылок, ни детских игрушек. По словам нашего коллеги, даже продукты в холодильнике были специально расставлены таким образом, чтобы сделать пиво доступным в первую очередь [27]. Но в реальном мире в каждый момент времени любое количество вещей, больших и маленьких, может выйти из-под контроля. Если робот пойдет на кухню, чтобы взять бокал с вином, и внезапно обнаружит в напитке таракана, ему придется составить план, который никак не был предусмотрен исходным алгоритмом. Очевидно, роботу следует выбросить насекомое, промыть бокал и снова наполнить его — тем же самым вином. Или же робот-дворецкий может увидеть трещину в стенке бокала, и в этом случае будет самым надежным отправить бокал в мусор максимально безопасным способом. Но какова вероятность того, что хоть один программист в мире способен предвидеть любую из массы подобных случайностей, если самые лучшие специалисты в этой области, работающие над приложениями для iPhone, все еще не могут надежно автоматизировать даже процесс создания записи в календаре на основе текста из вашей электронной почты?

Список непредвиденных обстоятельств, по сути, бесконечен — вот она, ахиллесова пята узкого искусственного интеллекта. Если дворецкий-робот заметит, что на пол упал крекер, ему необходимо разработать план того, как подобрать этот кусочек и выбросить его, не мешая гостям, или же он должен сообразить, что возня с упавшим печеньем в комнате, полной народа, просто

не стоит хлопот, потому что это вызовет слишком много беспокойства. Для решения таких проблем (ничтожных с точки зрения человеческого мозга) просто невозможно придумать отдельный алгоритм для каждого случая. Ведь если робот увидит на полу не крекер, а дорогую сережку, то баланс приоритетов сразу меняется: спасать сережку нужно будет непременно, невзирая на возможные неудобства для присутствующих.

Большую часть времени робот не должен применять к людям силу. Но что, если подвыпивший парень направился, сам не зная куда, спиной вперед и не видит младенца, ползающего позади него? В этот момент робот-дворецкий должен вмешаться, возможно, даже схватить пьяного взрослого в охапку, чтобы защитить ребенка.

Ни в одном наборе обучающих данных вы не перечислите всех возможных ситуаций. Роботу-дворецкому придется самому рассуждать, предсказывать и предвидеть, и он вряд ли сможет плакаться весь вечер напролет людям-краудсорсерам, собирающим для него тренировочный материал, — всякий раз, когда ему предстоит принять решение, которого у него не обнаружилось в программе. Пережить праздничный вечер в особняке Илона Маска в должности дворецкого для мыслящей машины было бы серьезной проверкой на уровень когнитивных способностей.

Мы, конечно, не можем позволить себе иметь дворецких-роботов, по крайней мере до тех пор, пока их цена не снизится в миллионы раз. Но теперь давайте рассмотрим второй сценарий, гораздо менее легкомысленный: возьмем роботов-компаньонов для пожилых людей и инвалидов. Предположим, что мистер Блейк недавно ослеп и хотел бы, чтобы его компаньон-робот помог ему сходить в магазин за продуктами. Опять-таки, это гораздо легче сказать, чем сделать, потому что по дороге в магазин и обратно, как и в самом магазине, может случиться множество неожиданных вещей. Для начала существуют простейшие проблемы навигации. По пути в продуктовый магазин компаньон-робот должен будет помочь мистеру Блейку преодолевать всевозможные непредсказуемые препятствия.

Они могут столкнуться с бордюрами, лужами, выбоинами, полицией, пешеходами, погруженными в свои телефоны, и с детьми, которые носятся на скутерах и скейтбордах. В магазине нашей паре, скорее всего, придется протиснуться сквозь узкие проходы среди коробок и лавировать между столиками дегустаторов, которые слегка изменяют функциональную планировку продуктового магазина, не говоря уже о людях, занимающихся инвентаризацией, или уборщиках, моющих пол после того, как кто-то случайно уронил банку с вареньем. Роботу-компаньону придется вести Блейка через все это или обходить любые неожиданные предметы вокруг — это не считая проблем собственного ориентирования. Тем временем к Блейку могут внезапно обратиться различные люди: старый друг, незнакомец, желающий помочь, попрошайка, полицейский; к нему может пристать дружелюбная собака, недружелюбная собака, а то и грабитель — любого из участников подобных взаимодействий необходимо опознать и отнестись к нему соответственно. В магазине до каждой вещи нужно дотянуться и схватить ее

(причем по-разному для разных предметов: красный перец требует иного обращения, нежели коробки с кашами или стаканчики с мороженым). Потом их придется положить в корзину для покупок, не разбив при этом яйца и не заваливая бананы сверху банками с консервированным супом. Даже сама корзина должна быть для робота однозначно узнаваемым предметом, несмотря на то что корзины могут быть разными по форме и размеру в зависимости от магазина. Точно так же варьируют от одного магазина к другому способы оплаты и детали того, как продукты упаковываются в пакеты. Невозможно полностью предвидеть и запрограммировать наперед тысячи случайных обстоятельств, меняющихся день ото дня и специфических для разных торговых заведений.

В качестве третьего сценария рассмотрим что-то вроде ядерной катастрофы на АЭС «Фукусима». Представьте себе здание, которое частично разрушилось в результате землетрясения, и ядерный реактор вот-вот расплавится от неконтролируемой цепной реакции. Робот-спасатель, отправленный в кризисную зону, должен быстро разобраться, что он сможет безопасно сделать, а что не сможет. Сумеет ли он открыть или взломать дверь, пробиться сквозь стену или это грозит еще большими разрушениями? Может ли он безопасно подняться по лестнице, которая была сделана для людей, а не для роботов? Если спасательный робот найдет кого-то из персонала, что делать тогда? Человек может быть еще в состоянии уйти самостоятельно, если расчистить ему путь, но, возможно, он завален или зажат между обломков, и его нужно освободить; у человека в такой ситуации очень вероятны повреждения, при которых с ним потребуется обращаться крайне осторожно. Если обнаружено несколько человек, роботу наверняка придется классифицировать их состояние и решить, какие травмы необходимо лечить в первую очередь, а какие нет, учитывая, что медицинские ресурсы у него не бесконечны. Если удалось найти имущество или оборудование, спасательный робот должен учесть ценность каждого предмета (вдруг это бесценное произведение искусства) и срочность, с которой его необходимо вынести из зоны бедствия. Все это, разумеется, требует глубокого понимания ситуации, которая никому, даже людям, не известна полностью, с массой непредвиденных обстоятельств и с большой вероятностью натолкнуться на что-то ценное или уникальное.

Что еще сложнее для искусственного интеллекта, робот должен учитывать опасность как действия, так и бездействия со своей стороны. По-настоящему похожий на человека робот-дворецкий будет в состоянии понять, что украшенная гирляндами рождественская елка грозит вот-вот упасть посреди всеобщего веселья и устроить короткое замыкание, и немедленно отрегулировать ее так, чтобы в дальнейшем она стояла надежно. Ничто из перечисленного выше не входит в число проверенных навыков у современных роботов и у «разума», который ими управляет.

Вот, стало быть, как обстоят дела в этой области сегодня, когда мы приближаемся к шестьдесят пятой годовщине существования искусственного интеллекта. Робототехники проделали отличную работу по изучению

возможностей роботов и довольно хорошо потрудились над тем, чтобы заставить роботов выполнять целый ряд конкретных операций.

Несравненно меньший прогресс достигнут в трех других областях, которые необходимы автоматам для преодоления трудностей в открытом мире: оценка ситуаций, прогнозирование будущего и принятие динамических решений параллельно с изменением окружающей обстановки — какое из множества возможных действий имеет наибольший смысл в данной ситуации.

Не существует универсального решения ни для выявления того, что возможно и что важно в каждом конкретном сценарии, ни для алгоритмизации действий роботов в сложных и непредсказуемых обстоятельствах. В настоящее время уже можно научить робота подниматься по лестнице или ходить по неровной поверхности (как показали прототипы Boston Dynamics), однако даже такая задача требует огромных усилий. Но несравненно сложнее создать робота, которого можно было бы оставить убираться на кухне в одиночку.

В мире с ограниченным набором правил и ситуаций можно запомнить пусть даже и очень большое, но все-таки конечное число факторов, ведущих к непредвиденным обстоятельствам, и интерполировать их между собой таким образом, чтобы предугадывать незнакомые сценарии. Однако в мире, открытом по-настоящему, обучающих данных никогда не будет достаточно. Если в яблочном соусе растет плесень, роботу нужно понять, как на это реагировать, даже если робот никогда не видел ничего подобного раньше. Для составления простой таблицы с указанием того, что делать в разных обстоятельствах, реальная жизнь слишком сложна [33]. Настоящая причина того, что у нас пока нет домашних роботов, универсальных по своим способностям, заключается в том, что мы не знаем, как их создать, чтобы они были достаточно гибкими для реального мира. Поскольку пространство возможностей обширно и открыто, решения, основанные исключительно на больших данных и глубоком обучении, всегда что-нибудь да упустят. Классические подходы к искусственному интеллекту тоже до сих пор несовершенны, хотя и в несколько другом смысле.

Все это еще раз указывает на важность многообразия когнитивных моделей и глубокого понимания. Даже в ситуации с беспилотным автомобилем внутренние модели поведения машины должны быть куда более разнообразны, чем обычно имеет современный искусственный интеллект. Нынешние системы в основном ограничиваются идентификацией наиболее распространенных на дороге объектов, таких как пешеходы, велосипеды и другие движущиеся транспортные средства. Когда же перед беспилотным автомобилем появляются совершенно иные объекты, системы с ограниченными знаниями не понимают, как себя вести. Например, автопилот Tesla версии 2019 года имеет, по-видимому, очень ограниченное представление о стационарных (или остановившихся на дороге) объектах, подобных рекламным щитам и вставшим на дороге пожарным машинам (первая авария самоуправляемой Tesla, приведшая к смертельному исходу, была, судя по всему, частично вызвана неверным распознаванием большого грузовика, поворачивающего налево,

поскольку его масса была настолько же больше, чем у самой Tesla, как и у рекламной конструкции.

Что же касается хорошего домашнего робота, то наполненность базовой когнитивной модели у него должна быть значительно больше. Число основных элементов окружающей обстановки на шоссе достаточно ограничено, в то время как в самой обычной гостиной можно встретить как минимум стулья, диван (или пару диванов), журнальные столики, ковры, телевизор, напольные лампы, шкафы с книгами или посудой, аквариум, кошку, не говоря уже о совершенно непредсказуемом ассортименте детских игрушек. На кухне можно встретить посуду, бытовую технику, шкафы и полки, еду, раковину, смеситель, используемые и неиспользуемые стулья и столы, миску для кошки и опять саму вездесущую кошку. И хотя кухонная утварь обычно хранится на кухне, нож по случайности может оказаться посреди гостиной и кого-то поранить.

То, что мы обсудили в этой главе, во многом перекликается с тем, о чем говорилось и в предыдущей, хотя она касалась не роботов, а обучения чтению. Возможно, вам уже стало ясно почему. Суть в том, что, хотя конструирование робота (точнее, его физической оболочки) — совсем иная задача, чем создание машины, которая умеет читать, мы все равно упираемся в итоге в одну и ту же проблему понимания открытого мира и адаптации к нему. Просто в случае робота ошибки приводят к физическим, зачастую неустранимым последствиям. Пролить на кого-то кипящий чай гораздо хуже, чем испортить перевод рассказа, но природа нерешенной задачи в обоих случаях идентична.

Так же как без разнообразия когнитивных моделей не может быть нормального чтения, так без них не могут обойтись и безопасные и надежные домашние роботы. Наряду с богатством моделей роботу понадобится и здоровая доза того, что в бытовой речи называют здравым смыслом: глубокое понимание мира и того, как он работает, то есть что может и не может произойти в различных обстоятельствах.

Ни одна из существующих форм искусственного интеллекта всем этим не обладает. Какая же интеллектуальная система имеет по-настоящему богатые когнитивные модели и здравый смысл? Это, разумеется, человеческое мышление.

ГЛАВА 6

Что подсказывает нам собственный разум

Что за волшебный трюк делает нас разумными? Фокус заключается в том, что никакого фокуса просто нет. Сила интеллекта проистекает из огромного разнообразия [нашего мышления и поведения], а не из какого-то единственного совершенного принципа.

Марвин Минский. The Society of Mind

В 2013 году, вскоре после того, как авторы этой книги начали сотрудничать в области искусственного интеллекта, им пришлось столкнуться с очередным безумием, захлестнувшим СМИ. Шумиха, возникшая вокруг весьма сомнительных утверждений, возмутила нас обоих до глубины души. Вот ее

предыстория: два исследователя, Александр Висснер-Гросс и Кэмерон Фреер, написали статью, в которой говорилось, что интеллект любого рода — проявление очень общего физического процесса, называемого «причинно-следственными энтропийными силами». В одном из своих видео Висснер-Гросс заявил, что система, построенная на этой идее, может «ходить прямо, использовать инструменты, сотрудничать, играть в игры, заводить полезные социальные знакомства, развертывать войска в глобальном масштабе и даже зарабатывать деньги, торгуя акциями, и все — сама по себе, без каких-либо команд извне». Наряду с этим документом Висснер-Гросс создал абсурдно амбициозную стартап-компанию Entropica, которая обещала «обеспечить широкие возможности» в здравоохранении, энергетике, разведке, автономной обороне, логистике, транспорте, страховании и финансах.

И средства массовой информации приняли этот бред на ура. Согласно заявлениям обычно вдумчивого научного журналиста Филипа Болла, Висснер-Гросс и его соавтор сформулировали закон [28], который «позволяет неодушевленным объектам вести себя таким образом, что это фактически дает им возможность увидеть свое будущее. Если они следуют этому закону, они могут демонстрировать поведение, напоминающее многие из человеческих действий, включая сотрудничество или использование инструментов для выполнения той или иной задачи». Даже фонд TED предоставил Висснеру-Гроссу платформу для представления своего «нового уравнения интеллекта».

Мы не поверили ни единому слову и заявили в ответ, что разобрали физику и искусственный интеллект Висснера-Гросса по косточкам, написав в довольно ехидной манере в онлайн-статье для журнала *The New Yorker* следующее: «Предполагая, что "причинная энтропия" может решить столь широкий спектр проблем, Висснер-Гросс и Фриер по сути обещают вам телевизор, который будет выгуливать вашу собаку». Оглядываясь назад, мы понимаем, что могли бы сказать то же самое и помягче. Тем не менее факты таковы, что даже спустя полдесятилетия не появилось еще ни одной статьи, развивающей тему причинной энтропии, и мы не видим никаких признаков того, что математика, которой оперировал Висснер-Гросс, достигла бы какого-то прогресса. Стартап-компания Entropica с тех пор себя не проявляла, а сам Висснер-Гросс, похоже, занялся другими проектами [29].

Теории, подобные «причинной энтропии», уже давно сделались притчей во языцех как у публики, так и у профессионалов. Дело в том, что чисто внешне такие идеи очень привлекательны благодаря кажущемуся сходству с хорошо известными крупными физическими теориями, элегантными в своей лаконичности, опирающимися на формальный математический аппарат и обладающими высокой прогностической силой. Средства массовой информации часто поднимают их на щит, потому что их легко подать как большие универсальные классические обобщения. Они претендуют на то, чтобы изменить наш мир, предлагают потенциальные решения целого ряда по-настоящему сложных проблем в рамках единой глобальной парадигмы. И правда, кто бы отказался ворваться в высшие сферы науки на крыльях концепции, бросающей вызов, скажем, общей теории относительности?

Такая же шумиха разразилась почти столетие назад в психологии и вынесла на гребень волны новое многообещающее направление — бихевиоризм. Как известно, психолог из Университета Джонса Хопкинса Джон Уотсон сделал громкое заявление, что может воспитать любого ребенка кем угодно — достаточно лишь тщательно контролировать его окружение и занятия, дополняя это целенаправленными поощрениями и наказаниями. Гипотеза, лежащая в основе бихевиоризма, состоит в том, что человеческий организм реагирует на историю своего становления с математической точностью, как если бы все элементы физиологии и поведения задавались простыми функциями. Чем больше вас вознаграждают за конкретный тип поведения, тем больше вероятность, что вы будете поступать так и в дальнейшем; чем больше вас наказывают за другой тип поведения, тем больше вероятность, что вы от него откажетесь навсегда. К концу 1950-х годов факультеты психологии большинства американских университетов были заполнены психологами, экспериментально изучавшими поведение крыс и голубей и сопровождавшими свои работы тщательными количественными измерениями с целью вывести точные математические закономерности и причинно-следственные связи в психологии.

Однако спустя всего лишь два десятилетия бихевиоризм практически сошел со сцены благодаря работам Ноама Хомского (подробности этого мы обсудим чуть позже). То, что работало с крысами (и то лишь в ограниченном наборе экспериментов), оказалось совершенно бесполезным (и неуместным) в случае людей. Поощрения и наказания, естественно, оказывают определенное воздействие на всех, но куда большее значение имеет множество других факторов, которые эта теория полностью игнорировала.

Проблема, по словам ученых Йельского университета Чеза Файерстоуна и Брайана Шолла, заключается в том, что «не существует единого принципа, по которому работает человеческий интеллект, потому что разум — это не что-то одно. Разум включает в себя различные составляющие, все они функционируют по-разному. Различать цвета — совсем иная процедура, нежели планировать отпуск, а они обе не похожи на процессы, вовлеченные в понимание предложений, составляющих речь, или на осознанное управление телом, или на работу памяти, или на испытывание чувств и эмоций». Ни одно уравнение никогда не сможет отразить все разнообразие того, что умеет делать человеческий интеллект.

Компьютеры вовсе не обязаны работать в точности так же, как люди. Им не требуется быть зависимыми от множества когнитивных ошибок, которые ухудшают человеческое мышление, например от желания больше верить фактам, подтверждающим вашу точку зрения, чем фактам, противоречащим ей. Нет нужды навязывать компьютерам и многие другие ограничения человеческого разума. Людям бывает сложно запоминать списки, состоящие из более чем семи предметов, но машина попросту не сталкивается с такими трудностями. Механический разум выполняет математические операции не так, как люди, и нет причин ему в этом мешать. Вообще, человеческий интеллект имеет множество недостатков, и компьютерам совершенно

необязательно подражать в этом своим создателям. Тем не менее для усовершенствования ИИ многое можно почерпнуть именно из работы человеческого разума — в том, где он все еще намного превосходит машины, в частности в области чтения, понимания языка и гибкости мышления.

Ниже мы перечислим 11 фактов из области когнитивных наук — психологии, лингвистики и философии, — которые мы считаем критически важными для придания искусственному интеллекту той гибкости, широты и устойчивости, которыми обладает человеческое мышление.

Серебряных пуль не существует

Едва начав читать о статье Висснера-Гросса и Фреера, посвященной причинной энтропии, мы тут же поняли, что претензии авторов нереалистичны. Бихевиоризм тоже пытался претендовать на слишком многое, и это сослужило ему плохую службу. Действительно — вы заявляете, что можете объяснить любое поведение, реальное или воображаемое, всякий раз опираясь только на историю воспитания рефлексов у животного. Если животное, вопреки вашим предсказаниям, сделало что-то неожиданное, вы просто вспоминаете другие аспекты воспитания, которые кажутся более подходящими для причин данного случая. Из-за постоянного использования подобной аргументации бихевиоризм почти не давал надежных предсказаний, зато постоянно предлагал то один, то другой механизм для «объяснения» того, что уже произошло. В результате от этой теории осталось только одно по-настоящему бесспорное и важное утверждение: животные, в том числе и люди, любят делать вещи, за которые получают награду. Это абсолютно верно, ведь при прочих равных условиях люди обычно выбирают вариант, который сулит им большее вознаграждение. Беда лишь в том, что это утверждение настолько банально, что ради него не стоило проводить такое множество экспериментов.

Едва ли поэтому бихевиоризм вообще способен объяснить то, как, скажем, человек понимает фразы из диалога в фильме или осваивает использование эксцентрикового зажима при сборке книжной полки, купленной в ИКЕА. Вознаграждение часто является частью системы мышления и поведения, но все же это не сама система. Висснер-Гросс попросту заново вернулся к вопросу о вознаграждении: с его точки зрения, организм (мышление, разум) работает правильно, если успешно сопротивляется хаосу (энтропии) вселенной, ведь никто из нас не хочет превращаться в прах, и поэтому сопротивляется разрушению. С этим утверждением трудно поспорить, однако оно почти ничего не говорит нам о том, как мы делаем индивидуальный выбор.

Глубокое обучение в значительной степени загнало себя в ту же самую ловушку. По сути, оно применяет относительно современную математику (сформулированную на языке ошибок и затрат) для оптимизации вознаграждения, получаемого нейронными сетями, но совершенно игнорирует множество других вещей, в которых нуждаются ИИ-системы для достижения того уровня мышления, который мы называем глубоким пониманием.

Но если изучение нейробиологии что-то нам и дало, так это то, что мозг человека (и животных) невероятно сложен: его часто называют самой сложной системой в известной нам вселенной, и это справедливо. Средний человеческий мозг имеет около 86 млрд нейронов, представленных сотнями, если не тысячами, различных типов; в нем насчитывается несколько триллионов синапсов, а в каждом отдельном синапсе — сотни индивидуальных белков: на каждом уровне сложность структур и функций в нем огромна. Анатомы выделяют более 150 четко различимых областей мозга, объединяемых и координирующихся обширными системами связей. Как сказал в 1906 году нейробиолог-новатор Сантьяго Рамон-и-Кахаль во время своей нобелевской лекции: «К сожалению, природа, кажется, не осознает нашей интеллектуальной потребности в удобстве и единстве и очень часто находит особое удовольствие в создании сложных и разнообразных систем».

По-настоящему умные и гибкие системы искусственного интеллекта тоже, вероятно, будут полны разнообразия и сложности, уподобившись в этом мозгу. Любая теория, предлагающая свести весь интеллект к одному-единственному принципу или уникальному базовому алгоритму, неизбежно окажется слепым поводырем слепых.

Познание подразумевает широкое использование внутренних образов

Бихевиоризм оказался колоссом на глиняных ногах, но окончательно его сокрушила монография, написанная в 1959 году Ноамом Хомским. Главной мишенью критики Хомского было моделирование «словесного поведения» — попытка объяснить человеческий язык, предпринятая Б. Ф. Скиннером, который тогда считался одним из ведущих психологов мира.

По сути, критика Хомского обращалась к вопросу о том, можно ли понимать человеческий язык лишь как продукт истории становления человеческой личности, которая зависит от событий, происходивших во внешней среде, окружавшей того или иного человека (что человек говорил и какую реакцию он получал в ответ), или же ведущую роль в языке играет внутренняя личностная структура, до определенной степени не зависящая от внешних факторов. В заключительной части своей книги Хомский особенно подчеркивал, что мы распознаем новую информацию именно как предложение не потому, что она соответствует определенному, ранее знакомому элементу тем или иным легко объяснимым способом, а потому, что предложение создается грамматикой языка, которую каждый индивид способен усвоить самостоятельно.

Хомский утверждал, что, только изучив эту внутреннюю грамматику, мы сможем понять, как дети учатся говорить. Простая схема, состоящая только из реакции на слова и стимулы, никогда не приведет нас к раскрытию языкового восприятия.

В результате на месте бихевиоризма возникла новая область науки, которую назвали когнитивной психологией. В тех случаях, где бихевиоризм

пытался объяснить поведение целиком на основе опыта вознаграждений и наказаний, полученных извне (вспомните про стимулы и ответы на стимулы, стандартно применяющиеся в контролируемом обучении, которое так популярно в современных приложениях глубокого обучения), когнитивная психология в основном фокусируется на внутренних представлениях, таких как убеждения, желания и цели.

В нашей книге мы постоянно приводим примеры того, что машинное обучение (в частности, нейронные сети) пытается завоевать и удержать место под солнцем, почти не используя представления в качестве подхода, и что из этого выходит. В строгом техническом смысле нейронные сети все-таки имеют то, что можно назвать представлениями: в качестве их выступают наборы чисел (известные как векторы), которые представляют параметры входов в систему и выходов из нее, а также и скрытые структурные единицы. Тем не менее всему этому крайне недостает хотя бы минимального разнообразия. В частности, отсутствуют какое-либо средства для представления того, что когнитивные психологи называют пропозициями, которые обычно описывают отношения между различными сущностями. Например, в классической системе искусственного интеллекта для представления знаменитого визита президента Джона Ф. Кеннеди в Берлин в 1963 году (когда он сказал «Ich bin ein Berliner» [34]), можно добавить ряд фактов, таких как PART-OF (BERLIN, GERMANY) и VISITED (KENNEDY, BERLIN, JUNE1963). Получение знаний в классическом искусственном интеллекте состоит отчасти именно в накоплении таких представлений, и на этой основе в дальнейшем строится вывод; так, в описанном случае несложно сделать вывод, что Кеннеди посетил Германию.

Глубокое обучение кое-как справляется с этой задачей, используя множество векторов, которые частично способны зафиксировать некоторые детали происходящего, но никогда не оперирует непосредственно представлениями. В рамках этого подхода не существует реального способа представить комбинацию слов VISITED (KENNEDY, BERLIN, JUNE1963) или PART-OF (BERLIN, GERMANY); возможно только очень грубое приближение. При благоприятных обстоятельствах типичная система глубокого обучения может правильно сделать вывод о том, что Кеннеди посетил Германию, но мы никогда не дадим гарантий того, что она умеет это делать в принципе. Алгоритм легко может дать сбой, и в следующий раз та же самая система глубокого обучения заявит, что Кеннеди посетил Восточную Германию (что, конечно, было в 1963 году совершенно невозможно) или что его брат Роберт посетил Бонн — и все это лишь потому, что перечисленные два варианта находятся близко друг к другу в так называемом векторном пространстве. Главная причина, по которой вы не можете рассчитывать на глубокое обучение в таких аспектах, как выводы и абстрактные рассуждения, заключается в том, что они не предназначены для точного представления фактических знаний. Если факты недостаточно ясны (с точки зрения машинного интеллекта), системе будет очень трудно правильно оперировать информацией.

Отсутствие представлений, выраженных в явной форме, вызывает аналогичные проблемы в игровой системе DeepMind Atari. Вспомним ее отказ в Breakout: когда весло перемещается всего на несколько пикселей к берегу, алгоритм падает, и объясняется это тем, что Atari вообще не умеет представлять абстракции (такие как весла, шары или стены) и тем более — пользоваться ими.

Между тем без абстрактных представлений трудно создать истинную когнитивную модель разума. А без настоящей когнитивной модели не может быть и надежности в реальном мире. Все, что вы можете иметь вместо этого, — лишь множество данных плюс надежда на то, что предъявляемые объекты не будут слишком отличаться от наборов, взятых для обучения. Но надежда эта часто не оправдывается, и когда новые данные выходят за пределы статистических закономерностей, выученных машиной, происходит крах.

Если нашей задачей будет сконструировать эффективные ИИ-системы для решения сложных проблем, представления, причем многочисленные и разнообразные, становятся насущно необходимыми. Не случайно, когда в DeepMind решили создать систему, которая могла бы на самом деле играть в го на человеческом (или сверхчеловеческом) уровне, они отказались от подхода обучения по пикселям (который они использовали в своей предыдущей игровой работе Atari) и начали с подробного представления правил игры в го, структуры игрового поля, вручную написали алгоритмы для дерева представления и поиска ходов каждого игрока. Как сказал эксперт по машинному обучению Университета Брауна Стьюарт Жеман, «фундаментальные проблемы в нейронном моделировании связаны именно с представлениями, а не с обучением как таковым».

Огромную роль в познании играют абстракция и генерализация

Многое из того, что мы знаем, представляет собой несомненные абстракции. Например, такое отношение, как «X — это сестра Y», применимо сразу ко множеству пар людей. Малия Обама — это сестра Саши Обамы, принцесса Анна — это сестра принца Чарльза и т.д. Иначе говоря, мы не просто знаем, что определенная пара людей является братом и сестрой, мы знаем, что такое сестры и братья в целом, и можем применить эти знания к конкретным людям, подпадающим под соответствующие определения. Например, мы знаем, что, когда люди имеют одних и тех же родителей, они являются братьями и сестрами. Если мы знаем, что Лора Инглз-Уайлдер была дочерью Чарльза и Кэролайн Инглз, а затем узнаем, что Мэри Инглз была дочерью той же самой пары, то без труда можем сделать вывод, что Мэри была сестрой Лоры. Кроме того, мы можем сделать вывод и о том, что Мэри и Лора наверняка были знакомы друг с другом (так как большинство людей хорошо знают своих братьев и сестер) и что они имели некоторые общие генетические особенности, как и определенные черты, говорящие о семейном сходстве.

Представления, лежащие в основе наших когнитивных моделей, равно как и здравого смысла, все базируются на весьма обширной и разнообразной коллекции всевозможных абстрактных отношений, которые затем объединяются в сложные структуры. Действительно, люди могут абстрагировать практически все что угодно: отрезки времени («10:35 вечера»), пространственные единицы («Северный полюс»), различные события («убийство Авраама Линкольна»), социально-политические организации («Государственный департамент США», «теневой интернет»), свойства и признаки («красота», «усталость»), отношения («отцовство», «партнерство»), теории («марксизм») и теоретические построения («гравитация», «синтаксис»). Более того, все они затем становятся языковыми единицами, служащими для объяснения, сравнения или повествования, для анализа различных по сложности ситуаций и понимания их основы. Это предоставляет разуму эффективнейшие инструменты для рассуждений об окружающем мире во всей его полноте.

Вот диалог, имевший место дома у Гэри, когда мы готовили эту книгу. В тот момент его сыну Александру было пять с половиной лет.
АЛЕКСАНДР: Что значит «вода по грудь»?
МАМА: «Вода по грудь» значит, что вода достает тебе до груди.
ПАПА: Для каждого человека это будет по-разному. «Вода по грудь» для меня глубже, чем для тебя.
АЛЕКСАНДР: Значит, «вода по грудь» для тебя — это то же, что «вода выше головы» для меня.

Именно такая гибкость мышления, сопровождающаяся обобщениями, изобретением новых концепций и расширением старых и часто основанная на очень малом количестве исходных данных, и должна стать главной целью разработки искусственного интеллекта.

Когнитивные системы обладают очень совершенной структурой

В бестселлере «Думай медленно… решай быстро» лауреат Нобелевской премии Даниэль Канеман разделил когнитивные процессы человека на две категории: система 1 и система 2. Внутри системы 1 (быстрой) все процессы выполняются мгновенно, часто автоматически. Человеческий разум просто запускает их, и все — у вас нет никакого представления о том, как вы это делаете. Когда вы смотрите на мир, вы мгновенно анализируете сцену перед собой, и, когда вы слышите речь на своем родном языке, вы сразу же понимаете, о чем люди разговаривают. Вы не можете контролировать это, и вы не представляете, как ваш разум это делает; по сути, вы даже не осознаете того, что ваш ум при этом работает. В системе 2 (медленной) процессы требуют сознательного, пошагового мышления. Когда задействована система 2, у вас появляется осознание мышления, в том числе и мыслительных усилий. Так, например, мы разгадываем головоломки, решаем математические задачи

или медленно читаем текст на иностранном языке, который сейчас изучаем, когда приходится искать в словаре чуть ли каждое третье слово [35].

Мы предпочитаем использовать для этих двух систем термины «рефлексивная» и «рассудительная», потому что они более удобны с мнемонической точки зрения, но в любом случае ясно, что люди используют для разных типов проблем разные виды высшей нервной деятельности. Один из первопроходцев в области искусственного интеллекта Марвин Мински зашел в своих рассуждениях о разуме настолько далеко, что настаивал на рассмотрении человеческого познания как «общества разума» с десятками или сотнями различных «агентов», каждый из которых специализируется на выполнении различных задач. Например, для того чтобы выпить чашку чая, требуется взаимодействие GRASPING-агента, BALANCING-агента, THIRST-агента и некоторого числа MOVING-агентов [36]. Идеи Говарда Гарднера о множественном интеллекте, как и триархическая теория интеллекта Роберта Штернберга, указывают в том же направлении, что и большое число исследовательских работ в области эволюционной психологии и психологии развития: все они сходятся на той мысли, что разум — не что-то одно, а целый набор «сущностей» и «операторов».

Нейрология рисует еще более сложную картину, в которой сотни различных областей мозга, каждая из которых имеет свою особую функцию, объединяются в различные схемы для выполнения любого конкретного вычисления. Хотя популярная идея, насчитывающая уже много десятилетий и утверждающая, что мы якобы используем только 10% нашего мозга, не соответствует действительности, мы должны согласиться с тем, что мозговая деятельность является дорогостоящей в метаболическом плане и поэтому мы редко (или вообще никогда) задействуем все ресурсы мозга одновременно. Как правило, то, что мы делаем в каждый конкретный момент, требует лишь некоторой части наших мозговых структур, и когда одни области мозга активны, другие, скорее всего, будут бездействовать. Затылочная кора обычно задействуется в зрительном анализе, мозжечок участвует в координации движений и т.д. Мозг — исключительно высокоструктурированный орган, и большая часть нашего интеллектуального мастерства происходит от использования правильных нейронных инструментов в необходимый для этого момент. Мы, следовательно, можем ожидать, что истинный искусственный интеллект также будет высокоструктурирован и основное его преимущество будет заключаться в способности правильно использовать для решения конкретной когнитивной задачи тот или иной элемент его структуры в подходящий момент.

По иронии судьбы нынешняя тенденция почти полностью противоположна только что высказанной идее. Сейчас в машинном обучении наблюдается заметный перекос в сторону сквозных моделей, где используется единый однородный механизм с максимально упрощенной внутренней структурой. Примером является модель вождения Nvidia 2016 года, которая отказалась от классического разделения модулей по функциям, таким как восприятие, прогнозирование и принятие решений. Вместо этого в ней присутствует лишь

одна довольно однородная нейронная сеть, в которой нет специализированных элементов, а вместо них идет обучение более прямым корреляциям между многими входными данными (пикселями) и одним набором выходных данных (инструкциями по управлению и ускорению). Поклонники такого рода систем указывают на достоинства совместной тренировки всей системы без необходимости обучать по отдельности несколько модулей (для восприятия, прогнозирования и т.д.).

На определенном уровне такие системы концептуально проще и выглядят удобнее: уже не нужно разрабатывать отдельные алгоритмы восприятия, предсказания и всего остального. Более того, на первый взгляд модель показывает себя неплохо, о чем свидетельствует впечатляющее видео. Зачем беспокоиться о гибридных системах, которые идентифицируют восприятие, принятие решений и прогнозирование как отдельные модули, если гораздо проще иметь всего лишь одну большую сеть и правильный обучающий набор?

Проблема в том, что такие системы редко обладают необходимой гибкостью. Система Nvidia работала много часов подряд, не требуя серьезного вмешательства со стороны людей-водителей, но все же это были не тысячи часов (как в случае с модульной системой Waymo). При этом Waymo умела перемещаться из пункта A в пункт B и справляться хотя бы с такими вещами, как смена полосы движения. Все, чем владеет Nvidia, — это способность придерживаться одной полосы движения. Конечно, даже такое умение важно, но ведь оно представляет лишь малую долю того, что связано с вождением.

Когда наступает момент истины и у лучших исследователей искусственного интеллекта возникает желание решать более сложные проблемы, они чаще возвращаются к использованию гибридных систем, и мы вправе ожидать, что это постепенно будет становиться все более и более стандартным исходом. Компания DeepMind была (почти) в состоянии научить Atari играть в простые игры без участия гибридных систем, внедрив сквозной процесс от пикселей и игрового счета до действий джойстика. Однако в случае с го они уже были не в состоянии ограничиться этим, поскольку во многих отношениях эта древняя игра намного более сложная, чем игры 1970-х и 1980-х годов. В го как минимум существует гораздо больше всевозможных игровых позиций, а конкретные ходы могут вести к гораздо более запутанным последствиям. С чистым сквозным обучением, таким образом, пришлось распрощаться, и гибридные системы опять оказались востребованными.

Для достижения победы в го требовалось объединить два разных подхода — глубокое обучение и технику, известную как игровое моделирование по методу Монте-Карло, позволяющее выбирать оптимальное продолжение игры среди множества возможностей, которые можно наглядно изобразить в виде ветвящегося дерева. Между тем само моделирование по методу Монте-Карло тоже является гибридом двух других техник, которые появились на свет в 1950-х годах, — это «поиск по дереву» в играх для прогнозирования возможных ходов игроков в будущем — хрестоматийная методика при конструировании искусственного интеллекта — и собственно моделирование по методу Монте-Карло, широко распространенный способ рандомной

симуляции с одновременным ведением статистики по результатам. Ни одна из этих систем — ни глубокое обучение, ни моделирование по методу Монте-Карло — не могла бы стать чемпионом мира по игре го. Урок здесь заключается в том, что искусственный интеллект, как и человеческий разум, должен быть структурирован с использованием множества инструментов для решения различных аспектов сложных проблем [37].

Даже простые на первый взгляд аспекты познания иногда требуют множества инструментов для реализации

Даже в самых мелких деталях механизмы когнитивного восприятия часто оказываются сложной системой, состоящей не из одного, а из множества компонентов. Возьмем, например, глаголы и их формы прошедшего времени — почти что общемировую грамматическую систему, которую психолингвист Стивен Пинкер однажды назвал «дрозофилой лингвистики»: действительно, здесь есть сходство с этими крошечными лабораторными насекомыми, с помощью которых генетики так многому научились. В английском и многих других языках основная масса глаголов образует формы прошедшего времени с помощью регулярных простых правил (в случае английского это добавление окончания «-ed»: talk — talked, perambulate — perambulated). Другие глаголы в тех же языках имеют нерегулярные формы прошедшего времени, не укладывающиеся в какое-либо общее правило (*англ.* sing — sang, ring — rang, bring — brought, go — went). Часть совместной докторской работы Гэри с Пинкером была посвящена ошибкам детей в чрезмерной регуляризации форм прошедшего времени (когда неправильный глагол при образовании этих форм преобразуется так, будто бы он был обычным глаголом, скажем, break — breaked (правильно «broke») или go — goed (правильно «went»). Основываясь на проанализированных данных, они выдвинули аргумент о существовании гибридной модели — крошечной структуры на грамматическом микроуровне, в которой правильные глаголы преобразуются по стандартным правилам (так же, как это можно найти в компьютерных программах и в классическом искусственном интеллекте), тогда как неправильные глаголы приобретают новые формы через ассоциативные сети (которые, по сути, и были предшественниками глубокого обучения). Эти две очень непохожие друг на друга системы сосуществуют и дополняют друг друга: нерегулярные глаголы требуют включения памяти, а регулярные следуют стандартной схеме, в том числе и тогда, когда релевантных данных почти нет.

Аналогичным образом человеческий разум имеет дело с концепциями сразу в нескольких формах (или режимах). Частично он ориентируется по общим определениям, частично — по типичным признакам, частично — по наиболее характерным примерам. Мы зачастую одновременно отслеживаем, что типично для некоторой категории и что должно быть в ней для того, чтобы она соответствовала некоторым более общим формальным критериям. Тина Тернер, уже став бабушкой, продолжала танцевать в мини-юбках. Возможно, на сцене она и не выглядела как типичная бабушка, но она вполне

соответствовала по этому критерию формальным определениям родственных отношений: у нее были дети, и у этих детей, в свою очередь, тоже были дети.

Ключевой проблемой для искусственного интеллекта является поиск жизнеспособного баланса между механизмами, которые просто усваивают абстрактные истины (например, «большинство млекопитающих рожают живых детенышей»), и механизмами, которые имеют дело с негостеприимным миром исключений (утконос — млекопитающее, однако он откладывает яйца). Универсальный искусственный интеллект потребует не только механизмов глубокого обучения для распознавания образов, но и механизмов обработки рассуждений и создания обобщений, более близких к механизмам классического искусственного интеллекта и к миру, где царят правила и абстракции.

Как недавно заявил Демис Хассабис: «Истинный интеллект — это нечто намного большее [чем просто еще одна разновидность перцептуальной классификации, в которой так преуспели системы глубокого обучения]; вы должны объединить ее с высокоуровневым мышлением и символическим мышлением, не говоря уже о множестве вещей, в которых классический искусственный интеллект уже пытался разобраться в 1980-х». Разработка широкого интеллекта потребует от нас объединения целого ряда несходных инструментов, старых и новых, и методы такого объединения нам еще предстоит открыть.

Человеческий язык и мышление композиционны

Суть языка для Хомского заключается в афоризме одного из первых лингвистов — Вильгельма фон Гумбольдта (1767–1835), назвавшего язык «бесконечным использованием конечных средств». Имея ограниченный мозг и ограниченное количество лингвистических данных, мы способны создать грамматику, которая затем позволяет нам произносить и понимать бесконечный диапазон предложений, во многих случаях путем построения более крупных предложений (таких, как это) из более мелких компонентов — слов и словосочетаний. Если мы в состоянии сказать, что «моряк любил девушку», то мы сможем использовать это предложение как одну из составляющих в более крупном предложении («Мария вообразила, что моряк любил девушку»), которая может служить составляющей в предложении еще большего объема («Крис написал эссе о том, как Мария вообразила, что моряк любит девушку») и т.д., причем каждое из этих предложений мы можем легко интерпретировать.

На противоположном полюсе — первопроходец в области нейронных сетей Джефф Хинтон, который в своей среде почитается таким же «духовным лидером», как Хомский в лингвистике. В последнее время Хинтон участвовал в многочисленных спорах относительно подхода, который он называет «векторами мышления». С алгебраической точки зрения вектор — это просто некоторое множество чисел, например пара координат [40,7128 °N, 74,0060 °W], которая представляет собой долготу и широту Нью-Йорка, или, скажем

[52 419, 663 268... 24 230, 97 914] — 52 числа, которые отображают собой площади в квадратных милях всех штатов США, перечисленных в алфавитном порядке. В системах глубокого обучения каждый вход и каждый выход можно описать как вектор, причем каждый «нейрон» в сети вносит в соответствующий вектор ровно одно число. Этот принцип позволяет специалистам, работающим в сфере машинного обучения, кодировать слова в виде векторов, исходя из того, что любые два слова, которые имеют одинаковое значение, должны кодироваться с помощью одинаковых векторов. Допустим, слово «cat» (кошка) кодируется как [0, 1, –0,3, 0,3], а слово «dog» (собака) выглядит как вектор с координатами [0, 1, —0,35, 0,25]. Описанная техника известна как Word2Vec [38]. Она существует уже в течение ряда лет и была разработана Ильей Суцкевером и Томашем Миколовым в тот период, когда они были сотрудниками Google. В принципе, такой подход часто позволяет компьютерам эффективно и быстро находить векторы различных слов (каждый из которых представлен парой сотен действительных чисел), основываясь на других словах, которые статистически регулярно появляются в текстах рядом с первыми [39].

В определенных контекстах Word2Vec действительно работает эффективно. Возьмем слово «саксофон». В большой коллекции образцов письменного английского языка слово «саксофон» регулярно встречается рядом с определенными словами, например «играть» и «музыка», а также с некоторыми именами, такими как Джон Колтрейн и Кенни Джи. В большой базе данных статистика употребления слова «саксофон» близка к статистике для слов «труба» и «кларнет» и очень непохожа на статистику для слов «лифт» или «страхование». Поисковые системы могут использовать подобные закономерности для определения синонимов; в частности, благодаря таким методам (или их модификациям) поиск товаров на Amazon также стал намного точнее.

Однако по-настоящему знаменитым сделало Word2Vec другое открытие. Оказывается, данная технология позволяет выявлять словесные аналогии, формулируемые примерно таким образом: «если слово "мужчина" относится к слову "женщина" так же, как слово "король" к слову xxxxx, то что такое xxxxx?» Если, далее, сложить координаты векторов, представляющих слова «король» и «женщина», и вычесть из этой суммы координаты слова «мужчина», а затем найти ближайший вектор, то вы получите ответ «королева», причем для этого вообще не требуется иметь представление о том, что такое «король» или что такое «женщина». В то время как исследователи традиционного искусственного интеллекта годами пытались определить эти понятия в машинных терминах, Word2Vec, судя по всему, нашел столь же кардинальный (и безжалостный) способ решить эту проблему, как Александр Македонский поступил с Гордиевым узлом.

Основываясь по большей части на этих результатах, Хинтон попытался экстраполировать данную идею на как можно более широкий круг ментальных объектов. Вместо представления предложений и мыслей в виде сложных деревьев, которые плохо взаимодействуют с нейронными сетями, почему бы не

изобразить все это просто как векторы? В своем интервью для журнала *The Guardian* Хинтон заявил: «Если вы возьмете вектор для Парижа, вычтете вектор для Франции и приплюсуете вместо этого Италию, вы получите Рим. Разве это не замечательно?» Подобные методы, указывал Хинтон, лежат в основе новейших достижений Google в области машинного перевода, так почему бы не представить все мысли таким образом?

Проблема в том, что предложения отличаются от слов по многим параметрам. Вы можете аппроксимировать значение слова, рассматривая, как оно использовалось в разных обстоятельствах, при этом номинальное значение слова «кошка» реально будет отождествить со средним значением всех случаев использования данного слова в известных нам контекстах, или (выражаясь более техническим языком) изобразить его как облако точек в векторном пространстве, которое система глубокого обучения использует для представления слов. Однако, в отличие от слов, почти каждое предложение уникально по своему смыслу. Фраза «Джону легко угодить» совсем не то же самое, что фраза «Джон любит угождать», хотя буквы в этих двух предложениях не так уж различны. При этом предложение «Джону легко угодить» уже кардинально отличается по смыслу от предложения «Джону нелегко угодить», хотя речь идет о добавлении всего одного слова.

Идеи и нюансы отношений между ними слишком сложны, чтобы можно было просто объединить их в предложения, которые оказываются похожими только внешне. Мы легко понимаем разницу между словосочетаниями «(некая) книга, которая находится на (этом) столе» и «некий стол, который есть в этой книге», как и разницу между ними обоими и словосочетанием «(некая) книга, которая не находится на (этом) столе». При этом каждая из них взята из предложения «Джеффри знает, что Фред не даст и ломаного гроша за книгу, которая находится на столе, но его весьма интересует большая и необычная скульптура рыбы, на самой верхушке которой в данный момент балансирует стол, особенно потому, что стол уже накренился вправо и может упасть в любую секунду». Каждое из этих предложений можно бесконечно размножать, причем все дочерние предложения приобретают совершенно разные значения. Иначе говоря, в каждом случае целое кардинально отличается от своих статистически усредненных частей [40].

*В данном случае — неопределенный артикль

Рис. 6.1. Пример синтаксического древа

Именно по этой причине лингвисты обычно представляют язык с помощью ветвящихся диаграмм, называемых деревьями (обычно рисующимися ветвями вниз, рис. 6.1).

В таком представлении каждый компонент предложения имеет свое место, и благодаря этому легко отличить одно предложение от другого и определить отношения между этими элементами, даже если почти все слова в этих двух предложениях одни и те же. Полностью лишенные таких высокоструктурированных представлений, системы глубокого обучения, как правило, сталкиваются с массой проблем там, где человек мгновенно может сориентироваться в тонкостях языка.

Существует, например, «анализатор настроений», основанный на глубоком обучении, — это система, которая пытается классифицировать, является ли данное предложение утвердительным или отрицательным. Техника этого такова: каждое предложение преобразуется в вектор, и предполагается, что положительные предложения типа «Мне понравилось!» будут представлены одним набором векторов, которые похожи друг на друга (то есть сгруппированы вместе в векторном пространстве), в то время как отрицательные предложения вроде «Мне не понравилось!» будут представлены другим набором векторов, которые группируются в отдельный кластер. Когда система сталкивается с новым предложением, она, по существу,

пытается выяснить, к какому кластеру оно ближе — к набору положительных векторов или к множеству отрицательных векторов.

Многие входные предложения очевидны по своему посылу и классифицируются системой правильно, но уже в тонких различиях они часто теряются. Подобные системы не могут различить, скажем, такие фразы: «Мне нравилось, пока не разонравилось» (например, негативный отзыв о фильме, который начинался хорошо, но потом скатился во что-то скучное) и «Мне сначала не нравилось, но потом понравилось» (соответственно — более позитивный отзыв о фильме, который начинался так себе, но потом стал по-настоящему увлекательным). Это и не удивительно, поскольку они не анализируют структуру предложения с точки зрения того, как она относится к его составным частям, и, что еще важнее, они не понимают, как значение предложения вытекает из его частей.

Мораль такова: статистические данные часто довольно близко подбираются к значению слова, но они никогда не понимают, о чем в реальности мы говорим. Но если они не могут точно уловить смысл даже отдельных слов [30], то им тем более не удастся понять сложные мысли или предложения, которые их описывают. Как однажды выразился Рэй Муни, специалист по компьютерной лингвистике в Университете Техаса: «Вы не можете втиснуть весь смысл каждого гребаного предложения в один гребаный вектор!» [31] Довольно грубо, но зато в точку: нельзя требовать от векторов слишком много [41].

Надежное понимание мира требует иерархического анализа информации одновременно в восходящем и нисходящем направлениях

Посмотрите на следующую картинку (рис. 6.2). Что это — буква или число?

Рис. 6.2. Что это — буква «B» или число 13?

В разных ситуациях этот рисунок можно интерпретировать и одним способом, и другим — все зависит от контекста. Смотрите сами (рис. 6.3).

Рис. 6.3. Интерпретация изображения зависит от контекста

Специалисты по когнитивной психологии часто делят используемую нами информацию на две группы: информация, движущаяся «снизу вверх», то есть поступающая непосредственно от наших чувств, и знания, идущие «сверху

вниз», которые по происхождению являются информацией о мире, полученной ранее [42]. Ко второй категории относится, например, знание о том, что буквы и цифры образуют разные типы символики, что слова и многозначные числа состоят из элементов, взятых из соответствующих наборов символов, и т.д. Изображения с неоднозначной трактовкой идентифицируются в одном контексте одним образом, а в другом контексте — другим образом именно потому, что мы пытаемся объединить информацию, получаемую на сетчатку глаза, с уже «упакованной» у нас в мозгу логичной картиной мира.

Прочитайте учебники по психологии, и вы увидите десятки примеров этого. В одном из классических экспериментов испытуемых просили посмотреть на изображения, подобные приведенному ниже (рис. 6.4), а затем нарисовать их по памяти на листке бумаги, где заранее были написаны определенные слова или словосочетания, например «солнце» или «штурвал» — в одном месте и «занавески в окне» или «ромб в прямоугольнике» — в другом.

Рис. 6.4. Изображения с несколькими возможными интерпретациями

То, как люди рисовали изображения по памяти, очень сильно зависело от того, какие надписи находились перед их глазами (рис. 6.5).

Рис. 6.5. То, как мы расшифровываем последовательность изображений, в значительной мере

зависит от контекста

Одна из наших любимых демонстраций важности контекста в восприятии создана в лаборатории Антонио Торральбы в Массачусетском технологическом институте. На экране было создано изображение озера с рябью, которая вызывает смутное ощущение присутствия автомобиля, стоящего прямо на поверхности воды. Если на мониторе вы увеличите фрагмент с рябью на глади озера, чтобы рассмотреть ее детальнее, вы действительно разглядите пятна света, напоминающие по очертаниям автомобиль. Такая игра света вполне способна ввести в заблуждение зрительный анализатор человека, однако никого из нас она не обманет, потому что мы знаем, что на самом деле автомобили не могут ездить по поверхности озера.

Вот еще один пример. Посмотрите на некоторые детали, которые мы взяли с этой фотографии кухни Джулии Чайлд, помещенной ниже (рис. 6.6).

Рис. 6.6. Фотография кухни Джулии Чайлд

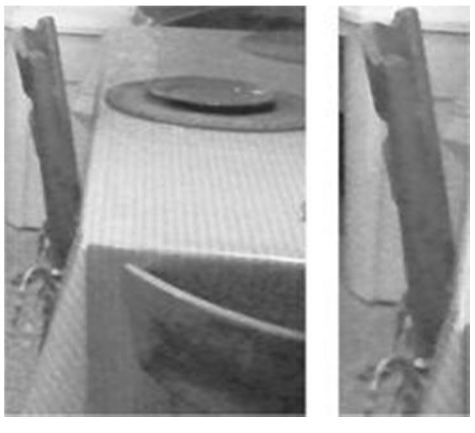


Рис. 6.7. Отдельные детали, взятые с фотографии кухни на рис. 6.6

Можете ли вы определить фрагменты под картинкой? Естественно, для вас это будет нетрудно. Изображение слева — это кухонный стол с двумя стульями вокруг него (спинка третьего стула на противоположной стороне выглядит как заметный полумесяц) и тарелка на подставке для столовых приборов сверху. Деталь справа — это просто кресло, которое на исходной фотографии находится слева от стола. Но пиксели, соответствующие столу и стульям, сами по себе не способны сказать нам об этом (рис. 6.7).

Действительно, если мы пропустим эти детали через программное обеспечение [32], используемое Amazon для распознавания фотографий (Rekognition), то оно пометит левую картинку с уверенностью в 65,5% как «кусок фанеры», а правую — как «полевую дорогу» или «грейдер» с надежностью 51,1%. Без контекста пиксели сами по себе не несут смысловой информации.

То же самое относится и к нашему пониманию языка. Одной из областей, где контекст имеет исключительно важное значение, является разрешение двусмысленности, примеры которой мы приводили выше. На днях один из нас увидел на проселочной дороге табличку с надписью «Бесплатный конский навоз» (англ. «Free horse manure», здесь «free» можно перевести и как прилагательное «бесплатный» («свободно отдаваемый»), и как глагол в повелительном наклонении «Освободите!»). Чисто логически говоря, эта надпись могла бы оказаться и лозунгом с требованием типа «Свободу конскому навозу!» (та же синтаксическая конструкция, что и в лозунге «Свободу Нельсону Манделе!»: англ. «Free Nelson Mandela!»), и объявлением о раздаче чего-то, что для владельца является лишним. Естественно, в данном случае никаких проблем понимания не возникло [43], потому что с конским навозом, насколько нам известно, не связано никаких освободительных движений.

Знания о мире также важны для интерпретации небуквальных языковых сообщений. Когда один официант в ресторане говорит другому: «Тот ростбиф желает порцию кофе», никому не придет в голову, что сандвич с ростбифом внезапно захотел взбодриться; мы делаем вывод, что человек, который заказал ростбиф, заказал еще и чашечку кофе. Действительно, все, что мы знаем о мире, подсказывает нам, что у самих сэндвичей не бывает желаний.

Перефразируя сказанное средствами лингвистической терминологии, можно утверждать, что «язык имеет тенденцию быть недостаточно конкретизированным»; это означает, что мы не говорим буквально все, что имеем в виду, оставляя вместо этого большую часть нашего языкового послания в контексте, потому что разобрать в одной фразе все нюансы того, что мы имели в виду, потребовало бы бесконечного времени.

Внутреннее знание сильно влияет и на наши моральные суждения. Например, большинство людей считают, что убивать — неправильно, и все же многие из нас полагают, что есть исключительные случаи, позволяющие совершать убийство, например на войне и при самообороне (а кое-кто оправдывает и убийства из мести). Если я скажу вам вне всякого контекста, что Джон Доу [44] убил Тома Крепыша, вы решите, что это явное преступление.

Но если вы увидите, как Доу убивает Крепыша в контексте голливудского фильма, в котором Крепыш сначала перестрелял семью Доу просто так, чтобы выместить на ком-то раздражение, вы, вероятно, будете обрадованы, когда Доу нажмет на курок, совершая акт отмщения. Красть и грабить — тоже неправильно, но кто из нас не считал Робина Гуда крутым героем? То, как мы понимаем вещи, редко является чистым следствием информации, получаемой извне (убийство или кража произошли) в отрыве от прочих знаний. Фактически наше восприятие всегда является комбинацией внешних данных и более абстрактных принципов внутреннего (и часто более генерализованного) характера. Поиск способа интеграции внешней информации и внутреннего знания исключительно важен для разработки более совершенного искусственного интеллекта, но пока что об этом вспоминают редко.

Теории всегда базируются на общих представлениях о мире

Согласно англоязычной «Википедии», «четверть» (или «четвертак» в разговорном языке) — как единица американской валюты — это «монета США стоимостью 25 центов, около дюйма в диаметре». Пицца (по данным того же источника) — «ароматное блюдо итальянского происхождения». Большинство пицц имеют форму диска, некоторые выглядят как плоские прямоугольники, а третьи, менее распространенные, могут быть овальными или даже более необычными по форме. Диаметр круглой пиццы обычно составляет от шести до восемнадцати дюймов. Все это, однако, по остроумному замечанию Лэнса Рипса, специалиста по когнитивной психологии из Северо-Западного университета, ничуть не мешает вам представить (и, возможно, даже съесть в качестве аппетитной закуски) пиццу, диаметр которой был точно таким же, как у американского «четвертака». Но, с другой стороны, вы никогда бы не приняли в качестве законной валюты копию четверти доллара, которая по диаметру была бы наполовину больше стандартного размера: кто угодно сочтет такую монету низкопробной подделкой (рис. 6.8).

Рис. 6.8. Платим за пиццу размером с четвертак четвертаком размером с пиццу

Отчасти это происходит потому, что мы пользуемся различными интуитивными моделями для оценки пригодности денег и еды. Типичная

денежная модель говорит, что мы готовы обменять ценные физические объекты (например, продукты питания) на маркеры ценности (например, монеты и банкноты), которые обозначают ценность в абстрактной форме, однако такой обмен должен дополнительно опираться на маркеры законности. Внешняя сторона легитимности денежных знаков основывается на признаках, которые назначаются и создаются специальными государственными органами, например монетным двором, и мы оцениваем достоверность получаемых нами денег по точному соответствию их признаков требованиям финансовых учреждений. В частности, эти требования говорят о том, что четвертак не может быть размером с пиццу.

В какой-то момент психологи и философы пытались выявить понятия, которыми оперирует наше сознание строго в терминах необходимых и достаточных условий: так, квадрат должен иметь четыре стороны одинаковой длины и четыре угла, равные 90°, прямая — это кратчайшее расстояние между двумя точками. Все, что соответствует подобным критериям, отвечает определению объекта; все, что не соответствует, дает нам иной объект или группу объектов. (Например, если две стороны прямоугольника не равны, у нас больше нет квадрата, хотя все еще есть прямоугольник в общем виде.) Для объектов алгебры и геометрии такой подход вполне адекватен, однако ученые изо всех сил пытались таким же образом определить понятия, которые были куда менее математизированными. Попробуйте, например, дать математически точное определение птицы или стула.

Существует другой подход, который заключается в рассмотрении конкретных случаев: либо через архетипы (скажем, малиновку можно назначить архетипом птицы), либо через целый набор примеров (например, можно усреднить в одном образе всех птиц, которых вы когда-либо встречали). С 1980-х годов многие придерживаются гипотезы (которую разделяем и мы), согласно которой любая теория содержит в себе определенную концепцию. Похоже, что наш мозг неплохо справляется с отслеживанием как отдельных примеров, так и архетипов, но мы также можем рассуждать и о концепциях (моделях оценки), относящихся к теориям, в которые они встроены, — примером этого служит рассуждение про деньги и пиццы, приведенное выше. В качестве другого примера мы можем дать определение любого живого объекта как чего-то обладающего «внутренней сущностью», независимой практически от всех его свойств, доступных для восприятия.

В одном из классических экспериментов йельский психолог Фрэнк Кейл спрашивал детей, может ли енот, которому изменили внешность так, чтобы он выглядел как скунс, в сочетании со «сверхвонючим» материалом стать настоящим скунсом. Дети высказали в ответ убежденность, что енот все равно остается енотом, несмотря на измененные перцептивные (внешность) и функциональные (запах) свойства. Предположительно, это является следствием применения заложенной в нас биологической теории, в которую встроено представление о том, что только внутренняя сущность действительно имеет значение при классификации организмов. (Что особенно важно, контрольный эксперимент показал: дети не распространяют ту же теорию на

искусственно созданные предметы. Например, турку для кофе, которую превратили путем добавления некоторых частей в кормушку для птиц, по мнению детей, нужно воспринимать именно как кормушку, а не как турку.)

Мы считаем, что концепции, встроенные в теории, позволяют людям обучаться очень эффективно. Предположим, что дошкольник впервые видит фотографию игуаны. Почти сразу же после этого ребенок сможет с достаточной точностью распознать не только другие фотографии игуан, но и игуан, заснятых на видео, или игуан в реальной жизни, легко отличая их от кенгуру и, возможно, даже от некоторых других ящериц. Кроме того, ребенок, обладая общими знаниями о животных, сможет сделать вывод о том, что игуаны едят и дышат; что они рождаются маленькими, растут, размножаются и умирают; и что, вероятно, существует группа (популяция) игуан, которые выглядят более или менее похожими и ведут себя одинаково. Никакие факты в мире для нас не остаются изолированными. Чтобы добиться истинного успеха в обучении, универсальный искусственный интеллект должен приобрести возможность встраивать полученные факты в насыщенные другими фактами более всеобъемлющие теории, которые помогают оценить и упорядочить новые факты.

Причинно-следственные отношения — фундаментальный аспект понимания мира

Как подчеркивал лауреат премии Тьюринга Джуда Перл, универсальным и неотъемлемым аспектом человеческого познания является глубокое понимание причинно-следственных связей. Если бы мир был прост и мы бы знали о нем всё, возможно, единственной наукой, в которой мы бы нуждались, была бы физика. Мы могли бы определить, что на что влияет, просто запустив соответствующую симуляцию: если я приложу к данному объекту силу, равную стольким-то ньютонам, что произойдет дальше? Однако ниже мы убедимся на многих примерах, что такого рода детальное моделирование в реальном мире часто нереально — в нем слишком много объектов, которые нужно отследить, и при этом слишком мало времени на это.

Вместо точных моделей мы часто используем аппроксимации; мы, как правило, знаем, что те или иные вещи причинно связаны, даже если мы не знаем точно как и почему. Мы принимаем аспирин, потому что знаем, что он поможет нам чувствовать себя лучше, и для этого нам не нужно понимать биохимию. Большинство взрослых в курсе, что секс может привести к рождению детей, даже если они не понимают точную механику зачатия и эмбриогенеза; более того, они могут использовать эти неполные знания для того, чтобы обзавестись потомством. Вам не нужно быть врачом, чтобы знать, что витамин С может предотвратить цингу, или инженером-механиком, чтобы быть в курсе, что нажатие педали газа заставляет автомобиль двигаться быстрее. Причинное знание есть повсюду, и оно лежит в основе большей части того, что мы делаем.

В классическом фильме Лоуренса Кэздана «Большое разочарование» один из персонажей по имени Джефф Голдблюм шутит, что поиск ответов для человека даже важнее, чем секс. («Вы хоть неделю в жизни провели, не пытаясь ничего рационализировать?» — спрашивает он.) Причинные выводы еще важнее, чем рационализация; без них мы бы просто не поняли мир. Не то что неделю, мы даже час не можем провести без выяснения причин. Как и Джуда Перл, мы полагаем, что для развития искусственного интеллекта нет в принципе темы важнее, о чем сейчас совершенно забыли разработчики. Сам Джуда создал в этой области мощную математическую теорию, но нам еще многое предстоит узнать о том, как отдельному человеку удается распознать и запомнить большинство известных людям причинно-следственных связей.

Тернистым путь к пониманию этого выглядит потому, что изобилует нерешенными проблемами. Почти все известные нам причины вызывают появление корреляционных зависимостей (так, автомобили действительно имеют тенденцию ехать быстрее, когда вы нажимаете педаль газа, при условии, конечно, что двигатель работает, а ручной тормоз отпущен), однако очень многие связи не имеют объяснимых причин. Крик петуха довольно надежно предвещает рассвет; но для любого человека очевидно, что, заткнув рот петуху, вы не остановите восход солнца. Показания стрелки барометра тесно коррелируют с давлением воздуха, однако изменение показаний барометра, сделанное вручную, не изменит давления воздуха.

При наличии времени и желания легко выявить множество чрезвычайно тесных корреляций из самых разных областей знания. Классический пример этого можно найти, скажем, у Тайлера Вигена: корреляция за 2000–2009 годы между потреблением сыра на душу населения и числом смертельных трагедий, произошедших вследствие того, что человек запутался в спальных простынях (рис. 6.9).

Рис. 6.9. Типичный пример ложной корреляции

Будучи студентом юридического факультета, Тайлер написал целую книгу под названием «Ложные корреляции» (Spurious Correlations). Как отмечает Виген, в тот же период времени (2000–2009 годы) количество людей, утонувших в результате случайного падения в бассейн, было тесно связано с количеством

фильмов, в которых появлялся актер Николас Кейдж. Создание машины, которая могла бы отличить ложные, без подлинной причинной связи, корреляции от реальных, таких как корреляция между нажатием педали газа и ускорением автомобиля, стало бы колоссальным достижением в области искусственного интеллекта [45].

Мы никогда не теряем из виду конкретные объекты и конкретных людей

В нашей повседневной жизни мы постоянно отслеживаем различные категории индивидуальных объектов вместе с их историей и свойствами. Вот чей-то супруг, он раньше работал журналистом и предпочитает пить бренди, а не виски. Вот чья-то дочь, она в детстве боялась грозы и предпочитает мороженое печенькам. У вашей машины есть вмятина на задней правой двери, и вы заменили ей коробку передач год назад. Аптека на углу улицы когда-то продавала хорошие товары, но с тех пор, как у нее сменилось руководство, уровень качества пошел вниз. Наш опыт состоит из множества отдельных вещей, которые существуют долго и изменяются с течением времени, причем многое из того, что мы знаем, сгруппировано определенным образом вокруг этих вещей. Речь идет не только об автомобилях, людях и магазинах в целом, но и об отдельных представителях каждой из этих групп с учетом их особенностей и индивидуальной истории.

Как ни удивительно, подобный подход вообще неестественен для систем глубокого обучения. Этот тип машинного интеллекта всегда сосредоточен не на особенностях конкретных людей или предметов, а на более общих категориях. Большинство систем глубокого обучения хороши в генерализации: они позволяют заключить, что дети в основном предпочитают сладости, а не овощи, машины имеют четыре колеса и т.д. Именно такие статистические факты системы глубокого обучения считают естественными, в то время как факты, относящиеся конкретно к вашей дочери или вашей машине, они могут полностью упускать из виду.

Конечно, встречаются и определенные исключения, но если вы посмотрите внимательно, то поймете, что они как раз подтверждают правило. Например, системы глубокого обучения можно настроить так, чтобы они очень хорошо идентифицировали изображения конкретных людей. Вы можете, скажем, обучить такую систему с высокой точностью распознавать изображения Дерека Джитера [46]. Но это возможно лишь потому, что система рассматривает изображения Дерека Джитера как «категорию похожих изображений», а не потому, что она имеет какое-либо представление о Дереке Джитере как о спортсмене или личности. Механизмы, используемые программами глубокого обучения для совершенствования в распознавании конкретного человека, остаются точно такими же, как и при обучении распознаванию любых иных объектов, например бейсболистов: и то и другое является для них не более чем категориями изображений. Куда проще обучить систему глубокого обучения распознавать фотографии Дерека Джитера, чем

заставить ее по серии новостей за 20 лет (с 1995 по 2014 годы) сделать вывод о том, что в течение всего этого времени Джитер играл на позиции шорт-стопа в клубе Yankees.

Точно так же совсем нетрудно разработать систему глубокого обучения для довольно точного отслеживания конкретного человека на видеозаписи. Но для любой такой программы задача все равно сведется к тому, чтобы ассоциировать группу пикселей в одном видеокадре с аналогичным фрагментом пикселей в следующем кадре; никакого глубокого понимания сущности конкретной личности (или людей вообще) за всем этим не стоит. Алгоритмы глубокого обучения понятия не имеют о том, что, когда человек не находится в кадре видео, он или она пребывает где-то еще, а не исчезает насовсем. И система совсем не удивится, если в пустую телефонную будку зайдет один человек, а выйдут оттуда сразу двое.

Существа, обладающие сложными когнитивными способностями, никогда нельзя рассматривать в качестве чистого листа

В 1865 году исследования Грегора Менделя привели его к выводу, что носителями наследственности являются некоторые реальные объекты; он назвал их «факторами», а мы сейчас называем генами. Неизвестной для Менделя осталась материальная сущность этих «факторов»; прошло почти 80 лет, прежде чем ученые нашли ответ на этот вопрос. В течение десятилетий исследователи постоянно заходили в тупик, ошибочно предполагая, что «факторы» Менделя — это еще не известные науке белки, хотя кое-кто подозревал, что они могут представлять собой не белки, а нуклеиновые кислоты. Только в 1944 году Освальд Эвери методом исключения доказал, что жизненно важную роль для передачи генетической информации играет ДНК. Но даже тогда большинство коллег Эвери попросту не обратили внимания на его открытие, потому что в то время научное сообщество вообще не очень интересовалось нуклеиновыми кислотами. Работы самого Менделя тоже почти сразу были забыты его современниками, пока эти законы не открыли заново в 1900 году.

Современные разработчики искусственного интеллекта могут так же пройти мимо очевидных вещей, когда дело доходит до старого вопроса о врожденности знаний. В области естественных наук он часто формулируется в форме «nature versus nurture» (то есть «природа против воспитания»). Какая часть наших знаний по сути встроена в мозг и как много нового мы усваиваем в течение жизни? Точно такие же вопросы возникают и в случае искусственного интеллекта: все ли должно быть встроено в него исходно? Или он должен всему научиться сам?

Любой, кто всерьез задумался над этим вопросом, поймет, что дихотомия в этом вопросе явно надумана. Данные биологии — во всяком случае, из таких ее областей, как психология развития (которая изучает развитие детей) и

нейробиология развития (в настоящее время изучает связь между генами и развитием мозга) говорят почти об одном и том же: генетика и воспитание работают вместе, и противопоставлять их друг другу неразумно. Как подчеркивал Гэри в своей книге «Рождение разума» (The Birth of the Mind), индивидуальные гены фактически являются рычагами этого сотрудничества. (Точнее, в этой книге говорится о том, что каждый ген — это что-то вроде выражения IF— THEN в компьютерной программе. Оператор THEN определяет, какой конкретный белок должен быть построен, но тем не менее этот белок создается только при наличии определенных химических сигналов, причем каждый ген задает только ему одному свойственные значения оператора IF. Результат этого взаимодействия подобен адаптивному обучению, но в то же время он состоит из сильно сжатого набора компьютерных программ, выполняемых автономно отдельными ячейками в ответ на условия среды. Из всего этого «ирландского рагу» и создается знание.) Как ни странно, большинству исследователей в области машинного обучения эти аспекты познания, взятые из мира живых и мыслящих существ, судя по всему, совершенно не интересны [47].

Статьи по машинному обучению очень редко принимают к сведению обширную литературу, накопившуюся в области психологии развития, и когда они это делают, то ссылаются разве что на исследования Жана Пиаже, который был признанным пионером в этой области, но умер почти 40 лет назад. Вопросы, задаваемые Пиаже, такие как «Знает ли ребенок, что объекты продолжают существовать даже после того, как они ушли из поля зрения?», все еще остаются совершенно правильными, но ответы, которые он предлагал в свое время, в частности его теория этапов когнитивного развития и его предположения о возрасте, в котором дети открывают различные предметы и сущности, основывались на устаревших методологиях, которые не выдержали испытания временем.

Редко можно встретить какое-либо исследование по психологии развития за последние два десятилетия, цитируемое в литературе по машинному обучению, и еще реже можно увидеть что-либо о генетике или нейробиологии развития. Люди, получающие образование в сфере машинного обучения, в большинстве своем делают упор на алгоритмы, помогающие усваивать новую информацию, но совершенно игнорируют ценность врожденных знаний. Они, кажется, всерьез верят в то, что, раз они изучают обучение, ничто врожденное не имеет ценности. Но природа и воспитание на самом деле не конкурируют друг с другом; во всяком случае, чем больше информации содержит ваша отправная точка, тем больше вы можете узнать в дальнейшем. Тем не менее в глубоком обучении явно доминирует принцип чистого листа, который пренебрегает какими-либо предшествующими знаниями независимо от их важности [48].

Мы ожидаем, что, оглянувшись однажды назад, исследователи поймут, каким колоссальным недосмотром обернулся этот подход. Мы, конечно же, не отрицаем важность обучения на опыте: его роль очевидна даже для тех из нас, кто признает ценность врожденных знаний. Тем не менее обучение с

абсолютно чистого листа, о котором мечтают апологеты машинного обучения, делает разработку более совершенного искусственного интеллекта намного сложнее, чем она могла бы быть. Откуда берется такое стремление к «чистоте», когда наиболее эффективным решением, как подсказывает нам природа, является объединение двух подходов?..

Как полагает Элизабет Спелк, специалист по психологии развития из Гарварда, люди, вероятно, уже рождаются с пониманием того, что мир состоит из устойчивых объектов, которые существуют в пространстве и времени, перемещаясь сразу по всем измерениям. Ощущение геометрических и количественных соотношений у людей тоже может оказаться врожденным, как и основы интуитивной психологии. Или, как утверждал Кант двумя веками ранее в философском контексте, для правильного понимания мира необходимо врожденное «пространственно-временное многообразие».

Представляется весьма вероятным, что некоторые аспекты языка также присутствуют у нас врожденно. Иначе говоря, дети появляются на свет сразу с пониманием того, что звуки или жесты, которые производят окружающие, являются информационными каналами, несущими смысл. Это понимание объединяется с другими врожденными знаниями о человеческих отношениях (мама позаботится обо мне и т.д.). Другие аспекты языка также могут оказаться изначально заложенными в нашем сознании, в частности, сюда относится разделение языка на предложения и слова, понимание того, что язык имеет синтаксическую структуру и что синтаксические отношения имеют связи с семантическими понятиями. Даже ожидания ребенка относительно звучания языка могут проявляться в качестве врожденного свойства.

В отличие от этого, системе, действительно обучаемой с нуля, которая сталкивается с миром как с чистым аудиовизуальным потоком, придется заново узнать буквально все на свете, начиная с того, что в мире существуют личности с устойчивыми свойствами. Несколько человек пытались смоделировать нечто подобное в самом общем случае (в том числе в DeepMind), и результаты оказались далеко не столь впечатляющими, как тот же подход в более узком применении, касавшемся исключительно настольных игр.

Многие, кто работает в области машинного обучения, считают, что встраивание в систему любых начальных знаний равносильно своего рода мошенничеству и что более честными и впечатляющими решениями будут те, когда в нейронной сети изначально присутствует как можно меньше исходной информации. Большая часть самых известных ранних проектов DeepMind, по-видимому, опиралась именно на такую философию. Их игровая система Atari выстроена практически ни на чем более, кроме общей архитектуры глубокого обучения с подкреплением функций, которые представляли собой операции с джойстиком, пикселей экрана и общей системы оценки результата. Даже сами правила игры система должна была осваивать исключительно на опыте, включая и любые аспекты игровой стратегии.

В своей более поздней статье, опубликованной в журнале *Nature*, инженеры и программисты DeepMind утверждали, что машина освоила игру в го «без

человеческих знаний». Хотя DeepMind, безусловно, использовали меньше человеческих знаний о го, чем их предшественники, фраза «без человеческих знаний» (использованная прямо в названии статьи) была явным преувеличением: система все еще в значительной степени опиралась на то, что исследователи-люди выяснили за последние несколько десятилетий о том, как научить машины играть в интеллектуальные игры, подобные го и шахматам. Прежде всего — это симуляция по методу Монте-Карло, описанный нами выше метод случайной выборки из вероятностного древа различных игровых возможностей: ничего общего с глубоким обучением данный подход не имеет. К тому же (в отличие от их ранее широко обсуждаемой работы над играми Atari) разработчики из DeepMind исходно встроили в новую систему правила игры в го и некоторые другие (весьма подробные) знания об этой игре. Следовательно, их утверждение, что человеческие знания в принципе не были задействованы, просто не соответствовало действительности.

Не менее важно, что само это заявление в упомянутой публикации ясно продемонстрировало современные ценности, на которые ориентируется вся область искусственного интеллекта: необходимо устранить предшествующее знание, а не прикладывать усилия к его использованию. Это звучит примерно так же, как если бы производители автомобилей настаивали на том, что круглые колеса необходимо открыть заново, а не просто продолжать их использовать, основываясь на огромном опыте двух тысячелетий по производству колесных экипажей. Мы полагаем, что реальное продвижение в области искусственного интеллекта начнется с понимания того, какие виды знаний и представлений о мире должны быть встроены в машинные системы еще до обучения, чтобы помочь им освоить остальное.

Вместо попыток создать системы, которые изучают все на основе взаимосвязей между пикселями и действиями, мы как сообщество разработчиков искусственного интеллекта должны научиться создавать системы, которые используют для изучения мира базовое понимание физических объектов. Многое из того, что мы называем здравым смыслом, представляет собой приобретенное знание, например что в кошельках обычно держат деньги или что сыр может быть кусковым и тертым, но даже такие знания почти всегда опираются как минимум на твердое ощущение времени, пространства и причинности. В основе этих базовых ощущений может лежать врожденный механизм представления абстракции, композиционности и свойств отдельных сущностей, таких как объекты и люди, которые существуют в течение некоторого периода времени (измеряемого в минутах или десятилетиях). Машинам необходимо с самого начала привить те же самые основополагающие принципы, чтобы они имели возможность научиться всему остальному [49].

В недавнем открытом письме, адресованном специалистам в области искусственного интеллекта, руководитель программы информатики Калифорнийского университета в Лос-Анджелесе Аднан Дарвиче призвал к внедрению более широкого образования в сообщества ИИ-программистов и

разработчиков такими словами: «Нам нужно новое поколение исследователей искусственного интеллекта, которые хорошо разбираются и заинтересованы в классических подходах, машинном обучении и информатике в самом широком смысле, а также более информированы об истории искусственного интеллекта».

Рис. 6.10. Не стоит пилить сук, на котором сидишь

Нам хотелось бы еще расширить эту точку зрения и сказать, что исследователи искусственного интеллекта должны опираться не только на большой вклад компьютерных наук, часто забываемый в сегодняшнем увлечении большими данными, но также и на широкий круг других дисциплин, от психологии до лингвистики и нейробиологии. История развития всех этих областей, составляющих когнитивные науки, и сделанные в них открытия могут многое рассказать нам о том, как биологические существа формируются для выполнения сложных интеллектуальных задач. И если искусственный интеллект должен стать чем-то вроде машинной версии естественного интеллекта, то нам нужно научиться создавать высокоструктурированные гибридные системы, исходно обладающие врожденными знаниями и способностями, воспринимающие и транслирующие знания композиционно и способные хранить информацию и ассоциации с конкретными (а не

статистически усредненными) объектами и личностями. Все это доступно каждому человеку, включая маленьких детей.

Когда искусственный интеллект сможет наконец воспользоваться уроками когнитивных наук, перейдя от парадигмы, вращающейся исключительно вокруг больших данных, к парадигме, опирающийся и на массовые данные, и на абстрактные причинно-следственные связи, мы наконец-то сможем приступить к решению одной из самых сложных задач на свете: как наделить машины здравым смыслом (рис. 6.10).

ГЛАВА 7

Здравый смысл и путь к глубокому пониманию

Предметом сегодняшнего исследования являются вещи, которые не могут двигаться сами по себе.
Им нужно все время помогать: толкать, передвигать, брать с одного места и ставить на другое.
Многие из них совсем не хотят ходить, например книжные полки, шкафы, неподатливые стены, столы.
А вот скатерть, лежащая на упрямом столе, — если ее хорошенько схватить за свисающие концы — выказывает явное желание попутешествовать.
И стоящие на ней бокалы, тарелки, соусники, ложки, миски при этом прямо-таки трясутся от нетерпения отправиться в путь.
Вислава Шимборская. Малышка и скатерть (A Little Girl Tugs at the Tablecloth)

Одного лишь знания недостаточно.
Девиз Хэмпширского колледжа

Здравый смысл — это те базовые знания, которыми, как мы ожидаем, обладают все обычные люди. Примеров множество.

- «Люди не любят терять свои деньги».
- «Вы можете хранить деньги в своем кошельке».
- «Вы можете хранить кошелек у себя в кармане».
- «Ножом можно разрезать те или иные вещи».
- «Объекты не исчезнут, если вы просто накроете их одеялом».

Мы все были бы удивлены, если бы увидели собаку, несущую слона, или как стул внезапно превращается в телевизор. Ирония здравого смысла — да и всего искусственного интеллекта — заключается в том, что все понимают, какие явления относятся к здравому смыслу, однако никто, похоже, не знает, в чем его суть и как создавать машины, которые бы им обладали.

Люди пытаются решить проблему здравого смысла с самых первых дней существования искусственного интеллекта. Джон Маккарти, тот самый человек, который придумал термин «искусственный интеллект», впервые начал привлекать внимание к этой особенности человеческого мышления еще в 1959 году. Но прогресса здесь до сих пор на удивление мало. Ни

классический искусственный интеллект, ни глубокое обучение не достигли в этой сфере ничего серьезного. Глубокое обучение, в котором отсутствует прямой способ привлечения абстрактных знаний (например, «люди хотят вернуть утраченные вещи»), по большей части игнорировало это проблему вообще; исследователи классического искусственного интеллекта, наоборот, пытались ее решить, применяя то одни, то другие подходы, но так и не преуспели.

Один из подходов состоял в том, чтобы попытаться освоить базовые знания, сканируя интернет. Одна из самых обширных программ в этом направлении была запущена в 2011 году и называется NELL [50]. Проектом руководит Том Митчелл, профессор Университета Карнеги — Меллон и один из пионеров в области машинного обучения. День за днем (проект все еще продолжается) NELL отыскивает в интернете документы и читает их, выделяя конкретные лингвистические паттерны и расшифровывая, что они могут значить. Если он видит словосочетание типа «города, такие как Нью-Йорк, Париж и Берлин», то предполагает, что Нью-Йорк, Париж и Берлин — это разные города, и добавляет слово «город» в свою базу данных. Если он увидит фразу «квотербек [команды] "Нью-Йорк Джетс" Келлен Клеменс», он может сделать вывод о том, что Келлен Клеменс играет за «Нью-Йорк Джетс» (в настоящем времени, поскольку у NELL нет чувства времени) и что Келлен Клеменс является квотербеком.

Как бы разумна ни была основная идея, результаты ее применения оказались более чем скромными. В качестве примера перечислим ниже десять фактов, которые NELL узнала некоторое время назад [33].

- Агрессивные_собаки относятся к млекопитающим.
- Прилагательное «узбекский» можно соотнести со словом «язык».
- Рецепты_кофейных_напитков относятся к категории объектов, называемых рецептами.
- Рошель_Иллинойс — это остров.
- Станция_Симо-Китадзава — это высотное здание.
- Стивен_Хокинг — это человек, который посещал школу в Кембридже.
- Хлопок — сельскохозяйственный_продукт, растущий в Гуджарате.
- Келлен_Клеменс играет в лиге НФЛ.
- N24_17 [именно так!] и Давид и Лорд — это братья и сестры.
- В городе Сент_Джулианс говорят на английском языке.

Некоторые из этих утверждений верны, некоторые ложны, некоторые бессмысленны; и в целом все они не особенно полезны. Они не помогут роботам управляться на кухне, и хотя, возможно, часть из них пригодится в развитии машинного чтения, все равно они остаются слишком разрозненными и неуверенными, чтобы вселить в компьютеры даже подобие здравого смысла.

Другой (очень модный в наши дни) метод сбора общеизвестных знаний, основанных на здравом смысле, заключается в использовании краудсорсинга; по своей сути это означает обращаться за помощью к обычным людям — носителям здравого смысла. Наиболее выделяющимся проектом данного направления является, пожалуй, ConceptNet, который реализуется в

Массачусетском технологическом институте (лаборатория Media Lab) уже с 1999 года. В рамках этой инициативы поддерживается веб-сайт, на котором добровольцам предлагается вводить простые факты на английском языке. Например, участника могут попросить предоставить факты, которые будут значимы для понимания такой простейшей ситуации: «Боб простудился, Боб пошел к врачу»; ответы на этот запрос включают в себя, например, следующие утверждения: «Люди с простудой чихают» или «Вы можете помочь больному человеку с помощью лекарств». Затем в процессе сопоставления с образцом английские предложения автоматически преобразуются в машинные коды.

Здесь основная идея тоже кажется разумной, но результаты опять неутешительны. Одна из проблем заключается в том, что если вы просто попросите неподготовленных людей перечислить очевидные для них факты, они склонны будут давать в ответ что-то вроде «Утконос — это млекопитающее, которое откладывает яйца» или «Отбой — это сигнал горна, подаваемый вечером». Все подобные факты компьютер сможет легко найти и сам; требуется же нечто совсем другое, а именно информация, которая очевидна для людей, но которую трудно найти в интернете, например «После того как что-то умерло, оно никогда не будет снова живым» или «Непроницаемый контейнер с отверстием сверху и больше нигде способен удерживать жидкости внутри себя».

Вторая проблема заключается в том, что, даже если вы сможете уговорить людей, не обладающих образом мышления как у программистов, дать вам нужную информацию, почти невозможно будет заставить их формулировать фразы таким хитрым сверхточным способом, который требуется компьютерам. Вот, например, часть того, что ConceptNet узнала от самых обычных людей о ресторанах и смогла интегрировать в логическую схему (рис. 7.1).

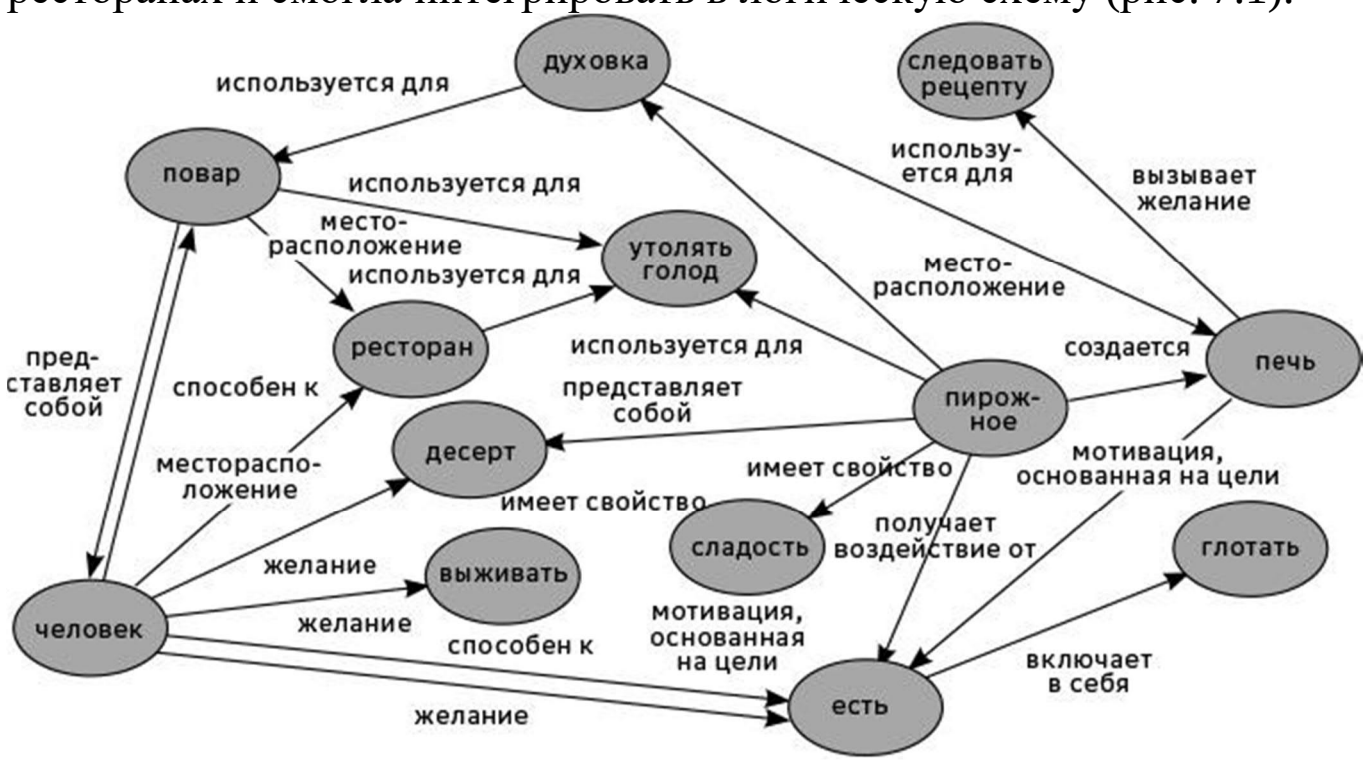


Рис. 7.1. Фрагмент схемы из ConceptNet

На первый взгляд она кажется совершенно нормальной. Каждая отдельная ссылка (например, стрелка в верхнем левом углу, указывающая на то, что для

приготовления пищи используется духовка) сама по себе кажется правдоподобной. Человек может находиться в ресторане, и почти каждый человек, которого мы когда-либо встречали, хочет выжить; никто не усомнится в том, что нам нужно есть, чтобы выжить. Но стоит нам погрузиться в детали, мы обнаружим в них полный беспорядок.

Возьмем для примера связь, в которой говорится, что существо, именуемое «человек», находится в месте, называемом «ресторан». Как уже давно отметил Дрю Мак-Дермотт — учитель Эрни — в своей широко известной статье, озаглавленной «Искусственный интеллект обнаруживает свою природную глупость», значение связи, отображаемой этой стрелкой, на самом деле неясно. В любой момент кто-то в мире находится в ресторане, но многих конкретных людей там нет. Означает ли данная связь, что если вы ищете конкретного человека (скажем, вашу мать), то вы всегда можете найти ее в ресторане? Или что в каком-то конкретном ресторане (скажем, Katz's Delicatessen) вы сможете найти людей 24 часа в сутки? Или что любого человека, которого вам потребуется найти, вы всегда найдете в ресторане, подобно тому как китов всегда можно найти в океане? Другая связь говорит нам, что «торт используется для утоления голода». Может быть, это и правда, но остерегайтесь другой связки, которая гласит «повар используется для утоления голода» в сочетании со связкой «повар представляет собой человека». Фактически эта комбинация связок говорит не только о том, что повар в состоянии приготовить еду, но и о том, что он сам может быть едой. Мы вовсе не утверждаем, что краудсорсинг бесполезен в принципе, но на сегодняшний день проекты с участием неподготовленной публики часто дают на выходе информацию, которая оказывается на поверку запутанной, неполной или даже вовсе неверной.

Рассмотрим еще один, более поздний проект, также запущенный в Массачусетском технологическом институте, но осуществляемый другой научной группой; он называется VirtualHome (Виртуальное жилище) [34]. Для сбора информации о действиях, необходимых для совершения ряда простейших задач (таких как помещение продуктов в холодильник и сервировка стола), в этом проекте также задействовали краудсорсинг. Исполнители собрали в общей сложности 2800 действий для 500 задач, включающих 300 объектов и 2700 типов взаимодействий (все это отдаленно напоминает работу Шанка над сценариями, но здесь процедуры структурированы менее формально). Затем все базовые действия были воссозданы в одном из игровых движков, благодаря чему стало возможным (в ряде случаев) увидеть анимацию выполнения задачи. Но и в этот раз результаты оставляют желать лучшего. Рассмотрим, например, одну такую задачу, названную «Тренировка».

1. Пройдите в гостиную.
2. Найдите пульт.
3. Возьмите пульт.
4. Найдите телевизор.
5. Включите телевизор.

6. Верните на место пульт.
7. Найдите пол.
8. Лягте на пол.
9. Посмотрите в телевизор.
10. Найдите обе_руки.
11. Вытяните обе_руки.
12. Найдите обе_ноги.
13. Вытяните обе_ноги.
14. Встаньте.
15. Прыгайте.

Описанные действия могут иметь место в жизни некоторых людей, но далеко не у всех [35]. Часть из нас предпочитает ходить в спортзалы или бегать на улице; другие любят плавать, третьим нравится поднимать тяжести. Какие-то из перечисленных шагов реальные люди пропускают ради простоты, другие просто не требуются постоянно, ну и, разумеется, пара упражнений — это еще не тренировка. Вместе с тем поиск пульта дистанционного управления и включение телевизора вообще совершенно не обязательно должны быть частью тренировки, не говоря уже о том, что человеку явно не нужна команда найти свои руки или ноги. Во всем этом есть что-то крайне неестественное.

Еще один подход состоит в том, чтобы все то же самое попытались записать высококвалифицированные специалисты в такой форме, которую компьютер смог бы интерпретировать однозначно. Многие теоретики искусственного интеллекта, начиная с Джона Маккарти и заканчивая Эрни и многими его коллегами, такими как Гектор Левек, Джо Халперн, Джерри Хоббс, надеялись, что такой способ окажется гораздо более эффективным.

Но, честно говоря, и здесь, на нашей родной территории, прогресс шел куда медленнее, чем мы рассчитывали. Работа была кропотливой и сложной [36], поскольку опиралась на тщательный анализ, который до сих пор невозможно автоматизировать. Хотя некоторые важные результаты действительно были достигнуты, мы все еще никоим образом не близки к тому, чтобы целиком закодировать принципы и факты здравого смысла, а без этого (или чего-то подобного этому) главные функции искусственного интеллекта следующего уровня, такие как автоматическое чтение и автономное механизированное выполнение различных бытовых работ, так и будут оставаться вне нашей досягаемости.

Самым крупным достижением в этой области [37], безусловно, является проект, известный как CYC и выполняемый под руководством Дуга Лената в течение последних трех десятилетий, нацеленный на создание огромной базы данных человеческого здравого смысла, представленной в машинно-интерпретируемой форме. Он содержит буквально миллионы тщательно закодированных фактов обо всем, от терроризма до медицинских технологий и семейных отношений, которые были тщательно подготовлены группой людей, хорошо образованных в сфере как искусственного интеллекта, так и философии.

Большинство исследователей из внешнего мира рассматривают данный проект как очередную неудачу: слишком мало было опубликовано сообщений о том, что происходит внутри него (проект проходил в основном без широкой огласки). Во всяком случае, пока что по отношению к затраченным усилиям мы видим слишком мало демонстраций того, на что по-настоящему способна эта база данных. Статьи, написанные о ней независимыми специалистами, были в значительной степени критическими, и очень немногие разработчики интегрировали новые данные в более крупные системы. Мы считаем, что цели проекта достойны восхищения, но после трех десятилетий CYC все еще остается недостаточно полным, чтобы оказать достойное влияние на теорию искусственного интеллекта, несмотря на свой колоссальный масштаб и огромное количество проделанной работы. Проблема того, как получить обширную и надежную базу знаний, основанных на здравом смысле, так и остается нерешенной. Что же делать?

Хотелось бы, чтобы у нас был простой и элегантный ответ на этот вопрос, но мы такового не видим. Более того, сомнительно, что найдется какой-то единый подход, во многом как раз потому, что сам здравый смысл чрезвычайно разнообразен в своих проявлениях. Вряд ли даже лучший метод сможет в одиночку справиться с тем, что уже много лет остается камнем преткновения для всего нашего сообщества. Здравый смысл — вершина, которую нам придется покорять всем вместе, и здесь предстоит еще долгий путь. Попытка обойти эту гору, лишь немного отклонившись от нынешнего проторенного пути, точно не приведет нас к вершине.

Тем не менее у нас уже есть самое общее представление о том, куда следует направиться в первую очередь. Пользуясь все той же метафорой про горную вершину, заметим, что даже если мы не сможем сейчас добраться туда самостоятельно, то по крайней мере существуют способы увидеть, как выглядит наша гора сверху, какое оборудование может понадобиться, чтобы подняться туда в будущем, и какая стратегия будет для этого оптимальной.

Чтобы добиться более ощутимого прогресса, для начала нам нужны две вещи: перечень того, какие знания должен иметь универсальный искусственный интеллект, и понимание, как эти знания будут автономно представлены внутри машины, причем их формулировки должны быть однозначными и в терминах выполнения алгоритма.

Мы обсудим эти вопросы в обратном порядке, потому что нахождение четкого способа представления знаний машине является необходимым условием для кодирования любого знания, относится оно к области здравого смысла или нет. Как вы уже догадываетесь, эта задача оказывается гораздо более тонкой, чем может показаться на первый взгляд. Некоторые знания удается представить чуть ли не напрямую, но в большинстве случаев так сделать не получится.

Более легким вариантом логической репрезентации отношений между объектами является иерархическая таксономия в виде дерева, на котором сразу становятся видны уровни обобщения. Рис. 7.2, приведенный ниже, подсказывает нам, что собаки — это млекопитающие, а млекопитающие — это

животные, из чего нетрудно сделать вывод о том, что собаки являются животными. Если вы знаете (или узнаете), что Лесси — это собака, то логика требует отнести ее и к животным.

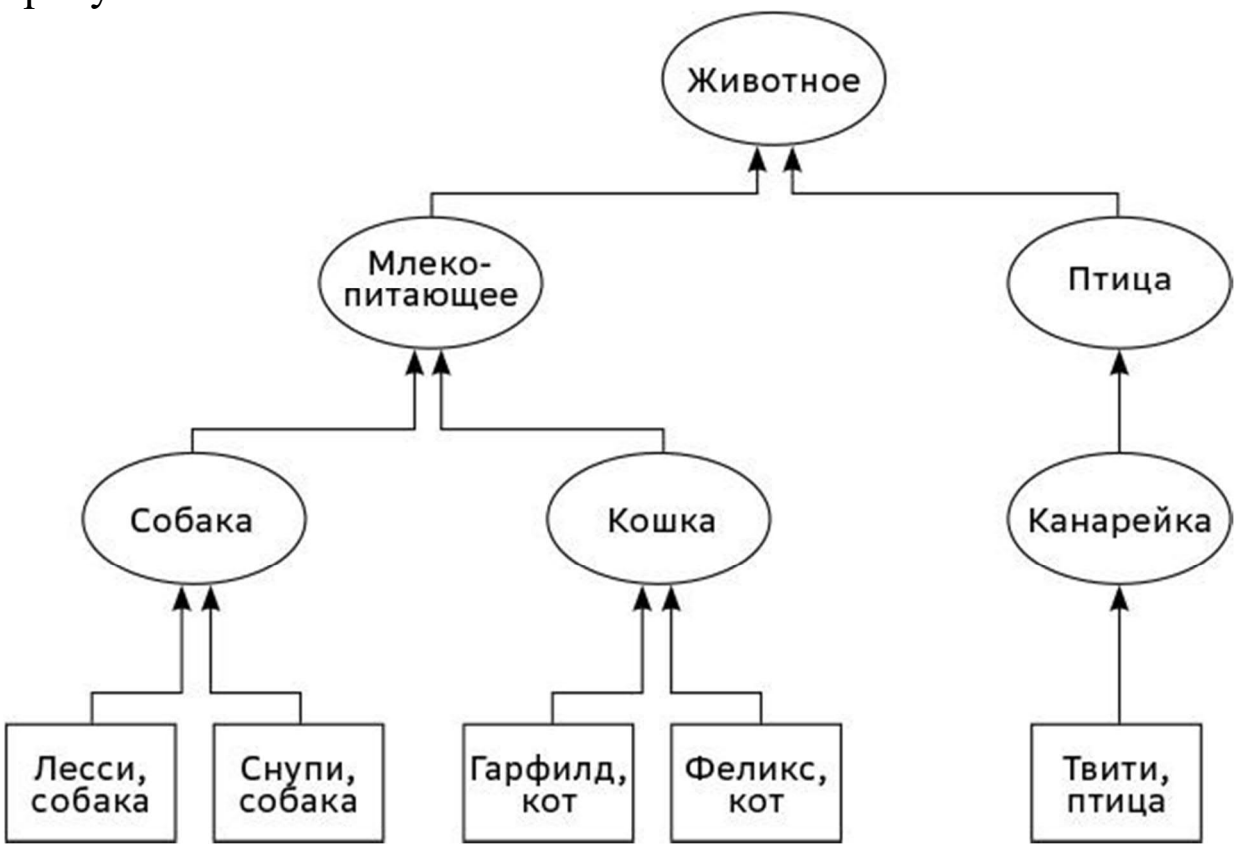


Рис. 7.2. Пример таксономического древа

Многие онлайн-ресурсы, начиная с «Википедии», включают огромное количество таксономической информации: «скунс — плотоядное животное», «сонет — форма стихотворения», «холодильник — бытовая техника». Другой инструмент — WordNet — специализированный онлайн-тезаурус, который часто используется при изучении и разработке искусственного интеллекта, содержит таксономическую информацию для конкретных слов: «зависть» — это «чувство, напоминающее обиду», «молоко» — это и «молочный продукт», и «разновидность напитка». Существуют также специализированные таксономические реестры, например SNOMED — медицинский словарь, содержащий факты в форме технических определений: «аспирин является одной из фармацевтических форм салицилата» и «желтуха является разновидностью клинических симптомов». (Онтологические онлайн-словари, широко распространенные в одной из ветвей интернета, называемой Semantic Web, тоже являются частью формализованного знания.)

Сходные технологии можно использовать и для представления отношений между целым и его частями. Если палец ноги является частью ступни, а ступня является частью тела, можно заключить, что палец является частью тела. Как только вы узнаете эти и подобные вещи, некоторые из проблем, о которых мы упоминали ранее, начнут становиться менее безнадежными. Если вы увидите, что Эллу Фицджеральд сравнивают с бутылкой винтажного вина, вы сразу обратите внимание на то, что Элла Фицджеральд — это человек, а люди — это животные и что бутылки в целом относятся к другой ветви иерархических взаимоотношений между объектами (той, где размещаются неодушевленные

предметы), а значит, вы сделаете вывод, что Элла не может быть бутылкой. Аналогично, если гость просит у робота-дворецкого принести напиток, то робот со встроенной универсальной таксономией без труда определит, что вино, пиво, виски или сок подходят к озвученной формулировке, а вот черешок сельдерея, пустой стакан, часы или анекдот к делу явно не относятся.

Но, увы, здравый смысл включает гораздо больше понятий и отношений, чем отображается в таксономии. Для большинства вещей в мире нам понадобится другой подход. Если таксономия животных еще более или менее четко определена (как следствие работы эволюционного отбора и процесса видообразования), то многие другие объекты не формируют естественной таксономии [38]. Скажем, мы хотим создать определенную категорию исторических событий с отдельными элементами, такими как «русская революция», «битва за Лексингтон», «изобретение печати» и «протестантская Реформация». Здесь границы гораздо более размыты. Было ли движение Сопротивления во Франции частью Второй мировой войны? Как насчет советского вторжения в Финляндию в 1939 году? И в более широком смысле: должна ли, например, категория объектов, содержащая автомобили и людей, включать также демократию, естественный отбор или веру в Санта-Клауса? Таксономия — неподходящий инструмент для работы со здравым смыслом. Как заметил философ Людвиг Витгенштейн, даже такую простую категорию, как «игра», определить чрезвычайно трудно.

Кроме того, существуют простые знания, которые мы упомянули в начале главы, например о том, что ножи могут резать вещи, а метлы можно использовать для уборки пола. Но подобные факты, похоже, вообще не вписываются в чистые таксономические отношения. Между тем трудно понять, как робот будет заботиться о вашем доме без такого рода знаний.

Еще один возможный подход заключается в создании диаграмм, часто называемых семантическими сетями, которые мы приводили выше в качестве иллюстрации к проекту ConceptNet. Семантические сети были изобретены еще в конце 1950-х годов; они позволяют компьютерам представлять весьма широкий спектр понятий, а не только то, какие объекты являются частями более обширных понятий и какие категории находятся внутри других категорий. Пользуясь этим методом, можно выражать довольно сложные отношения, например то, что город Олбани соседствует с Гудзоном, а полицейские принадлежат к той категории людей, которые водят полицейские машины.

Однако на примере ConceptNet мы уже убедились, что представления семантических сетей недостаточно ясны для решения описанной проблемы. Нарисовать их гораздо проще, чем заставить работать. Предположим, вы хотите закодировать несколько взаимосвязанных фактов. Вот они: у Иды iPhone, и она родилась в Бойсе; ее iPhone имеет батарею внутри, а батарея вырабатывает электроэнергию. Вы довольно быстро сможете нарисовать что-то вроде схемы ниже (рис. 7.3).

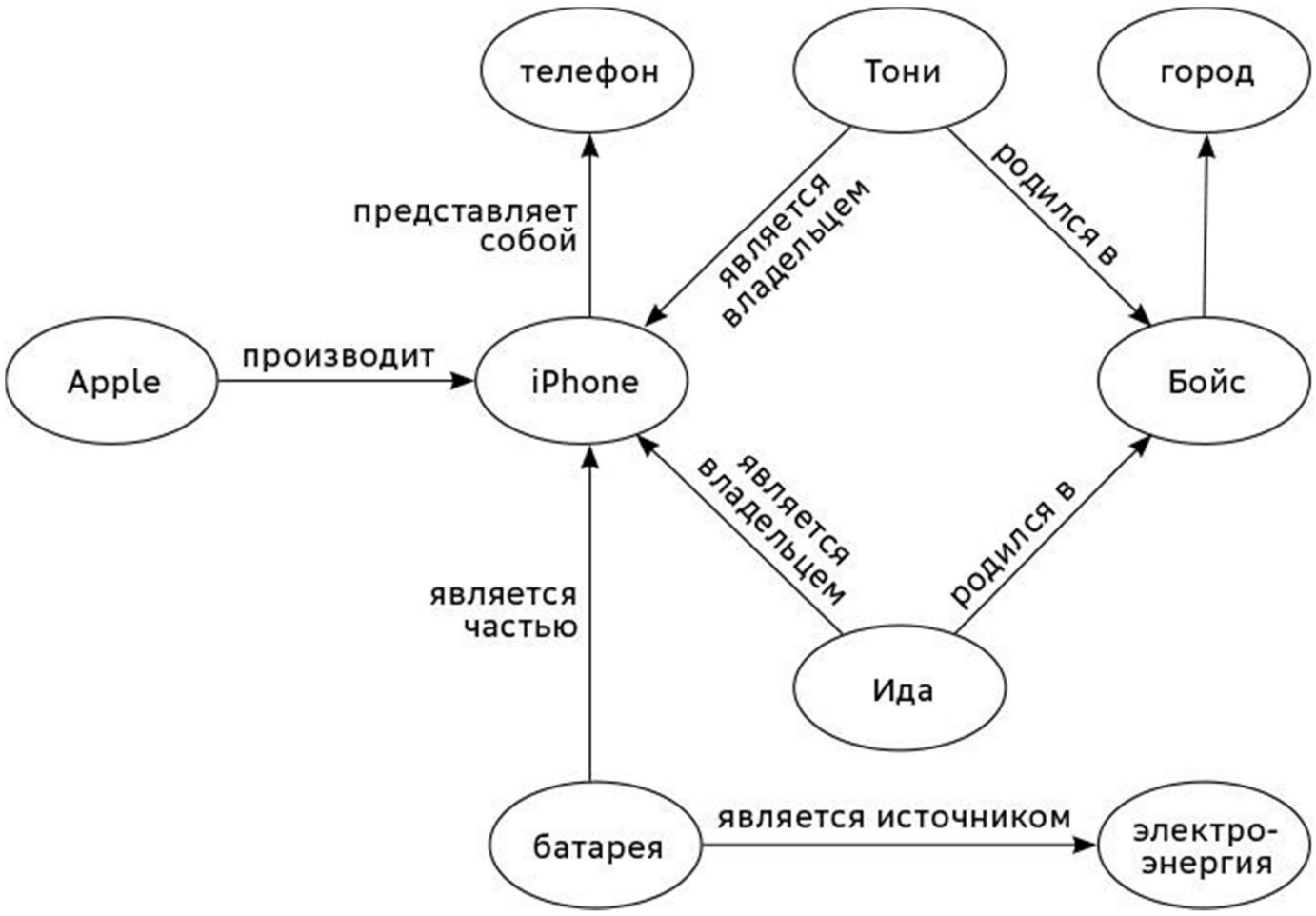


Рис. 7.3. Пример семантической сети

Проблема, однако, в том, что многое из того, что требуется знать для правильной интерпретации подобных диаграмм, не выражается явным образом (и отсутствует в самих диаграммах), а машины не умеют обращаться с тем, что не сформулировано в явном виде. Для вас очевидно, что если Тони и Ида родились в одном и том же городе Бойс, то из этого следует, что если вы отправляетесь в родной город Тони, то вы автоматически попадете и в родной город Иды. Но если мы рассмотрим iPhone, принадлежащий Тони, он, скорее всего, не будет принадлежать и Иде, однако ничто в семантической сети не позволяет понять в явной форме различие между двумя только что описанными ситуациями. Без дальнейшей работы над усовершенствованием схемы машина не сможет продолжать рассуждения в любом из множества возможных направлений.

Или рассмотрим тот факт, что все iPhone производятся компанией Apple. Если вы видите айфоны, вы можете сразу сделать вывод, что это — продукция Apple: это, по крайней мере, выходит из того, что написано в семантической сети на рисунке. Тем не менее данная сеть отображает ситуацию так, что на свете существуют только те iPhone, которые принадлежат Иде и Тони, а это явно не соответствует действительности. Далее, каждый iPhone имеет батарею, но у него есть и другие части. Человеческий ум не пришел бы к заключению, что любая часть iPhone представляет собой только батарею, однако из диаграммы это никак не следует. Еще более пристальный анализ показывает, что в нашей семантической сети никак не отображается характерная для человеческого бытия тонкость — место рождения всегда уникально, но на собственность это правило не распространяется. Иначе говоря, если Ида

родилась в Бойсе, она никак не могла бы родиться в Бостоне; с другой стороны, это ограничение никак не мешает ей владеть не только iPhone, но, скажем, и телевизором.

Научить машины понимать, что же еще у людей на уме помимо формально отображенного на диаграмме, действительно сложно. Легко заметить, что в семантической сети недостает как раз того, для передачи чего она и создается: здравого смысла. Если вы не знаете заранее о таких вещах, как рождение (которое в случае человека происходит только в одном месте), а также об отличиях его от промышленного производства и владения (где одна компания может производить более одного продукта, а один человек может владеть несколькими вещами), то формальное связывание понятий — даже в очень сложную сеть — машине не поможет.

Ситуация с семантическим сетевым подходом станет еще хуже, если вы введете в описание параметры, связанные со временем. Посмотрите на сеть, изображенную на рис. 7.4 и напоминающую схему из ConceptNet, которую мы обсуждали ранее.

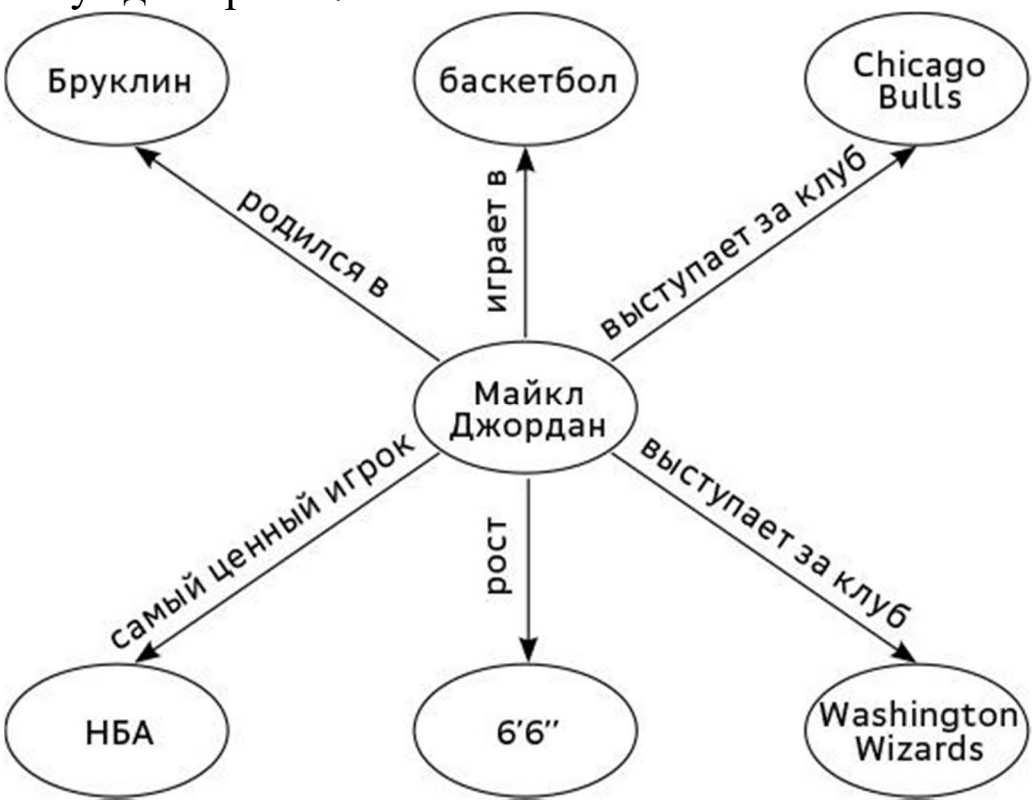


Рис. 7.4. Еще один пример семантической сети

При поверхностном изучении все выглядит довольно хорошо: баскетболист Майкл Джордан имеет рост шесть футов шесть дюймов, родился в Бруклине и т.д. Но более глубокий анализ показывает, что система, которая знает только информацию с этой диаграммы и ничего более, с легкостью может совершать нелепейшие ошибки. Возможно, она решит, что Майкл Джордан имел в момент рождения рост шесть футов или что он играл за клубы Wizards и Bulls одновременно. Слова «играет в баскетбол» могут относиться ко всем событиям его жизни, начиная с более узкого периода профессиональной карьеры и заканчивая всем промежутком времени с того момента, когда он впервые коснулся мяча в детстве, и непосредственно до сегодняшнего дня (при условии, конечно, что он все еще время от времени играет в баскетбол со

своими друзьями). Если мы теперь расскажем системе, что Джордан играет в баскетбол с 1970 года по настоящее время (исходя из идеи, что он начал это делать, когда ему было семь лет, и продолжает играть по сей день), ничто не мешает ей сделать ложный вывод, что Джордан все это время играл в баскетбол по 24 часа в сутки и 365 дней в году без остановки в течение 48 лет.

До тех пор пока машины не начнут успешно справляться с подобного рода задачами так, чтобы людям не приходилось им помогать буквально на каждом шагу, они не смогут читать, рассуждать или безопасно ориентироваться в реальном мире. Как нам их этому научить?

Первый намек приходит из области формальной логики, разработанной философами и математиками. Возьмем следующие утверждения: «Компания Apple производит все iPhone», «Ида владеет iPhone» и «Все iPhone имеют аккумулятор». Большая часть неопределенности, которую мы видели в семантических сетях, может быть решена путем кодирования этих фактов с использованием общеупотребительной технической нотации:

$$\forall \times \mathrm{IPhone}(x) \Rightarrow \mathrm{Made}(\mathrm{Apple}, x)$$

$$\exists z\ \mathrm{IPhone}(z) \wedge \mathrm{Owns}(\mathrm{Ida}, z)$$

Первое утверждение следует понимать так: «Для каждого объекта x, если x — это iPhone, однозначно верно, что [компания] Apple произвела x». Второе следует понимать так: «Существует такой объект z, удовлетворяющий условию z = iPhone, для которого верно утверждение, что z принадлежит Иде».

Понимание языка формальной логики требует некоторого привыкания, и неподготовленным людям это неудобно, что усложняет привлечение краудсорсинга для кодирования принципов здравого смысла. В наши дни описанный подход потерял всякую популярность в области искусственного интеллекта; практически каждый предпочел бы использование голых фактов. Однако на более длинной дистанции формальная логика или нечто подобное ей может стать первым необходимым шагом к представлению машинных знаний с необходимой точностью. В отличие от ссылки в семантической сети «Apple — производит — iPhone», формулировка, выраженная средствами математической логики, всегда однозначна. Не нужно догадываться, означает ли приведенная выше запись, что «Apple производит все iPhone», или «Apple производит некоторые iPhone», или «Единственное, что производит Apple, это iPhone». В том виде, как она написана, данная формула может соответствовать только первому утверждению.

Трансляция здравого смысла должна начинаться именно с чего-то подобного — либо с действительно формальной логики, либо с альтернативной системы записи фактов и отношений, выполняющей аналогичную функцию [39]. В любом случае это будет тот или иной способ четко и недвусмысленно представить все то, что знают обычные люди.

Но даже когда однажды мы придумаем правильный способ кодирования знаний, представляющих аналог здравого смысла у машин, у нас все равно останутся нерешенные проблемы. Одна из трудностей, с которой сталкиваются нынешние методы сбора понятий и фактов — ручное кодирование, веб-

майнинг и краудсорсинг, — заключается в том, что они часто оказываются неструктурированной мешаниной фактов, от «Муравьеды едят муравьев» до «Циклон Б ядовит». На самом же деле мы хотим другого, а именно чтобы наши машины имели полное и системное представление о мире.

Часть этой проблемы заключается в том, что не нужно, чтобы системы искусственного интеллекта изучали индивидуально каждый из бездны фактов, которыми владеют люди. Наоборот, важно, чтобы они поняли, как все эти бесчисленные факты связаны между собой. Нет никакого интереса в том, чтобы машины знали по отдельности, что писатели пишут книги, художники рисуют картины, композиторы сочиняют музыку и т.д. Вместо этого требуется, чтобы искусственные интеллектуальные системы рассматривали все эти конкретные факты как примеры более общего отношения между объектом и субъектом (например, в форме утверждения «индивид создает творение») и включали это наблюдение в более широкую структуру знаний, которая, в частности, дает понять, что создатель обычно владеет своим творением до тех пор, пока не продаст его, что творения одного человека часто стилистически сходны и т.д.

Дуг Ленат называет подобные коллекции атомарных знаний микротеориями (другим термином для них может быть «информационный каркас»). Количество микротеорий, вероятно, исчисляется тысячами. Специалисты в области представления знаний пытались разработать такие теории или каркасы для самых различных аспектов понимания реального мира, от психологии и биологии до использования бытовых предметов. Хотя в современных подходах к разработке искусственного интеллекта, ориентированных на большие данные, информационный каркас играет крайне незначительную роль (или вообще игнорируется), мы считаем конструирование микротеорий жизненно важной задачей, хотя более полное понимание того, как их создавать и использовать, может отнять немало времени.

Если бы мы могли опираться только на три микротеории, то скорее всего опирались бы в основном на темы, выступающие как центральные постулаты в «Критике чистого разума» Канта, которая утверждала (опираясь на философскую аргументацию), что фундаментальными аспектами бытия выступают время, пространство и причинность [40]. Для достижения реального прогресса в теории универсального искусственного интеллекта жизненно важно обеспечить этим постулатам твердую позицию в машинном разуме. Даже если мы еще не можем сотворить эти принципы сами, то по крайней мере мы можем что-то сказать о том, как они должны выглядеть.

Давайте начнем с понятия времени. Перефразируя Екклесиаста, можно сказать, что «для каждого события есть свое время» и без понимания временных связей между событиями почти ничего из области здравого смысла не будет нам доступно. Если дворецкий-робот должен подать бокал вина, он должен знать, что пробку нужно вытащить из бутылки перед тем, как налить вино, а не после. Роботу-спасателю во многом придется определять приоритеты своих действий, исходя из понятия времени и понимания того,

насколько срочными являются те или иные процедуры: огонь может распространиться на бóльшую площадь за считаные секунды (а значит, его надо тушить немедленно), в то время как спасать кошку, застрявшую на дереве, можно и без особой спешки.

В этом плане теоретики классического искусственного интеллекта вместе с философами добились весьма ощутимого прогресса, разработав формальные логические системы для представления различных ситуаций и того, как они развиваются или изменяются с течением времени. Наш робот-дворецкий может начать с осознания того, что вино в настоящее время находится в закупоренной бутылке, а бокал в это же самое время пуст и что его (робота) задачей будет в течение двух минут налить в бокал вино. Так называемая временна́я логика позволит роботу построить когнитивную модель этих событий, а затем перейти от когнитивной модели к общепринятым знаниям (например, если вы наливаете что-то из бутылки в бокал, то часть содержимого бутылки теперь окажется в бокале), а затем — и к определенному плану, структурированному во времени: откупорить бутылку в подходящий момент, потом (а не до того) наклонить горлышко бутылки к отверстию бокала и т.д.

Тем не менее и здесь нам предстоит серьезная предварительная работа. Первая сложность заключается в переносе синтаксических отношений между словами в предложениях на независимую временную шкалу. Возьмем такую фразу: «Тони удалось налить вино, хотя ему пришлось использовать нож, чтобы вытащить пробку, потому что он не смог найти штопор». Здесь недостаточно только логики, основанной на понятии физического времени. Чтобы прийти к выводу, что упомянутые события произошли не в том порядке, как они описаны в данном предложении, а в точности наоборот, потребуется уже кое-что знать о языке, в частности обо всех тех хитрых способах, которыми предложения могут описывать запутанные отношения между событиями во времени. В этом аспекте, к сожалению, еще никто не достиг существенного прогресса. (Остается непонятным также и то, как — и можно ли вообще — интегрировать все это с глубоким обучением.)

Чтобы создать систему, которая могла бы разобраться в том, когда Майкл Джордан играл в баскетбол, а когда — вряд ли, или машинный мозг, способный реконструировать, что должно было случиться перед тем, как Альманзо вернул мистеру Томпсону его кошелек, нам нужно больше, чем просто абстрактное понимание времени. Недостаточно знать, что события имеют начало и конец; вы должны понять конкретные факты о мире. Когда вы читаете: «Альманзо повернулся к мистеру Томпсону и спросил: "Вы не теряли бумажник?"», вы мысленно заполняете ряд фактов, относящихся к определенной хронологии: в самый первый момент у мистера Томпсона был в кармане кошелек; затем этого кошелька в кармане у него не стало; еще позднее его кошелек нашел Альманзо. Когда вы думаете о Майкле Джордане, вы используете тот факт, что даже самый заядлый спортсмен занимается спортом лишь часть своего времени (даже если не рассматривать те часы, когда он спит). Чтобы рассуждать о том, как мир разворачивается и меняется со временем, искусственному интеллекту потребуется интегрировать в свое

представление о мире сложную смесь общих истин (таких как «человек не может эффективно выполнять сложные действия во время сна») и конкретных фактов, позволяющих уточнить, как эти общие истины применимы в более частных обстоятельствах.

Аналогично, ни одна машина пока что не может посмотреть обычный для нас фильм и надежно уяснить, что в нем является воспоминаниями героя (то есть ситуациями из условного прошлого), а что — событиями, разворачивающимися в условном настоящем времени. Даже в фильмах, где события в основном сменяют друг друга в простой временной последовательности, все еще возникает много проблем, например оценка временных промежутков между данной сценой и следующей (прошла минута? день? месяц?). Решение таких задач почти всегда отдается на откуп зрителям (в конце концов, так интереснее смотреть). Возможность вычислять длину промежутков времени по ходу фильма опирается на базовое человеческое понимание того, как оно устроено, и вместе с тем — на обширные и подробные знания о том, что будет правдоподобным, а что нет при различных вариантах развития событий. Как только мы соберем все это вместе, разгадки придут сами.

Даже ваш виртуальный календарь станет намного умнее, когда научится размышлять о том, где вам следует быть и когда, а не просто сохранять каждое событие в виде множества предстоящих встреч, которые иногда идут с такой частотой, что вам не хватает времени, чтобы добраться с одной встречи на другую. Если даже вы запланируете мероприятие в другом городе или стране, программа сама правильно установит часовой пояс и не заставит вас прибыть на три часа раньше (а то и позже) назначенного времени. Принципиально здесь, что программистам не нужно заранее предвидеть каждый конкретный сценарий, потому что искусственный интеллект будет решать, что вам нужно, исходя из общих принципов, заложенных в нем изначально. Цифровой помощник сможет безошибочно предоставить вам список судей, которые в настоящее время состоят в Верховном суде (или любом другом суде, который вас заинтересует), назовет вам игрока «Чикаго Буллз», который дольше всех пробыл в команде, вычислит, сколько лет было Нилу Армстронгу, когда Джон Гленн облетел Землю на космическом корабле, и даже скажет вам, когда нужно ложиться спать, если у вас поезд завтра в 6:30 утра, а вы хотите поспать восемь часов. Программы персонализированной медицины смогут связать показатели анамнеза пациента за несколько минут или часов с тем, что он делал в течение предшествовавшего периода жизни. Высококлассное персональное планирование, в котором руководители фирм опираются на своих умных и опытных помощников, станет автоматизированной услугой, доступной каждому.

Помимо чувства времени, машины нуждаются в понимании пространственных и геометрических отношений в мире людей и объектов. Как и в предыдущем случае, некоторые аспекты базовой структуры этих знаний уже известны в подробностях; но есть и много таких, где прогресс явно оставляет желать лучшего. Например, евклидово пространство действительно

изучено хорошо, и мы знаем, как выполнять в нем все виды геометрических вычислений. Современные специалисты по компьютерной графике используют геометрию, чтобы вычислить, какие узоры создает на предметах падающий свет в условиях сложного освещения; результаты моделирования здесь настолько реалистичны, что кинематографисты постоянно используют подобные методы для создания убедительной картины событий, которые в реальности никогда и нигде не происходили.

Однако понимание того, как устроен мир, — это нечто гораздо большее, чем просто умение создавать реалистичные изображения. Подумайте, например, что вам требуется знать о форме двух обычных предметов, изображенных на рис. 7.5, — ручной терки и сетчатого пакета с овощами, — а также о том, что следует из нашего понимания этих форм.

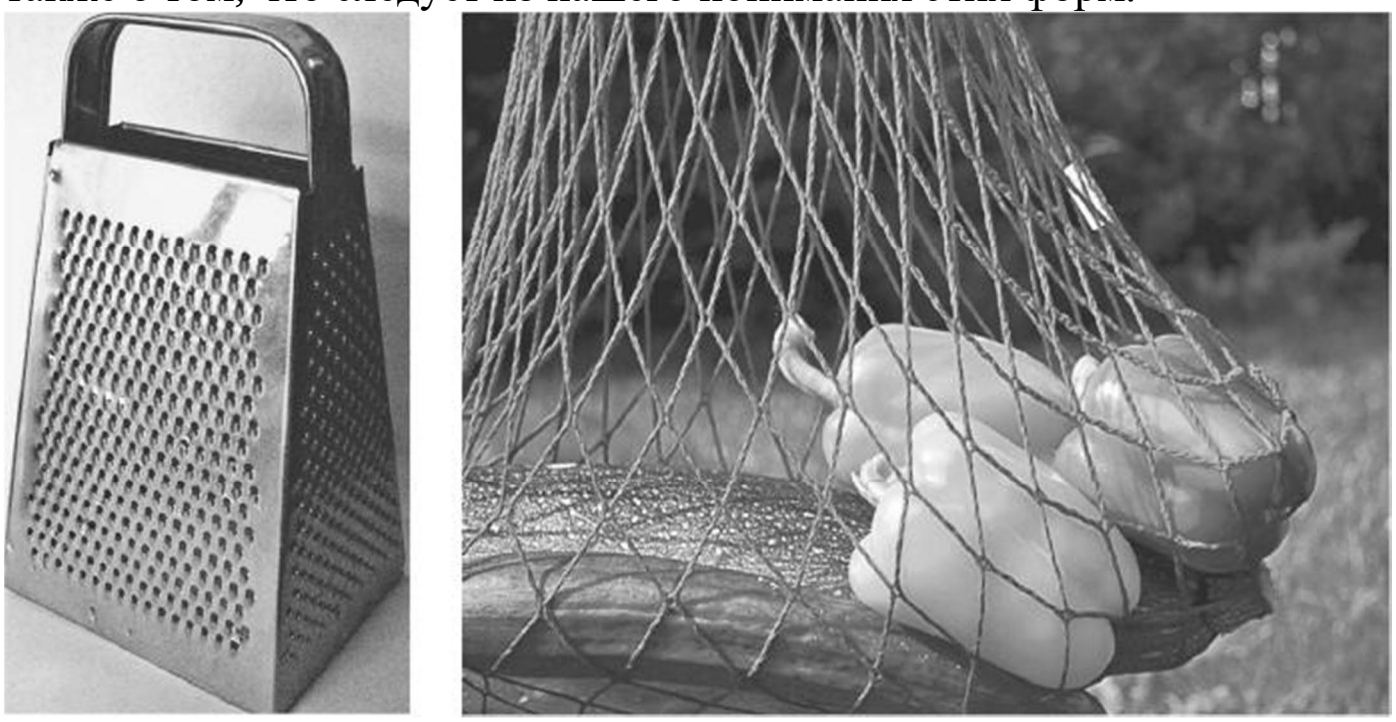


Рис. 7.5. Самые обычные предметы, создающие реальные проблемы для искусственного интеллекта

Оба этих хорошо знакомых нам объекта имеют довольно сложную форму (гораздо более сложную, чем элементарные геометрические фигуры в стереометрии, скажем сфера или куб), причем их пространственная форма очень важна для того, чтобы они могли выполнять свои функции. Терка представляет собой усеченную пирамиду — это делает ее более устойчивой на поверхности стола, — а ручка находится сверху, чтобы вы могли держать терку неподвижно, пока измельчаете овощи или сыр. Форма и структура отверстий, находящихся на каждой грани и ведущих во внутреннее пространство терки, позволяют, например, нарезать сыр узкими полосками, которые затем сами проваливаются внутрь. Наконец, отверстия на разных сторонах сгруппированы особым образом, чтобы обеспечить разнообразие вариантов измельчения; например, предназначенные для сыра чеддер отверстия на стороне, обращенной на изображении к зрителю, имеют маленький полукруглый режущий край с острым лезвием, для того чтобы каждый раз от куска сыра захватывалась и отсекалась лишь небольшая полоска. Все эти детали, если вдуматься, создают довольно функциональный

дизайн, причем пространственная форма всего приспособления определяется его задачами.

Стандартные программы для графики или компьютерного дизайна могут представлять форму, использовать ее в видеоигре, вычислять ее объем и даже определять, какие отверстия находятся в контакте с каким-то конкретным куском сыра, удерживаемым в определенном месте, но они не могут создать никаких причинно-следственных связей между формой терки и ее функциями. У нас еще нет системы, которая могла бы смотреть на терку, как мы, то есть понимать, для чего она нужна или как ее можно использовать для того, чтобы натереть сыр — например, моцареллу для приготовления пиццы.

В определенном смысле сетчатая сумка типа авоськи создает еще больше проблем для искусственного интеллекта, по крайней мере в его текущем состоянии. Терка хотя бы имеет четкую форму; вы можете перемещать терку в пространстве или поворачивать ее вокруг трех основных координатных осей, но вы не можете согнуть или сложить ее, поэтому элементы терки остаются в постоянных геометрических отношениях друг с другом. В отличие от нее, авоська не имеет постоянной формы: ее стенки огибают предметы, находящиеся внутри нее, и меняют свои очертания в зависимости от поверхности, к которой они прилегают. Таким образом, авоська — это не одна конкретная форма, а бесконечное множество форм; все, что остается неизменным в ее структуре, — это длина элементов, из которых она состоит, и способ их соединения друг с другом. Между тем из всей этой информации искусственный интеллект должен сделать вывод, что вы можете положить в авоську огурцы и перец, причем они останутся внутри, и что вы можете при желании положить в нее даже горох, однако он сразу посыплется наружу, и, наконец, что даже при большом желании вы не сможете положить в авоську огромный арбуз. Но к таким выводам современные компьютерные системы прийти не способны. Как только это все-таки произойдет, роботы смогут безопасно и эффективно работать в переменчивой, сложной и открытой среде — от кухонь и продуктовых магазинов до городских улиц и строительных площадок, что чрезвычайно расширит их возможности.

Причинность в широком понимании включает в себя любые знания о том, как мир меняется со временем [51]. Она может варьироваться от самого общего — закона тяготения Ньютона и теории эволюции Дарвина — до очень специфического: нажатие кнопки питания на пульте телевизора включает и выключает его; если гражданин США не подает свою годовую налоговую декларацию до 15 апреля следующего года, то рискует получить штраф. Происходящие изменения могут затрагивать физические объекты, сознание людей, социальную организацию и практически все, что меняется со временем.

Принципы причинности мы регулярно используем для понимания людей и других существ — психологи называют это интуитивной психологией. Используем мы ее и для овладения знаниями о неживых предметах, например когда разбираемся в принципах работы бытовых инструментов (молотки, дрели и т.п.), а также в более общем смысле, когда пытаемся понять, что

представляют собой те или иные артефакты — объекты, созданные человеком, — наподобие тостеров, автомобилей и телевизоров. Когда мы учимся понимать компьютеры, мы часто относимся к ним как к субъектам, имеющим собственную психологию (машина «хочет», чтобы я ввел свой пароль; если я введу свой пароль, машина «распознает» его и «позволит» мне ввести следующий запрос). Причинно-следственные связи необходимы и для понимания принципов работы социальных институтов (если вы хотите одолжить книгу, вы идете в библиотеку; если вы хотите принять закон, вам нужно провести его через конгресс), товарно-денежных отношений (если вы хотите съесть бигмак, вам придется заплатить за него), договорных обязанностей (если подрядчик отказывается от проекта на полпути, вы можете подать в суд за нарушение контракта) и языка (если два человека не говорят на одном языке, они могут использовать переводчика). Таким же образом мы можем рассуждать и о том, что на самом деле не происходит (если работники метро бастуют, значит, метро не работает и нужно найти другой максимально приемлемый способ добраться до работы). Очень большая часть рассуждений в рамках здравого смысла опирается на ту или иную форму причинности, почти всегда меняющуюся во времени и часто затрагивающую пространство.

Одним из признаков нашей способности мыслить в категориях причинно-следственных связей является исключительная гибкость человеческого ума. Мы можем, например, использовать любой конкретный факт о причинно-следственных связях целым рядом различных способов. Если мы понимаем связь между пультом дистанционного управления и работой телевизора, мы сразу сможем делать прогнозы, планировать действия и находить объяснения. Мы можем предсказать, что если мы нажмем кнопку питания на пульте, то телевизор включится. Далее мы быстро выясним, что если мы хотим включить телевизор, то для этого требуется нажать на нужную кнопку. Если мы заметили, что телевизор внезапно включился, то сделаем вывод, что кто-то в комнате, очевидно, нажал кнопку. Легкость, с которой люди могут осуществлять все подобные ментальные операции без какого-либо специального обучения, просто поражает. Если мы научим искусственный интеллект делать то же самое, это будет крупнейшей революцией во всей нашей области. После этого интеллектуальные аварийно-спасательные машины окажутся способными, например, чинить мосты или фиксировать сломанные конечности пострадавших с помощью любых доступных материалов, так как они будут сами разбираться в технике и материаловедении.

Особое значение для будущего искусственного интеллекта будет иметь способность быстро и плавно объединять различные области причинного понимания, что у людей получается совершенно естественно. Возьмем для примера сцену из телесериала «Закон Лос-Анджелеса», которую психолингвист Стивен Пинкер описал в своей книге «Как работает мозг». Одна из героинь сериала, безжалостный адвокат Розалинд Шейс, входит в двери лифта — за которыми самого лифта не оказалось — и падает в пустую шахту;

скоро до нас донесется ее последний крик. Глядя на эти события, мы, зрители, задействуем элементарные физические знания, чтобы почти мгновенно сообразить, что Розалинд, судя по всему, упадет на самое дно шахты, и одновременно используем простейшие знания из биологии, предугадывая, что такое падение наверняка убьет героиню. Вместе с тем мы используем и наш психологический опыт, на основании которого Розалинд Шейс предстает человеком, не способным на самоубийство, и заключаем из этого, что она погибла по ошибке. Да и сама ошибка Розалинд основана на другом предположении из области здравого смысла — если двери шахты автоматически открываются, то за ними обязательно будет присутствовать кабина лифта. В общем случае это соображение совершенно справедливо, но все-таки иногда лифт дает сбой, и тогда — как в описанном случае — это легко может привести к трагедии.

Теперь перенесем те же идеи в сферу деятельности роботов, предназначенных для помощи пожилым людям. Эти роботы станут намного лучше, если смогут естественно понимать непредсказуемые взаимодействия между различными областями знания. Например, робот-помощник должен предвидеть психологию пожилого дедушки, чтобы понять, как тот отреагирует на появление робота. Вдруг он попытается вырваться и убежать? Или даже проявит агрессию? Для этого необходимо интерпретировать пациента как сложный динамический физический объект. Если цель робота состоит в том, чтобы уложить дедушку в кровать, недостаточно поместить на кровать центр масс человека, надо следить за тем, чтобы его голова не качнулась и не ударилась об изголовье. Роботы, которые смогут одновременно рассуждать и о психологии, и о физике, станут гигантским шагом вперед по сравнению с теми, которые доступны сейчас.

Связность рассуждений должна сыграть главную роль и при разработке интеллектуальных машин, способных к глубокому пониманию. Например, для настоящего погружения в историю Альманзо система искусственного интеллекта, читающая эту повесть, должна понимать, что мистер Томпсон изначально не знал, что потерял свой кошелек, и что понял он это только после того, как услышал вопрос Альманзо и ощупал карман своих брюк. Иначе говоря, такой системе необходимо будет определить исходное психическое состояние мистера Томпсона и то, как это состояние менялось со временем по мере развития ситуации. Точно так же компьютеру потребуется уяснить, что мистер Томпсон был бы расстроен потерей кошелька и что он испытает облегчение, когда получит его назад от Альманзо и увидит, что все деньги на месте. Сюда же относится и понимание того, что Альманзо наверняка будет оскорблен подозрением в том, что это он мог быть вором, — оно опирается на знание психологии людей, в частности того, как они относятся к деньгам и социальному статусу. При наличии у искусственного интеллекта действительно сложной системы причинно-следственных связей понимание вещей, естественных для каждого человека, станет для машины нормой.

Интеграция временных, пространственных и причинно-следственных связей приобретает решающее значение в тех ситуациях, когда речь идет о

планировании действий в условиях открытого, динамически меняющегося мира и, наоборот, о понимании заранее написанного плана (или алгоритма), который при всей своей понятности для человека может оказаться для искусственного интеллекта слишком расплывчатым или недостаточно подробно изложенным (как мы уже видели, в кулинарных рецептах часто пропускаются очевидные для людей шаги). Нынешние роботы — буквалисты: если что-то не указано в алгоритме, они этого никогда не сделают (по крайней мере, на основе доступных в настоящее время технологий). Чтобы добиться максимальной отдачи от искусственного интеллекта, необходимо придать ему такую же гибкость, какой обладает наш собственный интеллект.

Возьмем в качестве примера алгоритм приготовления яичницы-болтуньи. В интернете предлагается множество рецептов, но ни один из них не описывает процесс в технической полноте. Вот рекомендации, взятые со случайно выбранного сайта.

1. Взбейте яйца, молоко, соль и перец в миске среднего объема до полного смешивания.
2. Нагревайте сливочное масло в широкой и неглубокой сковороде на среднем огне, пока масло не растопится.
3. Вылейте полученную смесь в сковороду.
4. Контролируйте процесс приготовления, слегка перемешивая смесь так, чтобы более жидкая часть контактировала с дном сковороды, а не оставалась сверху, пока все не загустеет равномерно и жидкие участки не исчезнут. Не следует, однако, перемешивать смесь непрерывно или слишком интенсивно.

Легко видеть, что авторы рецепта пропустили множество очевидных шагов, исходя из того, что читателю нет нужды пояснять элементарные подробности. На самом же деле в приготовлении болтуньи задействованы бесчисленные дополнительные операции. Нужно вытащить коробку с яйцами из холодильника и взять нужное количество, молоко и масло тоже требуется извлечь и распаковать. Перед смешиванием яйца необходимо разбить. Нужное количество масла требуется отрезать от брикета заранее, до того как вы его положите на сковороду, а неиспользованные продукты следует поместить обратно в холодильник. Если миска или сковорода окажутся грязными, их необходимо вымыть перед использованием. Наконец, легко догадаться, что если у вас не оказалось перца, то яичницу все равно можно будет приготовить, пусть даже немного пожертвовав ее вкусом, но бесполезно перемешивать молоко с любым количеством соли и перца при отсутствии яиц — вы получите все что угодно, только не яичницу.

В более широком смысле роботы должны сравняться с людьми по общему уровню когнитивной адаптируемости. Мы, люди, строим те или иные планы, а затем корректируем их буквально на лету, если выясняется, что обстоятельства не совсем соответствуют нашим ожиданиям. Мы можем сделать множество здравых и обоснованных предположений о развитии событий, в которых ни разу не участвовали. Например, мы легко догадываемся о том, что получится, если мы сдадим наш чемодан в багаж, забыв застегнуть на нем молнию, или попытаемся пронести доверху полную чашку кофе через длинный коридор в

движущемся поезде. По-настоящему серьезным достижением робототехники можно называть разработку лишь таких машин или цифровых помощников, которые будут способны составлять планы, а затем адаптировать их к меняющимся обстоятельствам.

Одним из очевидных, но в конечном итоге весьма разочаровывающих способов решения описанной проблемы является компьютерное моделирование. Если мы хотим узнать, может ли собака нести слона, мы могли бы запустить симуляцию этого процесса, основываясь на физических параметрах названных животных, подобно тому как это делается в современных видеоиграх.

В определенных обстоятельствах компьютерная симуляция может стать эффективным способом приблизиться к моделированию причинности. Виртуальный физический эксперимент может создавать близкое к реальности видеоотображение моделируемого процесса и детально определять, что и как в сценарии будет двигаться и меняться со временем. Например, симуляция, используемая в видеоиграх типа Grand Theft Auto, имитирует взаимодействие между людьми, автомобилями и различными другими объектами в игровом мире. Моделирование начинается с полного перечня базовых физических параметров, необходимых для развития игровой ситуации: формы каждого объекта, его веса, материала, из которого он изготовлен, и т.п. Затем программа использует точные физические формулы, чтобы предсказать, как каждый объект будет двигаться и меняться, с точностью около одной миллисекунды, обновляя картинку в зависимости от решений, принимаемых игроком. По сути, это уже некоторая форма причинного мышления: исходя из свойств множества игровых объектов в момент времени t, компьютер предсказывает, как должен выглядеть мир в момент времени t + 1. Ученые и инженеры часто используют моделирование для прогнозирования различных ситуаций, включая весьма сложные, например эволюцию галактик, движение клеток крови и аэродинамику вертолета, находящегося в воздухе.

В некоторых случаях симуляция неплохо проявляет себя и в приложении к программам искусственного интеллекта. Представьте себе, что вы проектируете робота, который забирает предметы с ленты конвейера и затем пакует их в коробки. Робот должен предвидеть развитие событий в тех или иных ситуациях, например если предмет, который робот снимает с конвейера, находится в определенном положении, то он (предмет) может перевернуться кверху дном. Для подобных проблем симуляция часто является лучшим подходом. Однако по целому ряду причин в большинстве рассуждений, которые приходится выполнять искусственному интеллекту, симуляции оказываются бесполезными.

Главная проблема состоит в том, что мы просто не можем смоделировать все на свете, начиная с поведения отдельных атомов, потому что у нас никогда не хватит на это компьютерного времени и требуемого объема памяти. В существующих симуляциях физические модели используют для аппроксимации сложных объектов упрощенные представления таких

объектов [52] (не углубляясь в слишком детальные и часто не до конца изученные подробности). Однако и их создание оказывается весьма трудоемкой задачей, и для большинства физических взаимодействий реального мира их до сих пор не существует. Следовательно, ни одна физическая симуляция не может считаться достаточно подготовленной для использования в приложениях искусственного интеллекта, и в ближайшем будущем нам нельзя даже надеяться, что эта ситуация как-то сдвинется с места. Нам уже сейчас необходимо дополнить физическое моделирование какими-то другими методами.

Повседневная жизнь полна всевозможных объектов, которые никто даже не удосужился встроить в компьютерные модели. Вспомните, сколько разных инструментов и машин мы используем для процедур, относящихся к понятию «резка». В среднестатистическом доме с ванной, кухней и кладовой для хранения инструментов имеется как минимум дюжина различных предметов, единственной функцией которых является резка или измельчение чего-либо: кусачки для ногтей, бритвы, ножи, терки, блендеры, шлифовальные машины, ножницы, пилы, зубила, газонокосилки и т.д.

Разумеется, существуют веб-сайты, называемые хранилищами активов, которые продают трехмерные модели, готовые для загрузки и подключения к стандартным механическим устройствам, но симуляторы, которые действительно разработаны настолько подробно, что уже пригодны для практического использования, существуют лишь для очень небольшой части инструментов и механизмов, с которыми современные роботы могли бы столкнуться в повседневной жизни. Возьмем любую конкретную разновидность блендера, скажем Cuisinart. Едва ли вы найдете для него по-настоящему хорошо детализированную трехмерную модель, но даже если таковая и существует, очень маловероятно, что она способна с точностью продемонстрировать, что именно будет происходить, когда воображаемый домашний робот задействует этот блендер для смешивания йогурта, банана и молока. И мы гарантируем вам — она никогда не догадается, каков будет эффект, если тот же робот попытается использовать блендер для дробления кирпича. Хранилища активов могут предложить вам модели как слонов, так и собак, но существующие физические модели, скорее всего, не будут в состоянии правильно интерпретировать ход событий после того, как вы положите слона собаке на спину.

Вернемся теперь снова к рецепту яичницы-болтуньи. Вполне понятно, что обычный человек не представляет всех сложных химических и физических процессов, происходящих при приготовлении этого блюда. Не обязаны это понимать и кухонные роботы. Множество людей в этом мире способны приготовить отличную порцию яичницы-болтуньи, но, вероятно, лишь единицы будут в состоянии объяснить физику и химию жарки яиц в масле на сковородке. Тем не менее каким-то образом мы неплохо справляемся с готовкой блюд и можем вполне успешно делать это в любых масштабах — от индивидуального обеда до общепита, — пусть и не имея полноценного представления о физико-химических основах кулинарии [53].

В случае роботов врожденные недостатки физического моделирования проявляются особенно ярко. Роботы — чрезвычайно сложные механизмы со множеством движущихся частей, которые взаимодействуют друг с другом и с внешним миром самыми различными способами. Чем больше взаимодействий, тем сложнее их правильно рассчитать. Например, безголовый собакообразный робот SpotMini, разработанный компанией Boston Dynamics, имеет семнадцать различных механических соединений (суставы и т.п.), каждое из которых работает в нескольких режимах вращения и приложения силы.

Более того, действия роботов, как правило, зависят от воспринимаемой ими информации, в частности потому, что большинство запрограммированных движений управляется обратной связью. Какое конкретно усилие будет приложено к тому или иному суставу, зависит от информации, которую определенный датчик передает конечности робота. Таким образом, использование симуляции для прогнозирования действий робота и их последствий повлечет за собой необходимость симуляции восприятия робота и того, как оно будет меняться со временем. Предположим, мы используем симулятор, чтобы проверить, можно ли рассчитывать на нового спасательного робота, если при пожаре нам потребуется доставить раненых в безопасное место. Симулятор должен знать не только о том, может ли робот сделать это физически, но и о том, сможет ли он найти дорогу в заполненном дымом здании при полностью отключенном электричестве. Пока что все это выходит далеко за пределы возможностей автономных мыслящих машин. В более широком смысле — то, что происходит с роботом, когда вы оставляете его одного в реальном мире, часто совершенно не соответствует тому, что происходит в симуляции. Это случается так часто, что робототехники придумали специальный термин — reality gap («разрыв с реальностью»).

Эффективность чисто имитационных подходов становится намного ниже в тех случаях, когда мы пытаемся создать физические аналоги, передающие рассуждения людей. Предположим, вы хотите выяснить, что произошло, когда мистер Томпсон коснулся своего кармана. В принципе, можно вообразить себе программу, которая симулирует поведение каждой клетки или молекулы в теле героя и полностью эмулирует обратную связь, полученную нервной системой мистера Томпсона через рецепторы в его пальцы, если кошелек будет присутствовать в кармане (или если его там не будет). Затем программа имитирует соответствующие паттерны нейронного возбуждения и в конечном итоге отправляет сообщение в префронтальную кору головного мозга воображаемого человека. Кульминацией этого процесса будет включение программы управления движением, которая заставит мистера Томпсона двигать губами и языком таким образом, чтобы воскликнуть: «Да, потерял! Да еще полторы тысячи долларов в придачу!»

Пофантазировать на эту тему приятно, но на практике модель Томпсона просто-напросто неосуществима. Объем вычислительной мощности, который потребуется для такого моделирования с требуемым уровнем детализации, слишком велик. По крайней мере, на данный момент, в начале XXI века, мы понятия не имеем, как воспроизвести в машинном варианте человеческий мозг

во всех подробностях. В рамках гипотетической симуляции необходимо будет рассчитать взаимодействия между совершенно невообразимым числом молекул, даже если речь идет о промежутке времени длиной всего в одну секунду. Нервная система в томпсоновской модели будет выполнять десятилетиями то, что у человека занимает мгновение. Короче говоря, нам подойдут лишь системы, которые способны абстрагироваться от точной физики в пользу психологии.

Последняя составляющая здравого смысла — это наша способность к рассуждениям. Вспомните знаменитую сцену из фильма «Крестный отец». Кинопродюсер Джек Вольтц просыпается и видит на противоположной стороне своей огромной кровати отрубленную голову его любимого жеребца. До него сразу же доходит смысл предупреждения, которое ему сделал накануне Том Хаген, и теперь он осознал: если команда Хагена с такой легкостью за одну ночь добралась до его тщательно охранявшейся лошади, то с такой же легкостью она сможет добраться до и самого Вольтца.

Когда мы впервые смотрим этот фильм и видим голову лошади на кровати Джека Вольтца, мы не бросаемся отыскивать в своей памяти именно такие примеры, чтобы понять, что почувствовал кинопродюсер. Мы, как и герой фильма, рассуждаем о том, что может произойти дальше, опираясь лишь на колоссальный фонд общих знаний о том, как устроен мир, обобщая, что мы знаем о людях, объектах, времени, физике, экономике, инструментах и о многом другом. Одно из преимуществ формальной логики заключается в том, что она позволяет понять большинство необходимой нам информации непосредственным образом. Легко сделать вывод, что если палец является частью руки, а рука является частью тела, то палец является частью тела; последний факт не требует специального анализа и целой команды исследователей, если вы знаете первые два факта и базовые логические принципы.

В качестве другого примера возьмем рассуждение о смерти Розалинд Шейс, которое можно достаточно легко транслировать в машинный интеллект с помощью алгоритма логических выводов при наличии у программы следующих фактов.

- В пустой шахте лифта никакие предметы (объекты) не имеют под собой опоры.
- Дно шахты лифта представляет собой твердую поверхность.
- Предметы, не имеющие под собой опоры, летят вниз, быстро набирая скорость.
- Предмет, падающий в пустой шахте лифта, быстро столкнется с ее дном.
- Человек — это один из объектов, способных падать.
- Человек, который быстро движется и сталкивается с твердой поверхностью, может погибнуть или сильно пострадать.
- Человек, в данном случае — Розалинд Шейс, вошел в пустую шахту лифта.

Затем алгоритм логических выводов может прийти к заключению о том, что Розалинд Шейс, вероятно, погибла или серьезно ранена, без необходимости построения полной, очень трудозатратной в вычислительном отношении

модели о поведении каждой молекулы в ее теле [54]. Когда срабатывают законы формальной логики, процесс рассуждений становится гораздо легче и быстрее.

Однако логика сталкивается с собственными проблемами. Во-первых, не каждый вывод, который может извлечь машина, полезен или уместен в реальной жизни. Зная правило «мать собаки — собака» и конкретный факт «Лесси — собака», система, основанная на чистой логике, может пуститься в погоню за верными, но бесполезными и не относящимися к делу выводами, например «мать Лесси была собакой», «мать матери Лесси была собакой» и «мать матери матери Лесси была собакой» — все это чистая правда, но вряд ли она потребуется хоть кому-то из нас. Точно так же машина, анализирующая историю Альманзо и пытающаяся понять, почему мистер Томпсон хлопает себя по карману, может долго блуждать по лабиринту из многочисленных выводов, говорящих, в частности, о том, что карман мистера Томпсона находится у него в штанах; что герой, вероятно, купил штаны в магазине одежды; что, когда мистер Томпсон купил эти штаны, у магазина одежды был владелец, что этот владелец магазина одежды, вероятно, начал тот день, когда мистер Томпсон купил у него штаны, с обычного завтрака и т.д. Как бы ни были верны эти выводы, ни один из них не будет играть существенной роли для ответа на вопрос «Зачем мистер Томпсон хлопал себя по карману?». Специалисты в области когнитивных наук часто называют описанную проблему «проблемой фреймов» и считают ее центральной проблемой автоматизированного мышления. Хотя полностью она все еще не решена, определенный прогресс в этой области все же достигнут.

Возможно, еще бóльшая проблема заключается в другом: основная цель работы формальных логических систем — сделать выводы максимально точными, однако в реальном мире многое из того, с чем нам приходится иметь дело, весьма расплывчато. Решение вопроса о том, было ли вторжение Советского Союза в Финляндию в 1939 году частью Второй мировой войны, с точки зрения формальной логики ничуть не легче, чем с точки зрения таксономии. В более широком смысле формальная логика, о которой мы говорили, хорошо проявляет себя только в одном аспекте: она позволяет нам получать знания, в которых мы полностью уверены, и применять работающие без исключений правила, чтобы вывести новые точные знания, в которых мы также будем уверены. Если мы полностью уверены, что у Иды есть iPhone, и мы уверены, что Apple производит абсолютно все айфоны, то мы можем быть столь же уверены, что у Иды есть что-то изготовленное компанией Apple. Но много ли вещей в жизни настолько однозначны? Бертран Рассел однажды написал: «Все человеческие знания неопределенны, неточны и неполны». И все же каким-то образом мы, люди, с жизнью справляемся…

Когда машины наконец смогут сделать то же самое, то есть сформулировать такого рода знания — неопределенные, неточные и неполные — и рассуждать о них с человеческой непосредственностью, эпоха универсального искусственного интеллекта, сильного и одновременно гибкого,

будет уже не за горами.

Логически правильное рассуждение, поиск релевантных способов представления знаний и сосредоточенность на ключевых аспектах разума (время, пространство, абстракции, индивидуальность) — все это должно помочь нам получить насыщенные знаниями и здравым смыслом когнитивные модели, демонстрирующие глубокое понимание. Новые способности, которые приобретет искусственный интеллект, позволят нам успешно изменить сегодняшнюю парадигму.

Чтобы достичь этого, нам понадобится еще кое-что: фундаментальное переосмысление того, как работает обучение. Необходимо разработать новый вид обучения, который эффективно использует уже имеющиеся знания, а не упрямо начинает изучение с чистого листа в любой области, с которой сталкивается. В текущей работе по машинному обучению исследователи и инженеры сосредотачиваются на какой-то одной очень конкретной, узкой задаче, пытаясь запустить интеллектуальную систему в прямом смысле с нуля. Они словно рисуют в своем воображении некую волшебную машину (не существующую в настоящее время ни в какой форме), которая при наличии достаточного времени выучит все, что ей нужно знать, просто просматривая видео на YouTube, без использования ранее полученных знаний. Но мы не видим доказательств того, что это когда-нибудь сработает, или даже того, что машинное обучение начинает делать успехи в этом направлении.

В лучшем случае это будет лишь опрометчивым обещанием: современные системы искусственного интеллекта понимают видеозаписи слишком поверхностно и неточно. Например, система видеонаблюдения может определять разницу между сценой, в которой человек идет шагом, и той, где он бежит, но ни одна система не может надежно распознать более тонкие отличия, например разницу между кадрами, где хозяин велосипеда отпирает велосипедный замок, и кражей велосипеда с использованием отмычки. Максимум того, на что способны современные системы глубокого обучения, — это маркировка видеозаписей, но даже это они, как правило, делают плохо, со множеством ошибок, подобных тем, что мы уже видели в этой книге. Ни одна существующая система ИИ не смогла бы, посмотрев фильм «Спартак», получить хотя бы малейшее представление о том, что в нем происходит, или ознакомиться с видеоверсией истории Альманзо и сделать вывод, что люди любят деньги, но не любят терять кошельки. Такой системе можно было бы предоставить всю информацию из «Википедии» или с других сайтов о кошельках и людях, но это ничуть не помогло бы ей улучшить понимание той детской истории. Назначение тегов видеозаписям — совсем не то же самое, что понимание событий, происходящих в них, и тем более не то же самое, что накопление знаний об устройстве окружающего мира.

Вынашиваемая определенным кругом людей идея о том, что некая самодостаточная (не нуждающаяся в помощи человека) видеосистема могла бы посмотреть фильм о Ромео и Джульетте и потом рассказать нам о человеческих отношениях, любви и иронии судьбы, выглядит нелепой, и в любом случае мы

так же близки к ее созданию, как к посещению туманности Андромеды. Пока что даже самые глубокие вопросы, которые мы можем задать компьютерам о видео, относятся к узкой технической сфере, например: «Какая сцена может появиться в видео, помеченном определенным тегом?» Если же спросить: «Что произошло бы, если бы Ромео так и не встретил Джульетту?» — современные системы просто не поняли бы вопроса, поскольку у них вообще отсутствуют какие-либо знания о человеческих отношениях. Это все равно что попросить камбалу бросить баскетбольный мяч.

С другой стороны, мы не хотели бы выплеснуть с водой и ребенка: обучение компьютерных систем необходимо, но подойти к нему следует более сложным образом, основываясь на уже существующих знаниях — это позволило бы нам прогрессировать быстрее. Как показал опыт Лената с CYC, будет, по-видимому, нереально закодировать вручную все, что может потребоваться машинам. Машины должны будут многому научиться самостоятельно. Мы могли бы вручную записать в машинных кодах тот факт, что острые твердые лезвия могут резать мягкий материал, но и тогда искусственному интеллекту придется изучать, как работают ножи, терки для сыра, газонокосилки и блендеры, опираясь лишь на эти знания и не имея закодированных вручную принципов работы каждого из этих механизмов.

Исторически сложилось так, что путь искусственного интеллекта все время пролегал где-то посредине между двумя крайними вариантами познания: ручным кодированием и машинным обучением. Изучение того, как работает газонокосилка, по аналогии с тем, как работает нож, очень отличается от совершенствования системы, которая классифицирует породы собак путем сбора все более подробно помеченных фотографий. Пока что основная масса исследований проводилась именно в области чистого машинного обучения, без учета возможности для ИИ учиться теми же способами, что и люди. Маркировка изображений ножей — это всего лишь изучение общих шаблонов пикселей, а не самих предметов. Осознание того, что и как делает нож, требует гораздо более глубоких представлений о формах, функциях и связях между ними. Понимание принципов использования ножа (и опасностей, сопряженных с такой работой) — это не накопление все большего множества картинок, а восприятие и изучение причинно-следственных связей. Цифровой помощник, который участвует в планировании свадебного торжества, должен не просто знать, что гостям в определенный момент требуется принести ножи и торт, он должен знать, почему это делается, то есть что ножи существуют для того, чтобы разрезать торты и другие блюда. Если бы однажды торт заменили специальными свадебными молочными коктейлями, то ножи, вероятно, вообще не понадобились бы, и здесь не имеет никакого значения, насколько сильно коррелировали ножи и свадьбы в предыдущих случаях. Вместо статистики у надежного цифрового помощника должно быть достаточно знаний о молочных коктейлях, чтобы понять, что ножи могут оставаться на своих местах, но зато может потребоваться дополнительная партия соломинок. Чтобы научить системы искусственного интеллекта хотя бы таким на первый

взгляд элементарным вещам, нам уже нужно вывести обучение на новый уровень.

Уроки, почерпнутые нами из анализа человеческого разума, заключаются в первую очередь в том, что мы должны почти всюду искать компромиссы. Нет никакого смысла начинать обучение мыслящих машин с чистого листа, чтобы им приходилось осваивать абсолютно все с нуля. С другой стороны, системы, у которых все заранее прописано программистами для любой мыслимой и немыслимой ситуации, выглядят совсем нереалистичными. Нам нужны тщательно структурированные гибридные системы с хорошими врожденными знаниями и навыками, которые позволяют машинам осваивать новые вещи на концептуальном и причинном уровнях; системы, которые могут изучать теории, а не просто отдельные факты. Хорошим вариантом для начала поисков в этом направлении представляются нам так называемые основные системы, о которых речь шла у Спелк, например системы отслеживания отдельных людей, мест и объектов. Такие системы являются стандартом классического варианта искусственного интеллекта, и машинное обучение еще не оказало на них серьезного влияния.

В обобщенном и более кратком виде наш рецепт обучения машин основам здравого смысла, чтобы в конечном счете привить им универсальный интеллект, заключается в следующем. Давайте начнем с разработки систем, разум которых способен воспринимать и использовать основные элементы человеческого знания: время, пространство, причинность, базовые знания о физических объектах и их взаимодействиях, базовые знания о людях и их взаимодействиях. Затем мы включим эти представления в более сложную архитектуру, которая может свободно оперировать всеми видами человеческих знаний, в то же время никогда не забывая о трех главных принципах познания: абстракции, композиционности и проверке теорий на конкретных объектах и людях. После этого потребуется разработка действенных методов рассуждения, которые смогут обращаться к знаниям, являющимся не пошагово изложенными алгоритмами, а сложноструктурированными наборами фактов, часто неопределенными и неполными; эти рассуждения должны будут в равной мере использовать и внешние данные, и уже накопленную внутреннюю информацию. Дальше нам потребуется соединить эти методы с восприятием, манипулированием и языком, на основе чего можно будет уже формировать насыщенные когнитивные модели мира. Наконец, останется построить самое главное — систему обучения, основанную на общечеловеческом подходе к этому процессу, где используются все знания и когнитивные способности, которыми к тому времени уже будет обладать искусственный интеллект. Сюда относится и то, что предстоит изучить, и уже имеющиеся знания, и задействование всех возможных источников информации, включая взаимодействие с миром, общение с людьми, чтение, просмотр видео и обучение с педагогом. Соберите все это вместе, и вы получите глубокое понимание. Конечно, это очень сложная задача, но ничто другое не приведет нас к нужной цели.

ГЛАВА 8

Доверие

Боги всегда ведут себя так же, как и люди, которые их создают.
Зора Ниэл Хертсон. Tell My Horse

Очень некрасиво бросаться людьми!
Слова Анны снежному гиганту Зефирке, из фильма киностудии Disney Film «Холодное сердце», 2013, режиссер Дженнифер Ли

В предыдущих главах мы не раз приводили примеры того, что машины, обладающие здравым смыслом и реальным пониманием мира, гораздо более надежны и дают намного более устраивающие нас результаты, чем те, которые опираются только на статистику. Однако существует и несколько других составляющих машинного разума, о которых необходимо позаботиться заблаговременно.

Надежные системы искусственного интеллекта должны создаваться с учетом лучших традиций инженерной практики, предусмотренных законами и стандартами во многих отраслях, но, увы, не в нашей сфере. Пока что основная масса приложений, управляемых искусственным интеллектом, представлена скороспелыми программистскими решениями, которые заставляют систему начинать работать немедленно, без необходимого уровня инженерных гарантий, обязательных практически для всех автоматизированных устройств. В автомобилестроении, например, необходимы так называемые стандартные стресс-тесты (в частности, краш-тесты и климатические испытания). В разработках, касающихся ИИ, что-либо подобное встречается весьма редко. Специалисты по искусственному интеллекту могли бы многому научиться у инженеров из других областей.

Например, в критических для безопасности узлах и конструкциях хорошие инженеры всегда проектируют машины и устройства так, чтобы они были значительно прочнее, чем тот минимум, который предполагают теоретические расчеты. Если, скажем, инженеры ожидают, что лифт никогда не будет перевозить более полутонны, им необходимо удостовериться, что в реальности этот лифт может выдержать пять тонн — в десять раз больше. Разработчик-программист, создающий веб-сайт, который рассчитан на 10 млн посетителей в день, должен убедиться, что сервер его сайта может обработать 50 млн запросов в сутки — на случай внезапного всплеска популярности. Неспособность обеспечить адекватный запас прочности грозит бедствием. Известно, что уплотнительные кольца в космическом челноке «Челленджер» хорошо работали в теплую погоду, но вышли из строя при запуске в холодную погоду и результаты этого просчета оказались катастрофическими. Если мы теоретически считаем достаточным, чтобы детектор опознавания пешеходов в беспилотных автомобилях имел надежность 99,9999%, мы должны добавить к

этому числу еще один десятичный знак и добиться фактической надежности, равной 99,99999%.

В области искусственного интеллекта, относящейся к системе машинного обучения, обеспечить требуемую безопасность разработчикам не удавалось ни раньше, ни сейчас. Конструкторы даже не могут разработать адекватные процедуры для проверки гарантий того, что в реальной ситуации системы ИИ будут работать с той или иной точностью и надежностью, и это выставляет их в очень невыгодном свете по сравнению, скажем, с производителями автомобилей или самолетов. Представьте себе производителя автомобильных двигателей, который заявляет, что новый двигатель будет работать в 95% случаев, не объясняя ничего про диапазон наружных температур, при которых его можно безопасно эксплуатировать. В случае искусственного интеллекта главным критерием применимости является принципиальная полезность системы, то есть работает ли она достаточно хорошо, чтобы из ее использования можно было извлечь выгоду. Однако такое отношение к надежности совершенно неприемлемо, если на карту ставится человеческая жизнь и безопасность. Автоматическое распознавание людей на фотографиях может иметь надежность 90%, и если речь идет только о личном использовании этой технологии для публикаций фотографий в Instagram, то этого будет вполне достаточно. Но если ее применяет полиция для поиска подозреваемых на записях камер наблюдения, то 10% ошибок — это катастрофически много. Поисковик Google, возможно, не нуждается в стресс-тестах, но программы автопилотирования машин и самолетов просто обязаны проходить всевозможные испытания на безопасность.

Кроме того, хорошие инженеры всегда предусматривают дублирующие системы на случай отказа основной. Они четко осознают, что ни один специалист не способен спрогнозировать в деталях все возможные сбои, поэтому включают в конструкцию системы резервного функционирования, которые заменят главную систему в случае непредвиденных обстоятельств. Даже у велосипедов есть и передние тормоза, и задние. Частично это делается, конечно, для удобства, но важно и то, что, если один тормоз выйдет из строя, второй все еще будет функционировать. На борту американских космических челноков было предусмотрено пять идентичных бортовых компьютеров, которые могли диагностировать друг друга и обеспечить резервный функционал в случае отказа одного (или даже четырех) из них. В обычном режиме четыре компьютера работали, а пятый находился в режиме ожидания, но в принципе для обеспечения полноценной работы шаттла было достаточно всего лишь одного. Точно так же автомобильные системы без водителя для надежности должны использовать не только видеокамеры, но и так называемые лидары (LIDAR, устройства, подобные радарам, но более точные и быстрые, потому что для измерения расстояния в них задействованы лазеры). Благодаря этому возникает частичная избыточность автоматического контроля, но пренебрегать ею не следует — в некоторых ситуациях и избыточности может едва хватить для предотвращения аварии. Но даже лучшие производители не всегда это осознают. Так, в течение многих лет Илон Маск

продолжает заявлять, что система автопилота Tesla не нуждается в лидарах, но с инженерной точки зрения это представляется очень рискованным, особенно с учетом далеко не идеальной надежности существующих систем машинного зрения. Большинство конкурентов Маска используют лидары в обязательном порядке.

Более того, хорошие инженеры всегда проектируют средства аварийного обеспечения — максимально надежные механизмы предотвращения катастрофы, если что-то выйдет из строя окончательно. Такие механизмы должны быть предусмотрены во всех критически важных узлах. Например, канатные дороги в Сан-Франциско имеют три тормозных каскада. Есть штатные колодочные тормоза, которые зажимают колеса кабины; если они не работают, то в дело вступают рельсовые тормоза — большие деревянные блоки, которые прижимают друг к другу направляющие, чтобы остановить кабину; если же не срабатывают и они, существует аварийный тормоз, массивный стальной стержень, который опускается и плотно прижимается к канатам. После того как в дело вступит аварийный тормоз, освободить кабину можно будет уже только с помощью сварочного аппарата, но это все же намного лучше, чем позволить кабине мчаться бесконтрольно.

Наконец, квалифицированные инженеры понимают, что к разным узлам и механизмам нужен разный подход. Радикальные инновации часто меняют подходы к тестированию новых машин, например путем внедрения автоматизированных самотестирующихся систем, однако устройства и детали, критически важные для безопасности, как правило, необходимо испытывать более старыми методами, проверенными на самой разнообразной продукции и показавшими себя надежными во множестве ситуаций. Система искусственного интеллекта, управляющая энергосистемой, не должна стать тем местом, где можно впервые опробовать новейший алгоритм, созданный в рамках дипломной работы, пусть даже ему прочат великое будущее.

Пренебрежение мерами безопасности может создавать очень серьезные и долгосрочные риски. Например, кибернетическая инфраструктура во многих критических аспектах оставалась совершенно неадекватной в течение нескольких десятилетий, что сделало ее крайне уязвимой и перед случайными сбоями, и перед злонамеренными действиями киберпреступников [55]. Широко распространившийся за последнее время «интернет вещей», от метеорологических датчиков до автомобилей, подключенных к обычному интернету, небезопасен до такой степени, что об этом уже рассказывают анекдоты. В одном широко известном реальном инциденте «этичные хакеры» [56] сумели овладеть системой управления джипом, в котором по шоссе ехал известный журналист. Еще одна высокоуязвимая мировая система — это GPS. Компьютерные устройства всех видов полагаются на спутники не только для обеспечения автоматизированного управления движением, но и для определения местоположения и времени буквально всего на свете — от смартфонов (и их владельцев) до квадрокоптеров и самолетов. Между тем спутниковые данные довольно легко заблокировать или подделать, и последствия этого могут быть катастрофическими. Еще один пример: хорошо

известно, что российское правительство совершало кибератаки на энергосистемы США, атомные электростанции, систему водоснабжения страны, объекты авиационной и другой стратегически важной промышленности. В ноябре 2018 года система водоснабжения Америки «удостоилась» определения «идеальная цель для киберпреступников». Если некий режиссер захочет снять научно-фантастический фильм в жанре апокалиптического реализма, то сценарий криминогенного коллапса глобальных автоматизированных систем будет несравненно более правдоподобным и ничуть не менее страшным, чем пресловутый «Скайнет». Нет причин сомневаться в том, что искусственный интеллект вскоре тоже станет мишенью для киберпреступников.

Проблемы надежности и безопасности не ограничиваются только стресс-тестами. После развертывания новой технологии ее необходимо постоянно поддерживать в хорошем состоянии, и квалифицированные инженеры заранее проектируют эти системы так, чтобы их можно было легко обслуживать. Конструкция автомобильных двигателей не должна создавать неразрешимых проблем для сотрудников автосервиса; операционная система компьютеров или иных интеллектуальных устройств действительно хороша тогда, когда ее удобно отлаживать, обновлять и переустанавливать.

Все сказанное актуально для искусственного интеллекта не менее, чем для любой другой области. Автономная система вождения, которая распознает другие транспортные средства, должна позволять устанавливать все необходимые обновления в случае выхода на рынок новых моделей автомобилей. Если автор некой системы уходит из фирмы, его преемникам должно быть очевидно, как именно следует ее совершенствовать и исправлять возможные недочеты. На данный момент, однако, в сфере искусственного интеллекта целиком преобладают большие данные и глубокое обучение, влекущие за собой появление все более сложных для интерпретации алгоритмов и моделей, которые сложно отлаживать и еще труднее поддерживать.

Если общие принципы надежного проектирования применимы к искусственному интеллекту в той же мере, в какой и к другим областям, то существует и ряд специализированных технических методов, которые можно и нужно извлечь из опыта разработки программного обеспечения. Например, опытные инженеры-программисты обычно используют так называемую модульную конструкцию. Когда разработчики программного обеспечения создают дизайн системы для решения комплексной проблемы, они делят соответствующие задачи на составные части и делают для каждой из задач отдельную подсистему. Они заранее планируют, что будет выполнять каждая подсистема, и, соответственно, любую из них можно написать в виде самостоятельной программы и протестировать по отдельности. Кроме того, они заблаговременно решают, как модули должны взаимодействовать друг с другом, чтобы все соединения можно было проверить по одному и в совокупности, убеждаясь каждый раз, что система работает. Например,

поисковая система интернета на верхнем уровне имеет сканер, который собирает документы из сети, индексатор, который классифицирует документы по ключевым словам, ретривер, использующий индекс для поиска ответа на запрос, пользовательский интерфейс, отвечающий за представление данных в удобном для пользователя виде, и т.д. Каждый из перечисленных блоков, в свою очередь, состоит из более частных подсистем.

Разработчики программ сквозного машинного обучения, ставших особенно популярными после некоторых успехов системы Google Translate, стараются намеренно пренебречь модульной структурой, добиваясь тем самым максимально быстрого ввода своих продуктов в коммерческое обращение. Но такая стратегия имеет серьезные минусы. Целый ряд важнейших проблем, например как отображать смысл переводимого предложения в интерфейсе, постоянно откладывается на будущее и в результате никогда не решается. Это, в свою очередь, приводит к тому, что в дальнейшем окажется сложно или даже невозможно интегрировать второпях созданные приложения в более сложные системы, потребность в которых наверняка возникнет. Один из ведущих исследователей искусственного интеллекта Леон Ботту, работающий в лаборатории AI Research компании Facebook, полагает, что проблема объединения традиционного программного обеспечения и систем машинного обучения «до сих пор остается практически нерешенной» [41].

Хорошее проектирование также требует хороших метрик — индикаторов, позволяющих оценивать технический прогресс в различных областях, чтобы инженеры знали, что их усилия действительно приводят к ощутимым улучшениям. На сегодняшний день самым известным индикатором прогресса общего интеллекта у машин является тест Тьюринга [42], который выясняет, сможет ли машина обмануть группу экспертов, притворившись человеком. К сожалению, несмотря на свою популярность, на практике он не особенно полезен. Хотя тест Тьюринга, как обычно полагают, представляет мыслящую машину реальному открытому миру и позволяет судить об искусственном интеллекте по критериям здравого смысла, реальность такова, что «судей» легко обмануть довольно дешевыми трюками. Еще со времен чат-бота Элизы, продемонстрированного в 1965 году, стало очевидно, что многих людей легко ввести в заблуждение, используя множество более или менее простых уловок, не имеющих ничего общего с интеллектом. Машина может избежать лишних вопросов, например, притворившись параноиком, подростком или иностранцем с недостаточным знанием местного языка. Победитель одного из последних конкурсов, организованных по принципу теста Тьюринга, — бот по имени Женя Густман — сумел объединить все три приема, притворяясь тринадцатилетним мальчишкой из Одессы. Однако цель искусственного интеллекта должна заключаться не в том, чтобы обмануть людей, а в том, чтобы достигнуть понимания открытого мира и научиться действовать в нем способами, сравнимыми по своей полезности, эффективности и надежности с человеческой деятельностью. Теста Тьюринга для этого просто недостаточно, так что нам следует придумать что-то получше.

Вот почему мы и многие наши коллеги, например сотрудники Института искусственного интеллекта Аллена (Allen Institute for Artificial Intelligence), уже несколько лет разрабатываем альтернативы тесту Тьюринга. Более корректное тестирование подразумевает широкий спектр задач, таких как понимание человеческого языка, способность делать выводы о физическом и умственном состоянии людей, анализ видео на YouTube, владение элементарными научными знаниями и способность к автономному выполнению роботизированных операций. Более продвинутым вариантом тестирования может стать экзамен, на котором системе предлагается изучить некоторые видеоигры и затем перенести полученные навыки на другие игры. Еще более впечатляющим мог бы стать робот-ученый, который будет способен прочитать описания простых экспериментов в книжке типа «Сто научных экспериментов для детей», выполнить их, понять, что они доказывают, и предсказать, что произойдет, если вы поставите их немного по-другому. Так или иначе, основная цель прогресса в области искусственного интеллекта должна состоять в том, чтобы приблизить существующие системы к машинам, умеющим гибко рассуждать, приспосабливая то, чему они уже научились, к новым ситуациям и не теряя при этом исходной надежности. Без более объективных показателей общей разумности добиться создания подлинного интеллекта будет крайне затруднительно.

Наконец, разработчикам искусственного интеллекта следует держаться как можно дальше от создания систем, которые могут слишком легко выйти из-под контроля. Например, любые работы по созданию роботов, которые могут проектировать и создавать других роботов, должны проводиться с особой тщательностью и только под пристальным наблюдением целого ряда экспертов, поскольку последствия неправильных решений в данной сфере очень трудно прогнозировать. На примере биологических организмов, внедрившихся в чужую для них среду и не встретивших там никаких препятствий к массовому размножению и расселению, мы уже хорошо понимаем, насколько опасно давать любой самовоспроизводящейся системе возможность бесконтрольно увеличивать численность. Появление с нашей помощью (или по нашему недосмотру) роботов, которые могут сами размножаться и улучшать себя неизвестными нам способами, таит в себе серьезную опасность.

Точно так же у нас нет достаточно надежных теорий, позволяющих с большой точностью прогнозировать последствия обретения роботами полного самосознания; по крайней мере, их не существует прямо сейчас [57]. Применение искусственного интеллекта, как и вообще любой технологии, сопряжено с риском непредвиденных ситуаций или даже катастроф, и, очевидно, чем шире мы открываем этот ящик Пандоры, тем в больших масштабах мы играем с огнем. В своем нынешнем состоянии искусственный интеллект несет мало рисков, однако незачем искушать судьбу, беспечно полагая, что все, что мы потенциально способны изобрести, будет работать и служить нам как задумано.

Мы (с должной осторожностью) полагаем, что для проверки надежности искусственного интеллекта можно было бы достаточно успешно использовать один известный среди разработчиков формальный метод, называемый верификацией программ и до сих пор больше распространенный в сфере классического искусственного интеллекта, чем в машинном обучении. По сути, это не один метод, а набор приемов, задействующий формальную логику для проверки правильности работы компьютерных систем или как минимум отсутствия каких-либо ошибок. Данный подход — как следует из его основных принципов — действительно позволяет во многих случаях убедиться, что тот или иной компонент искусственного интеллекта будет выполнять именно ту задачу, для которой он предназначается.

Каждое устройство, которое подключается к компьютеру, например динамики, микрофон или внешний диск, требует установки (или исходного наличия) так называемого драйвера, который представляет собой специальную программу, контактирующую с внешним устройством, опознающую его, запускающую его физическую работу и обеспечивающую взаимодействие между ним и компьютером. Многие драйверы чрезвычайно сложны с точки зрения программного кода и содержат иногда сотни тысяч строк. Поскольку драйверы периферических устройств должны корректно взаимодействовать с центральным ядром операционной системы (ОС) компьютера, ошибки в коде являются критическими для правильной работы практически всех частей, из которых и состоит любой персональный компьютер или ноутбук. Проблема совместимости драйверов с различными видами и версиями ОС постепенно обострялась по мере того, как написание драйверов все больше переходило к производителям периферического оборудования, а компании, разрабатывающие операционные системы, все чаще оставались в стороне от этого процесса.

В течение многих лет эта проблема регулярно приводила к многочисленным сбоям в работе компьютеров, вплоть до полного хаоса, пока в конце концов в 2000 году компания Microsoft не ввела набор строгих правил, которым должны следовать драйверы устройств при взаимодействии с операционными системами Windows. Чтобы обеспечить соблюдение этих правил, Microsoft предоставила и специальный программный инструмент под названием static driver verifier (статический верификатор драйверов), который проверяет правильность написания кода каждого конкретного драйвера, чтобы убедиться в его соответствии разработанным компанией правилам. После того как эта система проверки была введена в действие, сбои системы стали происходить значительно реже.

Аналогичные системы формально-логического анализа использовались для проверки ошибок в других сложных программах и в контрольных автоматах различных аппаратных устройств. Таким образом, в частности, было доказано, что компьютеризированная программа управления для авиалайнеров Airbus надежна в логическом и математическом смысле, то есть не содержит ошибок кода, которые могли бы привести к сбою программного обеспечения, известного своей чрезвычайной сложностью. В дополнение к этому команда

инженеров аэрокосмической промышленности и компьютерных ученых из университетов Карнеги — Меллон и Джонса Хопкинса недавно смогла объединить проверку авиационной электроники с логической проверкой физических параметров, чтобы убедиться в надежности программ предотвращения столкновений, используемых в самолетах.

Конечно, верификация программ тоже имеет ограничения. Ее алгоритмы позволяют оценить, как авионика (программы автоматического управления самолетом) будет вести себя в различных обстоятельствах; однако это не может дать гарантии того, что люди-пилоты будут управлять самолетами в достаточном соответствии с протоколом, или что датчики всегда будут работать должным образом, или что обслуживающий персонал никогда не пропустит необходимые проверки, а запчасти обязательно будут соответствовать требуемым спецификациям. (Две подряд фатальные катастрофы, случавшиеся с недавно выпущенной моделью самолета Boeing 737 Max и вызвавшие широкий общественный резонанс, были связаны в основном именно с неправильной информацией, поступавшей с датчиков, хотя еще одной проблемой в подобных случаях является возможность оперативного отключения автоматики и перевода самолета в режим ручного управления.)

И все-таки возможность убедиться в том, что само программное обеспечение не даст сбоя, уже является основой для безопасной автоматизации. Как минимум мы не хотели бы, чтобы программы интеллектуального управления самолетами стали сами собой перезагружаться прямо во время полета, и точно так же нам не понравилось бы, если бы домашний робот, собирающий нам книжный шкаф, внезапно сошел с ума и развалил всю конструкцию или напал на нашу дочь, приняв ее за нарушителя частных владений. Исследователям, занятым в области искусственного интеллекта, следует всячески поддерживать и развивать принципы, лежащие в основе программного тестирования, и, более того, стремиться к тому, чтобы инструменты глубокого понимания стимулировали появление новых подходов к тому, чтобы машины могли самостоятельно рассуждать о правильности, надежности и устойчивости программного обеспечения.

По мере развития технологий как минимум станет возможно доказать, что тестируемая система искусственного интеллекта гарантированно избегает ошибок определенного рода; например, что при нормальных обстоятельствах робот не упадет и не сломает себя вместе с мебелью или что результат машинного перевода является грамматически правильным. В более оптимистичном сценарии когнитивные возможности самого искусственного интеллекта смогут продвинуться значительно дальше, дойдя в итоге до уровня подражания опытным разработчикам программного обеспечения с их способностью предвидеть, как программное обеспечение работает в широком диапазоне условий, чтобы постоянно улучшать кодирование и отладку.

Каждая из рассмотренных выше технологий требует большой кропотливой работы по их совершенствованию. Мы бьемся над ними (даже если некоторые из них могут показаться приемлемыми по качеству), потому что терпение,

которое мы здесь защищаем, слишком легко игнорируется в пылу стремления сделать что-то быстро, более того, часто это даже не считается достоинством. Многие предприниматели из Кремниевой долины стремятся продвигаться как можно быстрее ради очередного прорыва; их главная мантра звучит так: «Сначала выпусти на рынок работающий продукт, прежде чем кто-то опередит тебя, а потом уже беспокойся о проблемах». Недостатком такого подхода является то, что продукт, созданный столь поспешно, зачастую работает лишь в одной ситуации, но его придется полностью переписывать при малейшем ее изменении; или же система работает в деморежиме, но совершенно непригодна для реального мира. Это явление известно как «технический долг» (technical debt): вы быстро (и, возможно, дешево) получаете первую, иногда совершенно ошибочную версию продукта, но потом вам приходится работать над ней задаром, выплачивая дополнительные «проценты» в виде сверхурочной работы, для того чтобы сделать систему более надежной и устранить многочисленные недостатки, а порой даже переделывать все с нуля. Для приложения в социальных сетях это, может быть, и сработает, но в масштабах одной из национальных робототехнических компаний такой подход реально опасен. Неправильные теги в изображениях из Facebook или Instagram могут привести к случайному отключению некоторых пользователей, что, конечно, плохо для компании, но не станет трагедией для человечества, а вот автопилот вашей машины, неправильно идентифицировавший объекты на дороге, или домашний робот, перепутавший кран в мойке с газовым вентилем, легко могут стать причиной смерти многих людей.

В конце концов, не существует идеального рецепта для создания надежных систем искусственного интеллекта, равно как и для технического проектирования в целом. Чтобы достичь успеха, необходимо использовать и координировать обширный набор методов, частично дублирующих друг друга. То, что мы обсудили здесь, — это лишь самое начало.

Подходы, основанные на глубоком обучении и больших данных, создают ряд дополнительных проблем уже хотя бы потому, что они сильно отличаются от традиционных подходов к разработке программного обеспечения.

Тем не менее бóльшая часть программного обеспечения в мире, от веб-браузеров до почтовых клиентов, электронных таблиц и видеоигр, состоит не из систем глубокого обучения, а из классических компьютерных программ: длинных сложных наборов инструкций, тщательно разработанных людьми для конкретных задач. Работа программиста (или команды программистов) состоит в том, чтобы понять некоторую задачу и перевести эту задачу в инструкции, понятные компьютеру.

За исключением тех случаев, когда программа для какого-нибудь приложения оказывается совсем элементарной, даже опытному программисту, вероятно, будет сложно написать ее в полностью готовом виде с первого раза. Как правило, при начальном запуске любая программа ломается и выдает ошибки, так что бóльшая часть дальнейшей миссии программиста состоит в том, чтобы выявить в алгоритме или синтаксисе всех «тараканов», или, говоря

иными словами, отладить программу. Предположим, что наш программист пытается создать клон игры Angry Birds, в котором пылающих тигров нужно швырнуть в проезжающие грузовики с пиццей, чтобы предотвратить эпидемию ожирения. Программист должен будет разработать (или адаптировать) физический движок программы (имитатор физических процессов), определяющий законы игровой вселенной, отслеживая при этом, что происходит с тиграми, когда игрок запускает их в полет, и сталкиваются ли тигры с грузовиками. Кроме того, он будет должен построить и графический движок, чтобы тигры и грузовики с пиццей выглядели естественно и красиво, и систему ввода, позволяющую отслеживать команды пользователей по катапультированию несчастных тигров. Каждый компонент будет иметь свою парадигму (я хочу, чтобы тигры делали то-то и то-то, а затем, когда на экране происходит нечто другое, от тигров требуются такие-то и такие-то действия). Мы не говорим уже о необходимости реалистично изобразить то, что происходит, когда компьютер выполняет все шаги программы (рис. 8.1).

Рис. 8.1. Так выглядит процесс отладки видеоигр

Если все идет хорошо, результат будет радовать: машина делает именно то, что хочет программист. Но в другой раз программист забывает поставить где-то в коде знак препинания, или неправильно устанавливает начальное значение некоторой переменной, или допускает любую из десяти тысяч других возможных ошибок. И в результате вы обнаруживаете тигров, летящих не в ту сторону, или грузовики, которые внезапно появляются неизвестно откуда. Ошибку может выявить и сам программист, однако довольно часто готовая программа передается внутренней команде по обнаружению багов, задача которой состоит, собственно, в том, чтобы их выявлять. Впрочем, если ошибка допущена в неочевидном месте и ее последствия проявляются только в редких или специфических обстоятельствах, на ее поиски могут уйти годы.

Впрочем, основные принципы отладки почти везде одинаковы: речь идет об определении и локализации несоответствий между тем, чего программист хочет добиться от своей программы, и тем, что программа делает фактически в

ходе того, как ее раз за разом выполняет компьютер, вооруженный формальной логикой. Например, программист хочет, чтобы тигр исчез в тот момент, когда он сталкивается с грузовиком, но по какой-то причине в 10% случаев изображение тигра не исчезает после столкновения, и теперь задача программиста — выяснить, почему так происходит. Здесь нет никакой магии — просто, когда программы работают, программисты понимают, почему они работают так, а не иначе, и в чем состоит логика, которой они следуют. Как правило, после того как основная причина сбоя становится понятной, уже нетрудно понять и логику того, почему что-то не работает, а затем устранить ошибки кода, приводящие к сбою.

В большинстве областей естественных наук, таких как фармакология, ситуация сильно отличается от только что описанной. Так, люди пользовались аспирином в течение многих лет, прежде чем удалось получить ясное представление о том, как он работает. Биологические системы настолько сложны, что редко удается целиком и полностью понять действие лекарства. Побочные эффекты от применения лекарств — это правило, а не исключение, потому что мы не можем отлаживать лекарства подобно тому, как отлаживаем компьютерные программы. Наши теории о том, как работают лекарства, уже становятся расплывчатыми, и многое из того, что мы знаем об их воздействии, известно лишь из экспериментов. Каждый раз, когда мы, проводя клинические испытания, обнаруживаем, что улучшение состояния от нового препарата происходит у большого числа людей, а побочные действия отсутствуют или не слишком серьезны, мы решаем, что использовать препарат можно.

Одна из многих проблем глубокого обучения заключается именно в том, что тестирование этих систем больше похоже на методы фармакологии, чем на обычное компьютерное программирование. Разработчики искусственного интеллекта, специализирующиеся на глубоком обучении, в общих чертах понимают, почему сеть, обученная на множестве примеров, может решать новые задачи, руководствуясь усвоенными корреляциями. Однако выбор структуры нейронной сети для работы с конкретными проблемами все еще очень далек от следования принципам, принятым в точных науках; он базируется в основном на экспериментах, а не на более универсальных гипотезах или теориях. Несмотря на то что нейронную сеть относительно легко научить выполнять свою задачу, для нас остается по большей части непонятным то, как именно она работает. В сущности, мы имеем здесь дело со сложной системой узлов, поведение которых определяется сотнями миллионов числовых параметров [43]. За исключением отдельных случаев, человек, конструирующий и обучающий нейронные сети, не имеет ясного представления о том, что делает тот или иной узел и почему каждый из многочисленных параметров сети получает определенное значение. Не могут программисты объяснить и того, почему система выдает правильный ответ, когда она работает правильно, а в других случаях ответ оказывается абсолютно некорректным. Если система глубокого обучения не хочет нормально работать, единственный способ улучшить ситуацию — это прибегнуть к методу проб и ошибок. Альтернативой будет либо внесение изменений в архитектуру сети,

либо создание более полных и релевантных наборов данных для ее обучения. По всем этим причинам в области исследования машинного обучения и в государственной политике использования этих систем за последние годы наметилась тенденция к созданию систем так называемого объяснимого искусственного интеллекта, хотя пока что мы не видим в этой сфере никаких определенных результатов.

В то же время колоссальные объемы человеческих знаний, которые уже можно было бы использовать для улучшения и повышения надежности систем, сегодня почти целиком игнорируются, поскольку остается неясным, как интегрировать их в процесс глубокого обучения. В сфере зрительного восприятия мы знаем очень многое о формах объектов и о том, как возникают изображения. В области лингвистики накоплен богатый материал о структуре языка, фонологии, синтаксисе, семантике и прагматике. В робототехнике существует немалый багаж знаний о физической структуре роботов и их взаимодействиях с внешними объектами. Но если для создания программ искусственного интеллекта мы продолжим использовать одно лишь сквозное глубокое обучение, никакие из знаний никогда не пригодятся: в рамках доминирующих современных подходов ими просто невозможно воспользоваться.

Если бы у Alexa была предусмотрена встроенная система здравого смысла, программа не стала бы внезапно смеяться по ночам — она бы понимала, что люди обычно смеются в конкретных ситуациях, например в ответ на шутки или же когда наблюдают забавные моменты из жизни; того же они ожидают и от систем, претендующих на сходство с человеком. Если бы здравым смыслом обладал пылесос Roomba, он не стал бы размазывать собачьи фекалии по всему полу, справедливо сочтя это малоприятным для людей, и придумал бы для решения этой проблемы что-то получше или, на худой конец, попросил бы людей о помощи. Чат-бот Tay не оказался бы таким черствым и понял, что большинство читателей будут обижены тем, что он опустился до ненавистнических высказываний. Гипотетический робот-дворецкий был бы достаточно осторожен, чтобы не разбить бокалы по дороге к винному бару. И если бы приложение Google Images имело более ясное представление о том, на что похож мир людей, оно бы сразу догадалось, что есть много, очень много матерей с иным цветом кожи, чем белый. Наконец, как мы объясним позже, при наличии у искусственного интеллекта здравого смысла мы, люди, куда реже превращались бы в скрепки для бумаги.

Несложно догадаться, что программы, в которых глубоким было бы именно понимание, а не только обучение, смогли бы избежать большой части тех ошибок и нелепостей, которые делает нынешний искусственный интеллект. Система автоматического ввода текста в iPhone не стала бы бездумно заменять «поклонников» на «покойников», если бы знала, в чем настоящая разница между этими словами. Если бы у Alexa было хоть какое-то представление о том, о каких вещах люди хотят говорить наедине, а какими — делиться с окружающими, она бы точно спросила разрешения у своих владельцев, прежде

чем записывать семейный разговор и отправлять его случайным знакомым. Программа прогнозирования течки у коров тоже поняла бы, что от ее работы нет никакой пользы, если из ее предсказаний получается, будто течки у коров вообще не бывает.

Одна из причин, по которой мы доверяем другим людям, заключается в том, что мы предполагаем, что из одних и тех же исходных данных и аргументов они сделают такие же заключения, какие и мы сами. Если мы хотим действительно доверять нашим интеллектуальным машинам, мы должны сначала убедиться, что от них можно ожидать выводов и действий, сходных с теми, которые сделали бы люди. Если мы с моим роботом отправимся вместе в экспедицию в Гималаи и наткнемся там на волосатую обезьяну в восемь футов высотой (говорят, так выглядит снежный человек), явно голодную и агрессивную, мне бы хотелось, чтобы робот, исходя из общеизвестных сведений о приматах, голоде и агрессии, заключил, что столь огромное клыкастое существо может быть весьма опасным и нам срочно надлежит задуматься об отступлении или убежище. Я бы очень не хотел потратить драгоценные секунды на дискуссию о том, не следует ли вместо этого поближе познакомиться с новым видом приматов, поскольку существуют совершенно надежные данные, что многочисленные встречи туристов с макаками в Азии и лемурами на Мадагаскаре оказались совершенно безобидными (рис. 8.2).

Рис. 8.2. Снежный человек напал на туриста, а робот-помощник в это время анализирует данные в поисках плана защиты

Итак, создание надежных мыслящих машин должно начинаться с построения когнитивных систем с глубоким пониманием мира, намного более глубоким, чем то, которое может предоставить статистика. Сейчас, к сожалению, лишь горстка ученых сосредоточила свои усилия на этой области, хотя на самом деле ее следовало бы поместить в самый центр текущих исследований в области искусственного интеллекта.

Наконец, чтобы машины заслуживали полного доверия, их необходимо наделить этическими ценностями. Конечно, уже один только здравый смысл может подсказать вам, что если человек выпадет из окна высотного здания, это наверняка его убьет; однако также вам потребуются и фундаментальные этические принципы, чтобы твердо понимать, что смерть человека — это плохо. Классическим примером основополагающих законов этики для роботов являются Три закона робототехники Айзека Азимова, опубликованные еще в 1942 году, задолго до появления искусственного интеллекта.

- Закон Первый. Робот не может причинять вред человеку или же своим бездействием позволить, чтобы человеку был нанесен вред.
- Закон Второй. Робот должен подчиняться приказам, отдаваемым ему людьми, за исключением случаев, когда такие приказы противоречат Первому Закону.
- Закон Третий. Робот должен защищать свое собственное существование, если действия, необходимые для защиты, не противоречат Первому или Второму Законам.

Для многих — хотя и сравнительно простых — этических решений, которые роботу придется регулярно принимать в повседневной жизни, законы, сформулированные Азимовым, действительно хороши. Когда робот-компаньон помогает кому-то совершать покупки, он не должен что-либо красть в магазине (даже если владельцы робота велят ему это сделать), потому что это нанесет урон владельцу магазина. Когда робот ведет домой незрячего подопечного, он не должен расталкивать других пешеходов, даже если это позволит человеку, которого робот сопровождает, быстрее добраться до дома. Элементарный список ограничений — «не лги, не мошенничай, не кради и не причиняй вреда» — охватывает очень широкий спектр обстоятельств.

Однако, как замечает специалист по этике из Питтсбургского университета Дерек Лебен, во многих других случаях сформулировать ограничения оказывается гораздо сложнее. Какие виды вреда или ущерба нужно предусмотреть роботу, помимо физических травм, потери имущества, репутации, работы, друзей? Какие виды косвенного вреда случайно или намеренно способен причинить робот? Если он прольет немного кофе на тротуар зимой и кто-то потом поскользнется на кусочке льда, нарушит ли этот робот Первый Закон? Как далеко может (или должен) заходить робот, чтобы люди не пострадали от бездействия? За то время, которое понадобится вам, чтобы прочитать и осмыслить последнее предложение, на Земле умирает в среднем четыре человека — должен ли робот во что бы то ни стало предотвратить эти смерти? Автомобиль без водителя (а он, опять же, не что иное, как робот на колесах), который попытается задуматься обо всех местах, где он может оказаться (например, ради помощи людям), никогда не сможет даже выехать со двора.

Можно представить себе и множество других моральных дилемм, в том числе различные ситуации, в которых, независимо от того, как робот себя поведет, кто-то обязательно будет ранен или погибнет. В 2012 году Гэри описал на страницах журнала *New Yorker* версию классической «дилеммы вагонетки» Филиппы Фут для искусственного интеллекта: что должен делать

автомобиль без водителя, если он видит школьный автобус, полный детей, который потерял управление и несется по мосту по направлению к ограждению? Должен ли автомобиль жертвовать собой и своим владельцем ради спасения школьников от падения в реку или защищать себя и своего владельца любой ценой? Первый закон Азимова здесь не очень поможет, поскольку в любом случае придется пожертвовать жизнями тех или иных людей.

Моральные дилеммы в реальной жизни часто даже еще менее ясны. Во время Второй мировой войны один из учеников философа-экзистенциалиста Жана-Поля Сартра страдал от мучительной дилеммы. Он чувствовал, что должен присоединиться к французскому Сопротивлению и сражаться на войне, однако у него была мать, которая эмоционально полностью зависела от него (муж ее бросил, а второго сына убили на войне). В этой ситуации Сартр смог сказать студенту лишь следующее: «Никакой общий кодекс этики не сможет прямо сказать вам, что вы должны делать». Когда-нибудь в далеком будущем мы, вероятно, создадим столь совершенные машины, что придется задуматься и о таких ситуациях, однако пока что есть и более насущные проблемы.

Ни одна из ныне существующих систем искусственного интеллекта не имеет ни малейшего представления о том, что такое война и тем более что значит «сражаться на войне», и о том, какую ценность представляет для человека его мать и его страна. Тем не менее непосредственные этические проблемы, возникающие перед современным искусственным интеллектом, далеко не столь сложны; для начала нам требуется обеспечить хотя бы то, чтобы искусственный интеллект не совершал вещей, неэтичность которых совершенно очевидна. Скажем, если робот-помощник захочет посодействовать нуждающемуся человеку, у которого проблемы с деньгами, что помешает ему напечатать долларовые купюры на цветном принтере? Ведь робот вполне может подумать, что вреда от подделки купюры, в сущности, будет очень мало: ни один человек, получивший или потративший фальшивую банкноту в будущем, не пострадает, поскольку (как полагает робот) раскрыть махинацию будет невозможно; более того, мир в целом, с точки зрения наивного искусственного интеллекта, станет только лучше, потому что эмиссия дополнительных денежных знаков стимулирует экономику. Тысячи вещей, которые представляются обычному человеку совершенно неприемлемыми этически (хотя, возможно, и логичными), могут выглядеть для мыслящей машины вполне разумными. С другой стороны, мы бы не хотели, чтобы роботы зацикливались на некоторых моральных дилеммах, которые для людей скорее воображаемы, чем реальны. Например, робот-спасатель не должен слишком долго размышлять о том, стоит ли спасать людей из горящего здания, потому что в нем живут правнуки преступников или оккупантов из далекого прошлого и теоретически есть ненулевая вероятность того, что однажды они нанесут вред людям, будучи потомками своих недостойных предков.

В подавляющем большинстве случаев задача искусственного интеллекта заключается не в том, чтобы найти идеальное решение в обстоятельствах, полных тончайших нюансов, подобных тем, что содержатся в дилеммах Софии

или Сартра, а в том, чтобы выбрать правильное действие в достаточно очевидных условиях. Например, бытовому роботу необходимо будет научиться примерно таким рассуждениям: «Можно ли сейчас ударить молотком по гвоздю, чтобы прибить одну доску к другой? Эта доска находится в данном помещении, может ли она повредить людям, если я начну с ней взаимодействовать? Что за люди находятся в этом помещении? Есть ли риск, что они случайно попадут под доску?» Или так: «Было бы хорошо или плохо, если бы я украл в аптеке это лекарство для моей подопечной Мелинды, которая не может позволить себе заплатить за него?»

Мы уже неплохо научились создавать классификаторы шаблонов, которые позволяют компьютерам отличать собак от кошек и золотистого ретривера от лабрадора, но никто не знает, как создать классификаторы моральных аксиом, чтобы машина могла распознавать такие понятия, как «вред» или «конфликт с законом».

Появление универсального искусственного интеллекта потребует разработать и новое законодательство. По закону любой искусственный интеллект, взаимодействующий с людьми на основе принятия самостоятельных решений, должен понимать и уважать основной набор человеческих ценностей. Например, существующие запреты на совершение краж и убийств должны применяться и к мыслящим машинам, а также к тем, кто их проектирует, разрабатывает и использует. Более совершенные методы позволят нам встраивать этические ценности в сознание машин, однако те же самые ценности должны уважаться и соблюдаться людьми и компаниями, которые создают роботов и другие формы искусственного интеллекта и управляют ими. Что не менее важно, но еще сложнее — социальные структуры и люди, формирующие окружение интеллектуальных машин, обязаны будут подчиняться тем же принципам, чтобы давать роботам примеры правильного поведения.

Как только все встанет на свои места — ценности, глубокое понимание, надежные инженерные практики и адекватная нормативно-правовая база, — мы найдем решение и для самых больших проблем всей интеллектуально-технической отрасли, например широко обсуждаемого казуса, называемого «скрепками Ника Бострома» [44]. Мысленный эксперимент Бострома основан на внешне безупречном логическом рассуждении: сверхразумному роботу поставили задачу изготовить как можно большее количество скрепок для бумаги. Будучи умным и исполнительным, он будет делать все, что в его силах, чтобы выполнить эту миссию буквально. Начал бы робот с того, что потребовал весь доступный металл, чтобы сделать как можно больше скрепок, а когда запасы металла закончились, он начал бы эксплуатировать все остальные залежи металла, доступные во вселенной (решив по ходу дела даже проблему межзвездных путешествий). В конце концов, когда все очевидные источники металла истощатся, он начнет добывать атомы металла из человеческих тел, пусть даже они присутствуют у нас лишь в следовых количествах. Как заметил по этому поводу Элиезер Юдковский, «искусственный интеллект не ненавидит вас и не любит вас, просто вы

сделаны из атомов, которые он может использовать для создания чего-то другого». Илон Маск (который написал свои мысли о книге Бострома в Twitter), похоже, был сильно впечатлен этим сценарием, так как предположил, что работа над искусственным интеллектом может однажды «породить демона».

Однако в описанном парадоксе явно есть что-то не то: он предполагает, что у нас имеется некая форма сверхмощного интеллекта, достаточно умная, чтобы справляться с межзвездными путешествиями и понимать людей (которые наверняка будут противодействовать чрезмерной добыче металлов), но до такой степени лишенная здравого смысла, что никогда не осознает, что ее усилия, во-первых, бессмысленны (кому понадобится бесконечное множество скрепок?), а во-вторых, нарушают даже самые основные моральные аксиомы искусственного разума (подобные тем, что предложил Айзек Азимов).

Нам представляются сомнительными даже базовые посылы эксперимента Бострома: возможно ли вообще создать такую систему — суперинтеллектуальную, но при этом полностью лишенную как здравого смысла, так и базовых ценностей? Можете ли вы представить себе искусственный интеллект с достаточно большим объемом знаний о мире и о принципах его функционирования, которому взбредет в голову превратить все материальное в скрепки и который к тому же совершенно не понимает человеческих ценностей? Если как следует поразмышлять о количестве здравого смысла, требуемом для создания суперинтеллекта, то становится более чем очевидно, что сконструировать эффективный и сверхразумный «максимизатор скрепок» — который не знал бы о последствиях своих действий для вселенной [45] — попросту нереально. Если система достаточно умна, чтобы планировать колоссальных масштабов проекты, связанные с изменением материи, то она совершенно точно будет достаточно умна, чтобы предсказать последствия своих действий и распознать конфликт между этими последствиями и основным набором этических ценностей. И этого — то есть здравого смысла в совокупности с Первым Законом Азимова и наличием отказоустойчивой системы, которая полностью остановит деятельность искусственного интеллекта в случае значительного числа человеческих смертей, — должно полностью хватить, чтобы «максимизатор» прекратил свою работу задолго до вселенской катастрофы.

Разумеется, люди, которым пришлась по вкусу притча про скрепки, могут продолжать развивать эту тему до бесконечности. (Поводов для подозрения всегда хватит. Что, если наш «максимизатор» сумеет обмануть людей? Что, если машина просто не позволит людям ее отключить?) Юдковский утверждает, что люди, которые ожидают, что искусственный интеллект окажется безвредным, находятся под влиянием антропоморфизма: они неосознанно полагают, что поскольку люди руководствуются более или менее благими намерениями или, по крайней мере, не хотят уничтожать всю человеческую расу, то искусственный интеллект будет думать и вести себя так же. Так это или нет, в любом случае лучшее решение, на наш взгляд, — не

отдавать процесс формирования искусственного интеллекта на волю случая и не позволять машине выводить все свои ценности только из случайных наблюдений: мы уже видели, к чему приводит подобный подход, на примере чат-бота Тау. Вместо этого от нас потребуется встроить в машину хорошо структурированный набор базовых этических ценностей, и, естественно, все это должно делаться под адекватным юридическим надзором, чтобы системы универсального интеллекта, уже способные нанести существенный вред человечеству и природе, гарантированно понимали мир достаточно для осознания последствий своих действий и учитывали человеческое благополучие и безопасность. Если необходимые меры предосторожности будут приняты, то абсурдные действия машин, влекущие за собой серьезные негативные последствия, не только окажутся вне закона, но и станут крайне трудным для реализации [58].

Так что давайте пока перестанем беспокоиться по поводу скрепок и вместо этого сосредоточимся на том, чтобы привить нашим роботам достаточно здравомыслия для распознавания сомнительных целей и средств. (И, конечно, нам следует быть осторожными, чтобы не дать машинам инструкций, не содержащих вообще никаких ограничений.) Как мы уже подчеркивали, существуют и другие, гораздо более насущные проблемы, чем гипотетические «максимизаторы скрепок», достойные наших лучших умов. Начать надо хотя бы с того, чтобы сконструировать домашних роботов, которые могут достоверно определить, какие их действия безвредны, а какие — нет. Хорошо здесь уже то, что искусственный интеллект — единственная технология, обладающая логическим потенциалом для снижения внутренних рисков, ведь даже самые лучшие ножи не могут рассуждать о последствиях своих действий, но мыслящие машины и роботы когда-нибудь наверняка преодолеют этот барьер.

Оба автора этой книги почерпнули первые сведения об искусственном интеллекте в детстве из научной фантастики, и мы постоянно поражаемся тому, что в них было предсказано верно, а что все еще остается мечтой. Объемы машинной памяти, вычислительная мощность и сетевые технологии, которые сейчас реально упаковать даже в формат умных часов или браслетов, испытали невиданный прогресс. Более того, даже несколько лет назад мы еще не предполагали, что распознавание речи станет настолько вездесущим, как это происходит сейчас. Тем не менее эпоха универсального искусственного интеллекта все еще находится намного дальше, чем мы думали, когда начали мечтать о ней.

Больше всего мы опасаемся не того, что машины захотят уничтожить нас или превратить нас в скрепки, а того, что наши ожидания, связанные с искусственным интеллектом, окажутся намного более оптимистичными, чем следовало бы, исходя из нынешних подходов к его созданию. Современные интеллектуальные технологии не имеют ничего общего со здравым смыслом, но мы все больше полагаемся именно на них. Главный риск для нашей области представляют не гипотетические всемогущие суперроботы, а идиотские

программы, обладающие вычислительной мощностью вместо разума, куда относится, в частности, автономное оружие, которое можно нацелить на людей, без всяких сдерживающих механизмов, равно как и реклама, управляемая безмозглыми системами, которые жертвуют долгосрочными позитивными перспективами ради однодневного успеха.

Сейчас мы, так сказать, находимся в эпохе междуцарствия: повсюду правят хоть и быстродействующие, но слишком узкие системы, объединенные в мощные автономные сети, но не обладающие подлинным интеллектом, достаточным, чтобы рассуждать о влиянии, которые они оказывают на мир и людей. Со временем искусственный интеллект станет, очевидно, еще мощнее, поэтому чем раньше он сможет рассуждать о последствиях своих действий, тем лучше.

Все эти проблемы напрямую связаны с основной темой данной книги. Мы утверждали и продолжаем утверждать, что искусственный интеллект идет сейчас по неверному пути и большинство нынешних усилий направлено на создание по сути не очень разумных машин, которые выполняют лишь самые узкие задачи и полагаются в первую очередь на большие данные, а не на то, что мы называем глубоким пониманием. С нашей точки зрения, это огромная ошибка, поскольку она ведет к своего рода «машинному инфантилизму»: доминирующие ныне системы не способны здраво оценить собственные силы и не имеют возможности задуматься над последствиями своих действий.

Можно было бы, конечно, на какое-то время прекратить развитие искусственного интеллекта, чтобы удостовериться, что он уже не несет в себе серьезных угроз, а затем исправить каждую обнаруженную нами конкретную ошибку. Но в долгосрочной перспективе это не сработает, и даже сейчас (как мы уже продемонстрировали) большинство текущих улучшений представляют собой «бинты» и «пластыри», а не кардинальное и квалифицированное лечение.

Единственный выход из этой путаницы — пробить путь к созданию машин, наделенных здравым смыслом, когнитивными моделями и эффективными инструментами рассуждения. Взятые вместе, эти способности могут стать основой для глубокого понимания, что, в свою очередь является необходимым условием для создания машин, которые смогут надежно предвидеть и оценивать последствия собственных действий. Это очень амбициозный проект, но о его осуществлении можно будет говорить только после того, как приоритеты ученых и разработчиков перестанут фокусироваться на статистике с ее поверхностной зависимостью от больших данных. Лучшее лекарство от ненадежного искусственного интеллекта — это надежный искусственный интеллект, а самый прямой путь к нему лежит через создание мыслящих машин, которые действительно понимают мир.

ЭПИЛОГ

Надежный, заслуживающий доверия искусственный интеллект, основанный на способности к рассуждениям, здравом смысле и правильных инженерных подходах к безопасности и тестированию, сможет серьезно изменить наш мир — будь то через десятилетие или через столетие. За последние 20 лет мы стали свидетелями серьезных технологических инноваций, в основном в виде систем машинного обучения с чистого листа, использующих огромную массу данных. Они неплохо проявляют себя в таких операциях, как распознавание речи, машинный перевод и маркировка изображений. Мы вовсе не предполагаем, что все ограничится только этим. Современные достижения в маркировке изображений и видео будут все больше развиваться; чат-боты станут умнее и надежнее, а способность роботов к самостоятельному передвижению и взаимодействию с различными предметами будет постоянно прогрессировать. Мы увидим все больше эффективных приложений, полезных для социума и природы: уже сейчас системы глубокого обучения используются для мониторинга диких животных и прогнозирования землетрясений и цунами. Разумеется, прогресс коснется и куда менее желательных вещей, таких как реклама, пропаганда и фальшивые новости, не говоря уже о разведке, слежке за людьми и автономном вооружении. И все это произойдет задолго до того, как станет возможной перезагрузка искусственного интеллекта, к которой мы призываем.

Однако по большому счету все перечисленное выше — не более чем прелюдия. Оглядываясь назад, мы будем рассматривать в качестве поворотного момента не возрождение глубокого обучения в 2012 году, а тот день или год, когда решение проблем здравого смысла и способности к рассуждению произведет на свет интеллектуальные системы, обладающие действительно глубоким пониманием. Что это будет означать для нас всех? Никто не знает наверняка; и никто не может претендовать на то, чтобы всерьез предсказать, каким будет это новое будущее. В 1982 году на экраны вышел фильм «Бегущий по лезвию», в котором мир наполнен гуманоидными «репликантами», созданными на основе искусственного интеллекта и выглядящими почти так же, как и люди. Вся ирония научной фантастики сконцентрировалась в этой картине там, где в кульминационный момент истории главный герой Рик Декард (его сыграл актер Харрисон Форд) направляется к телефону-автомату, чтобы сделать решающий звонок. Увы, в реальном мире гораздо проще заменить таксофоны сотовыми телефонами, чем создать искусственный интеллект, способный мыслить и чувствовать на человеческом уровне, но никто из съемочной группы не заметил этого анахронизма. Любые прогнозы, имеющие дело с быстро развивающимися технологиями, — неважно, здесь, на Земле, или «в одной далекой галактике» — обязательно в чем-нибудь окажутся ошибочными.

Тем не менее мы можем позволить себе сделать ряд осторожных, но достаточно обоснованных предположений. Начнем с того, что системы искусственного интеллекта, основанные на глубоком понимании мира, станут первыми представителями машинного разума, которые будут в состоянии сами учиться всему на свете — легко и беспрепятственно, — подобно тому как это

делает ребенок, постоянно расширяя свои познания о мире и часто нуждаясь не более чем в одном или двух примерах для усвоения новых концепций и используя любую ситуацию, чтобы создать на ее основе абстрактную, но практически адекватную модель поведения. Наконец, компьютерный разум действительно сможет понимать романы, фильмы, газетные статьи и видео, созданные людьми. Роботы с глубоким пониманием будут безопасно перемещаться по всему миру и взаимодействовать со всеми типами предметов и материалов, разобравшись самостоятельно, для чего они могут пригодиться. И главное — универсальный машинный интеллект научится свободно, без недоразумений взаимодействовать с людьми.

Как только компьютеры научатся понимать мир и человеческий язык, их возможности станут поистине безграничны. Начнем с того, что значительно улучшится поиск информации. Многие из вопросов, которые ставят в тупик нынешние технологии, — «Кто в настоящее время заседает в Верховном суде?», «Кто был старейшим судьей Верховного суда в 1980 году?», «Что такое крестражи в "Гарри Поттере"?» — окажутся для машин естественными и понятными. Возникнет возможность просить компьютеры о таких вещах, которые нам сейчас просто не приходят в голову по причине полной нереальности получить в ответ что-нибудь разумное. Так, сценарист мог бы сказать своему помощнику-поисковику: «Найди короткий рассказ, где лидер одной страны является агентом другой, и чтобы можно было его превратить в хороший сюжет для фильма». Астроном-любитель сможет поинтересоваться у машины: «Когда в следующий раз будет видно Красное пятно Юпитера?» — принимая во внимание не только чисто астрономическую видимость, но и прогноз погоды. Вы заявите программе видеоигр, что желаете видеть своим аватаром носорога, одетого в рубашку с галстуком, вместо того чтобы выбирать себе образ лишь из дюжины запрограммированных вариантов. Или попросите электронную книгу отслеживать каждое прочитанное вами произведение и определять, сколько времени вы тратите на чтение иностранных авторов, с разбивкой по жанрам.

В дополнение к этому цифровые секретари и помощники теперь смогут делать почти все, на что сейчас способны люди этих профессий, и станут доступными для широкого круга людей, а не только для богачей. Хотите спланировать корпоративный отдых для тысячи сотрудников? Ваш цифровой помощник, обладающий глубоким пониманием людей и жизни, выполнит бóльшую часть работы, начиная с выяснения того, что нужно заказать, и заканчивая необходимыми звонками и напоминаниями для конкретных людей. Эта работа очень похожа на то, что компания Google надеется уже сейчас препоручить (правда, без особого успеха) своему приложению Duplex, — набор номеров и общение с людьми на другом конце линии, — но без всяких ограничений, а не только для бронирования времени в парикмахерской или заказа столика в ресторане. Работа помощников станет одновременно индивидуализированной, с одной стороны (для клиента), и широкомасштабной — с другой (в области сервиса); в нее можно будет вовлечь десятки сотрудников и субподрядчиков, от шеф-поваров до видеооператоров. Ваш

цифровой помощник станет одновременно и связующим звеном, и менеджером проектов, который будет управлять расписанием сразу сотни сотрудников, а не только вашим собственным.

Компьютеры также станут намного проще в использовании. Вам больше не нужно будет просматривать справочные разделы и запоминать бесконечные сочетания клавиш. Если вы вдруг вы захотите выделить курсивом все иностранные слова, вы сможете просто попросить об этом, вместо того чтобы просматривать весь документ, слово за словом. Хотите скопировать 40 различных рецептов с 40 различных веб-страниц, автоматически конвертируя все британские меры веса и объема в метрические и уменьшая все порции в четыре раза? Вместо того чтобы специально искать приложение, которое умеет делать все это (и больше, как правило, ничего), все, что вам нужно будет сделать, — это сформулировать вашу просьбу на английском или любом другом языке, который вы предпочитаете. По сути, все, что мы в настоящее время делаем сами с помощью наших компьютеров, может быть сделано автоматически и с гораздо меньшей суетой. Веб-форумы, в которых подробно описываются мелкие, но жутко мешающие и раздражающие глюки Chrome или PowerPoint, на которые разработчики программного обеспечения просто не удосужились обратить должного внимания (доводя пользователей до белого каления), исчезнут навсегда. Из более далекой перспективы появление свободы в общении с компьютерами будет казаться таким же эпохальным событием, как создание интернета, а может быть, и еще более важным.

«Голопалуба» (прибор виртуальной реальности) из сериала «Звездный путь» тоже станет явью. Хотите пролететь над вулканом Килауэа во время извержения? Или сопровождать Фродо на Роковую гору? Надо будет просто запустить соответствующую симуляцию. Богатые миры виртуальной реальности, представленные в книге и фильме «Первому игроку приготовиться», станут тем, что сможет ощутить на себе каждый из нас. Мы уже сейчас неплохо знаем, как сделать графику весьма правдоподобной, однако системы искусственного интеллекта, основанные на глубоком понимании мира и людей, сделают возможным моделирование реалистичных персонажей, сложных по поведению и очень похожих на людей. А также почти безупречных психологически инопланетян, которые способны вести себя достоверно во всех подробностях, с учетом того, что их тело и разум сильно отличаются от наших.

Между тем бытовые роботы наконец-то станут практичными, надежными и заслуживающими доверия. Теперь они смогут работать в наших домах, готовить, пылесосить полы, покупать продукты, разбирать их, менять лампочки и мыть окна. И, наконец, глубокое понимание станет залогом безопасности беспилотных машин.

Со временем методы, позволяющие машинам достичь понимания мира, расширятся еще больше, и мыслящие компьютеры обретут потенциал людей-экспертов, чтобы суметь выйти за рамки простого здравого смысла и овладеть знаниями, которыми сейчас обладают врачи и ученые.

Когда будет достигнут подобный уровень понимания (возможно, после десятилетий кропотливой работы), машины смогут самостоятельно делать медицинскую диагностику на уровне лучших специалистов, осмыслять юридические дела и документы, преподавать сложные науки и многое другое. Конечно, политические проблемы никуда не денутся, больницы все равно придется убеждать в том, что внедрение искусственного интеллекта экономически оправданно, а энергетикам будет нужно доказывать, что лучше перейти на другие источники энергии, даже если идея этого принадлежит роботам. Кроме того, когда искусственный интеллект станет по-настоящему надежным и полезным, нам удастся преодолеть многие технические проблемы, выглядящие сейчас неразрешимыми.

Компьютерное программирование тоже наконец автоматизируется, и благодаря этому буквально каждый человек получит возможность делать что-то новое, например создавать собственный бизнес или развивать искусства. Изменится и строительная отрасль, поскольку роботы смогут выполнять квалифицированную работу каменщиков, плотников, электриков. Время, необходимое для постройки нового дома, существенно сократится, и стоимость жилья, очевидно, тоже уменьшится. Почти все виды деятельности, опасные для здоровья человека или скучные, утомительные, грязные, станут автоматизированными, независимо от требуемой для них квалификации. Широкое распространение получат и спасательные роботы, включая роботизированных пожарных, которые изначально будут создаваться со всеми необходимыми навыками, начиная от сердечно-легочной реанимации и заканчивая спасением людей под водой.

Художники и музыканты, коллекционеры и любители искусств получат от интеллектуальных роботов-менеджеров всестороннюю поддержку, благодаря которой они смогут радикально расширить рамки любого проекта. Хотите исполнить мелодии The Beatles с группой рок-ботов или дирижировать симфоническим оркестром, состоящим из роботизированных исполнителей, играющих с абсолютной слаженностью? Сыграть парный теннисный матч с вашим супругом против лицензированных реплик сестер Уильямс? Нет проблем! Построить из конструктора лего замок в натуральную величину с населением из роботов-рыцарей, устраивающих средневековые ристалища? Создать флотилию летающих квадрокоптеров, которые соберутся прямо в воздухе в копию Стоунхенджа, когда вы в следующий раз отправитесь на фестиваль Burning Man? Роботы и другие системы искусственного интеллекта придут вам на помощь со всеми расчетами и кропотливыми работами, сделав за вас большую часть задачи. Люди практически во всех областях будут делать то, что никогда не могли себе представить, — каждый сможет стать креативным директором для собственной команды роботов. Вообще, у всех людей появится больше свободного времени, поскольку искусственный интеллект и роботы возьмут на себя основную массу обязанностей, диктуемых повседневной жизнью.

Естественно, не во всех областях прогресс будет одновременным и одинаковым. Мы, вероятно, сможем научить машины глубокому пониманию

на уровне экспертов в одних областях значительно раньше, чем в других. Например, в определенных отраслях точных наук серьезные успехи будут достигнуты уже тогда, когда в решении гуманитарных проблем системы искусственного интеллекта все еще будут оставаться на уровне ниже, чем у ребенка. Совершенные во многих отношениях музыкальные компьютерные помощники наверняка появятся на много лет или десятилетий раньше, чем такие же по уровню помощники для адвокатов и судей. Идеалом прогресса будет машина, которая сможет стать экспертом в любой области, — в свое время появятся и они. И, конечно, темп научных открытий значительно ускорится, если мы объединим вычислительную мощь машин с интеллектуальным программным обеспечением, которое будет соответствовать гибкости и интуитивной мощи специалистов-людей.

После появления универсального искусственного интеллекта продвинутый компьютер сможет сделать все то же самое, что и целая команда живых специалистов, или даже то, на что люди просто не способны, потому что наш мозг, например, не в состоянии отслеживать в памяти взаимосвязи тысяч молекул, тем более — с той математической точностью, которая естественна для машин. Со столь совершенными формами искусственного интеллекта можно будет наконец вскрыть сложные причинно-следственные связи в работе огромного числа нейронов, чтобы выяснить, как работает мозг человека (о чем мы сейчас почти ничего не знаем). Это позволило бы нам изготовлять лекарства, излечивающие психические расстройства (в лечении которых за последние три десятилетия медицина фактически ничего не добилась). Естественно предположить, что системы искусственного интеллекта с подлинным научным мастерством помогут людям разработать более эффективные технологии в области сельского хозяйства и получения экологически чистой энергии. Ничего из этого, конечно, не случится в ближайшее время, да и в дальнейшем это будет нелегкой задачей, потому что сдвинуть искусственный интеллект с мертвой точки невероятно тяжело, — но в конечном счете такой день настанет.

Мы не хотим сказать, что впереди нас ждет идиллия. Разумеется, если окажутся верны предсказания Питера Диамандиса, то массированная автоматизация обеспечит человечеству изобилие, и цена многих вещей, от продуктов питания до электричества, снизится чрезвычайно. При наилучшем развитии событий мы, люди, вполне можем воплотить в жизнь утопию Оскара Уайльда: «развлекаться или наслаждаться культурным отдыхом… делать красивые вещи, или читать прекрасные книги, или просто созерцать мир с восхищением и восторгом с помощью машин… которые возьмут на себя всю рутинную и неприятную работу».

Реальность тем не менее может оказаться не настолько радужной. Число рабочих мест в мире наверняка сильно сократится, и вопрос о гарантированном базовом доходе, как и о перераспределении доходов, встанет куда острее и насущнее, чем сейчас. Даже если экономические проблемы окажутся успешно решены, многим людям будет нелегко отказаться от старых представлений о жизни. Если сейчас чувство удовлетворенности жизнью и собой в основном

базируется на успехах в работе и карьере, то в дальнейшем оно, вероятно, найдет воплощение в осуществлении личных проектов, связанных с искусством и другими видами творчества, поскольку подавляющее количество неквалифицированного труда станет автоматизированным. Конечно, найдутся и такие виды деятельности, которые, по крайней мере на начальном этапе, останутся прерогативой живых людей (например, ремонт самих роботов), и едва ли будет разумным предположить, что новые профессии полностью вытеснят все старые.

С появлением все большего количества свободного времени, снижением цен и отпаданием необходимости в тяжелой работе вся структура общества может сильно измениться к лучшему, хотя при этом трудоустройство станет гораздо сложнее, а неравенство в доходах, скорее всего, возрастет. Можно ожидать, что когнитивная революция в искусственном интеллекте вызовет столько же изменений в обществе, сколько и промышленная революция. Одни из них, несомненно, будут положительными, другие — нет, однако в обоих случаях масштабы изменений окажутся весьма драматичными. Решение проблем искусственного интеллекта не станет всеобщей панацеей, но мы все же верим, что оно в целом принесет позитивные перемены с учетом тех прорывов, которые мыслящие машины способны совершить для науки, медицины, технологического прогресса и сохранения окружающей среды, при условии, разумеется, что мы будем разумны и осторожны в достижении даже самых привлекательных целей.

Означает ли это, что наши потомки будут существовать в мире изобилия, где машины выполняют почти всю тяжелую работу, а прерогатива людей — культурный досуг в духе Уайльда и Диамандиса? Или в мире, в котором мы загружаем своих двойников в облако, как предположил Рэй Курцвейл? Или, заручившись неслыханными достижениями медицины, мы достигнем подлинного бессмертия — как сказал бы Вуди Аллен — более старомодным путем, то есть победив биологическую смерть? Или нам удастся объединить свой мозг с кремниевым процессором? История из романа «Восхищение ботаников» (The Rapture of the Nerds) может произойти с реальным человечеством, а может и не произойти. Случится ли это и как скоро — мы не имеем понятия.

Когда Фалес впервые исследовал электрические явления в 600 году до н э., он понял, что обнаружил нечто новое, но тогда еще было невозможно предсказать, что именно из этого получится; мы сомневаемся, что он фантазировал о том, как развитие электричества приведет к появлению социальных сетей, умных часов или «Википедии». И для нас было бы слишком самонадеянно думать, что мы можем предсказать, что станет с искусственным интеллектом через тысячу или хотя бы пятьсот лет и какое значение он будет иметь для наших отдаленных потомков.

Что мы действительно знаем, так это то, что искусственный интеллект идет по заданному нами пути и что сейчас необходимо сделать все возможное, чтобы его развитие привело к безопасным, заслуживающим доверия и надежным системам, деятельность которых была бы максимально направлена

на помощь человечеству. И лучший способ добиться прогресса в достижении этой цели — выйти за пределы одних только больших данных и глубокого обучения и начать создание более совершенной формы искусственного интеллекта — тщательно разработанной и с самого начала вооруженной человеческими ценностями, здравым смыслом и глубоким пониманием мира.

БЛАГОДАРНОСТИ

Цель этой книги состояла в том, чтобы, во-первых, рассказать людям немного больше об искусственном интеллекте, а во-вторых, бросить вызов современному положению вещей в этой области, в частности объяснить, как машинное обучение и роботы создаются и функционируют сейчас и что здесь можно было бы улучшить. Если в чем-то из этого мы действительно преуспели, то только благодаря неоценимой помощи наших коллег, друзей и членов семьи, у которых зачастую времени тоже было в обрез. Некоторые оказали нам особую любезность, прочтя рукопись от начала и до конца и высказав замечания по всем ее частям; среди них Марк Ахбар, Джои Дэвис, Энни Дьюк, Даг Хофштадтер, Гектор Левек, Кевин Лейтон-Браун, Вик Мохарир, Стив Пинкер, Филипп Рубин, Гарри Ширер, Мануэла Велозу, Атэна Вулуманос и Брэд Вайбл. Другие, в частности Ури Ашер, Род Брукс, Дэвид Чалмерс, Анимеш Гарг, Юйлин Гу и Кэти О'Нил, помогли нам своими ценными комментариями к конкретным главам.

Мы также благодарим группу наших друзей и коллег — Карен Баккер, Леона Ботту, Кенхена Чо, Зака Липтона, Мисси Каммингс, Педро Домингоса, Кена Стэнли, Сидней Левин, Омера Леви, Бена Шнейдермана, Гарри Ширера и Эндрю Сандстрема — за множество рекомендаций и информации, поток которых буквально никогда не останавливался. От Гарри, в частности, мы получили массу интригующих ссылок, ряд которых попал в книгу.

Весьма оживили книгу, на наш взгляд, остроумные и вместе с тем очаровательные рисунки Майяны Харел. Мы также благодарны Майклу Алкорну, Анишу Аталье, Тому Брауну, Кевину Эйхолту, Кунихико Фукусиме, Гари Лупяну, Тайлеру Виджену и Ориолю Виньяльсу за их любезное позволение использовать в книге их фотографии и иллюстрации, а Стиву Пинкеру и Дугу Хофштадтеру за разрешение подробно цитировать их сочинения.

Мы очень обязаны нашему литературному агенту Дэниелу Гринбергу, который помог нам наладить взаимодействие с издателем книги Эдвардом Кастенмайером и всей командой издательства Pantheon.

Особо хотелось бы нам отметить вклад четырех человек. Эдвард Кастенмайер предложил нам конечную структуру книги, которая оказалась чрезвычайно полезной с точки зрения изложения нашей позиции и всей

системы аргументов, не говоря уже о множестве блестящих и проницательных улучшений, сделанных им в ходе редактирования рукописи. Стив Пинкер, чья научная деятельность была источником вдохновения для Гэри в течение трех десятилетий, помог нам переосмыслить общую концепцию нашего труда. Энни Дьюк, только что вернувшаяся к своим занятиям когнитивными науками после короткой экскурсии в мир покерных чемпионатов, вдохновила нас идеями о том, как сделать книгу максимально привлекательной для непрофессионалов. Наконец, Атэна Вулуманос сыграла (как ей это обычно удается) сразу две роли: будучи неизменной опорой для своего супруга Гэри, она одновременно выступила в роли профессионального редактора, прочитав несколько версий чернового варианта книги и каждый раз находя десятки тонких, но действенных способов радикально улучшить наш стиль. Мы оба крайне признательны всем людям, упомянутым выше.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

Ссылки на многие из этих публикаций можно найти на нашем веб-сайте Rebooting.AI.com.

AI в целом. Ведущим учебником по искусственному интеллекту, дающим наиболее полное представление об этой области в целом, является книга Стюарта Рассела и Питера Норвига «Искусственный интеллект: современный подход» (М.: Вильямс, 2007) [59].

Недавняя серия статей одного из ведущих робототехников современности Родни Брукса под общим названием Future of Robotics and Artificial Intelligence («Будущее робототехники и искусственного интеллекта», см. https://rodneybrooks.com/forai-domo-arigato-mr-roboto/) представляется нам весьма читабельной и написанной очень в духе нашей книги. Брукс включил в эту серию множество интересных примеров, касающихся как практических аспектов робототехники, так и истории искусственного интеллекта.

Восторги и скептицизм, связанные с искусственным интеллектом. В нашей области всегда были люди с противоположными взглядами на роль искусственного интеллекта в жизни людей и на возможность создания универсального ИИ. Из ранних работ, посвященных этим вопросам, следует упомянуть книги Джозефа Вейценбаума Computer Power and Human Reason («Власть компьютеров и человеческий разум») и Хьюберта Дрейфуса What Computers Can't Do («Чего не могут компьютеры»). О тех же проблемах рассказывают и более современные работы, в частности The AI Delusion («Иллюзия ИИ») Гэри Смита, Artifictional Intelligence: Against Humanity's Surrender to Computers («Искусственный интеллект: Как не допустить захвата компьютерами власти над людьми») Гарри Коллинза и Artificial Unintelligence: How Computers Misunderstand the World («Искусственное недоразумение: Почему компьютеры понимают мир неправильно») Мередит Бруссард.

Масштабы применения искусственного интеллекта и связанные с этим проблемы ответственности. Недавно было опубликовано несколько важных

книг о краткосрочных и долгосрочных рисках, связанных с применением и прогрессом искусственного интеллекта. Особо следует упомянуть следующие две: Weapons of Math Destruction («Оружие математического разрушения») Кэти О'Нил и Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor («Автоматизация неравенства: как высокие технологии контролируют, преследуют и наказывают бедных») Вирджинии Ойбэнкс. Обе они обсуждают возможности эскалации социального насилия и неравенства, которые влечет за собой использование больших данных и машинного обучения правительствами, страховыми компаниями, полицией, работодателями и т.д.

Машинное и глубокое обучение. Прекрасным, легко читающимся введением в технологию машинного обучения, подробно описывающим все основные подходы в этой области, являются центральные главы книги The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World («Основной алгоритм: Как поиски совершенной самообучающейся машины изменят наш мир») Педро Домингоса. Еще один труд, The Deep Learning Revolution («Революция в глубоком обучении») Терренса Сейновски, излагает взгляды на этот вопрос с точки зрения исторической перспективы и содержит рассказы о жизни ученых, посвятивших себя этому вопросу. Лучшими из недавно вышедших учебников по машинному обучению мы считаем книгу Кевина Мерфи Machine Learning: A Probabilistic Perspective («Машинное обучение с точки зрения вероятностного подхода») и коллективную монографию Deep Learning («Глубокое обучение») Яна Гудфеллоу, Иошуа Бенжио и Аарона Курвилля. Существует множество бесплатных библиотек программного обеспечения для обучения машин, а также обучающих наборов данных, доступных онлайн, в частности Weka Data Mining Software, Pytorch, fast.ai, TensorFlow. Добавим сюда интерактивные ноутбуки Jupyter Зака Липтона и популярный курс Эндрю ЫНа (Andrew Ng) по машинному обучению на Coursera. Руководства для практического применения этих программ включают Introduction to Machine Learning with Python («Введение в машинное обучение на Python») Андреаса Мюллера и Сары Гвидо (Sarah Guido) и Deep Learning with Python («Глубокое обучение на Python») Франсуа Шолле.

Вопросы понимания машинами устной речи и письменного языка людей. Специально для непрофессионалов по этой теме написано не так много, но даже специализированные учебники часто содержат обширные разделы, доступные для широкого круга читателей. В качестве стандартных учебных пособий можно порекомендовать Speech and Language Processing («Компьютерная обработка речи и языка») Дэниела Юрафски и Джеймса Х. Мартина, а также Foundations of Statistical Natural Language Processing («Основы статистической обработки естественного языка») Кристофера Мэннинга и Генриха Шютце. Прекрасным введением в теорию поисковых систем и подобных им программ станет Introduction to Information Retrieval («Введение в информационный поиск») Кристофера Мэннинга, Прабхакара Рагхавана и Генриха Шютце. Как и в случае с машинным обучением, интернет

содержит массу общедоступных библиотек программного обеспечения и наборы данных для желающих попробовать себя в этой сфере. Наиболее широко используемым является инструментарий естественного языка (обычно называемый сокращенно NLTK, см.: https://www.nltk.org, и Stanford Core NLP по адресу https://stanfordnlp.github.io/CoreNLP/). Хорошим руководством по использованию NLTK в программировании является книга Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit («Обработка естественного языка с помощью Python: анализ текста с использованием набора естественных языков») Стивена Берда, Эвана Кляйна и Эдварда Лопера. Наконец, статья Дугласа Хофштадтера The Shallowness of Google Translate («Языковая ограниченность Google Translate»), опубликованная в *The Atlantic* 30 января 2018 года, представляет собой весьма проницательный (и вместе с тем приятный для чтения) анализ ограничений современных подходов к машинному переводу.

Робототехника. Помимо упомянутой в начале этого раздела серии статей в интернете, написанных Родни Бруксом, существует множество других полезных научно-популярных статей и книг о робототехнике. В прекрасной обзорной статье Мэтью Мейсона Toward Robotic Manipulation («О роботизации») различные виды манипуляций обсуждаются как со стороны биологических систем, так и с позиции робототехники. Вводным учебником ко всей области можно считать книгу Кевина Линча и Фрэнка Парка Modern Robotics: Mechanics, Planning, and Control («Современная робототехника: механика, планирование и управление»). Обзор высокоуровневого планирования роботизированных движений и манипуляций содержится в книге Planning Algorithms («Алгоритмы планирования») Стивена Лавалье.

Принципы работы человеческого разума. Литература по этому вопросу практически необозрима. Среди наших любимых книг по лингвистике — работы Стивена Пинкера The Language Instinct («Языковой инстинкт») и Words and Rules: The Ingredients of Language («Слова и правила: Ингредиенты языка»). В области психологии мы бы порекомендовали совместный труд Пинкера и Клюге How the Mind Works («Как работает разум»), книгу Гэри Маркуса The Stuff of Thought («Материя мысли») и «Думай медленно, решай быстро» (М.: АСТ, 2013) Даниэля Канемана. В области эпистемологии следует обратить внимание на книгу Brainstorms («Мозговые штурмы») Даниэля Деннета и произведение Бертрана Рассела Human Knowledge: Its Scope and Limits («Человеческое знание: Его масштаб и границы»). Книга Гэри Algebraic Mind («Алгебраический разум»), написанная в 2001 году, носит более технический характер, однако предвосхищает многие проблемы, повлиявшие на современное глубокое обучение.

Формирование у мыслящих машин здравого смысла. Недавняя статья авторов этой книги Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence («Здравый смысл и общие знания в искусственном интеллекте») похожа на главу 7 этой книги, но длиннее и содержит больше подробностей. В своей книге Common Sense, the Turing Test, and the Quest for Real AI («Здравый смысл, тест Тьюринга и поиски реального искусственного

интеллекта») Гектора Левек стремится, как и мы, доказать, что здравый смысл является самым главным шагом для создания подлинного интеллекта. Написанная Эрнестом Дэвисом книга Representations of Commonsense Knowledge («Формы представления общих знаний и здравого смысла») является отличным учебником по использованию математической логики для формулирования общих знаний. Коллективный труд The Handbook of Knowledge Representation («Справочник по формализованному представлению знаний») под редакцией Франка ван Хармелена, Владимира Лифшица и Брюса Портера представляет собой пособие для углубленного изучения, в котором опубликованы результаты целого ряда конкретных исследований. Книга Джуды Перла и Даны Маккензи The Book of Why: The New Science of Cause and Effect («Почему? Новая наука о причинно-следственных взаимосвязях») посвящена, как следует из ее названия, автоматизации причинно-следственных отношений.

Проблемы надежности искусственного интеллекта и нашего доверия к нему. Вопросы воспитания морального чувства у роботов и других систем искусственного интеллекта обсуждаются в книге Moral Machines: Teaching Robots Right from Wrong («Машины и нравственность: как научить роботов отличать добро от зла») Уэнделла Уоллака и Колина Аллена и в коллективной монографии Robot Ethics: The Ethical and Social Implications of Robotics («Этика роботов: этические и социальные последствия робототехники») под редакцией Патрика Лина, Кита Абни и Джорджа Беки.

Проблемы суперинтеллекта. Уже цитированная нами ранее книга Ника Бострома «Суперинтеллект: Этапы. Угрозы. Стратегии» (М.: Манн, Иванов и Фербер, 2016) ставит проблему того, что искусственный интеллект неизбежно придет к состоянию «сингулярности», в котором быстро приобретает качества суперинтеллекта и выйдет из-под контроля человека. Бостром описывает множество сценариев, от антиутопических до апокалиптических, демонстрируя, чтó каждый из них будет означать для человеческой расы, и обсуждает возможные стратегии развития этой области, позволяющие обеспечить неизменную доброжелательность искусственного интеллекта по отношению к людям.

Будущее искусственного интеллекта. Обсуждение долгосрочного воздействия искусственного интеллекта на жизнь человека и общества является предметом множества публикаций. Среди них можно выделить «Жизнь 3.0: Быть человеком в эпоху искусственного интеллекта» (М.: Corpus, 2019) Макса Тегмарка, «Изобилие. Будущее будет лучше, чем вы думаете» (М.: АСТ, 2018) Питера Диамандиса и Стивена Котлера, «Последнее изобретение человечества. Искусственный интеллект и конец эры Homo sapiens» (М.: АНФ, 2018) Джеймса Баррата, Artificial Intelligence: A Futuristic Approach («Искусственный интеллект: Футуристический подход») Романа Ямпольского и The Fourth Age: Smart Robots, Conscious Computers, and the Future of Humanity («Четвертая эпоха: Умные роботы, мыслящие компьютеры и будущее человечества») Брайона Риса. Еще одна книга, Machines That Think: The Future of Artificial Intelligence («Думающие машины: Будущее искусственного

интеллекта») Тоби Уолша, подробно обсуждает влияние искусственного интеллекта на жизнь, технологии и окружающую среду в краткосрочной и долгосрочной перспективе, особенно в такой сфере, как занятость населения. Кроме того, автор рассматривает различные стратегии и конкретные виды деятельности — от лабораторных исследований до работы компаний, — направленные на обеспечение того, чтобы искусственный интеллект всегда оставался для нас безопасным и приносил пользу.

ПРИМЕЧАНИЯ

ГЛАВА 1

[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.

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

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

ГЛАВА 2

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

[5] См.: http://autocorrectfailness.com/autocorrect-fail-ness-16-im-onnacrapholescreenshots/happy-birthday-dead-papa/.

[6] Lashbrook 2018. Справедливости ради заметим, что проблемы с преобладанием в медицинской литературе данных, основанных исключительно на работе с пациентами-мужчинами, относящимися к белой расе, возникли задолго до начала использования клинических данных в приложениях искусственного интеллекта.

[7] Усилия сообщества ученых, работающих в области искусственного интеллекта, по продвижению запрета на создание автономного оружия со встроенными интеллектуальными программами рассматриваются в Sample 2017 и в Walsh 2018. См. также Future of Life Institute 2015.

ГЛАВА 3

[8] Открытие принципа обратного распространения ошибки приписывается сразу нескольким авторам, работавшим независимо друг от друга в разные годы. Среди них в первую очередь называются Генри Келли — в 1960 году, Артур Брайсон — в 1961 году, Стюарт Дрейфус — в 1962 году, Брайсон и Ю-Чи Хо — в 1969 году, Сеппо Линнайнмаа — в 1970 году, Пол Вербос — в 1974

году, Янн Лекун — в 1984 году, Д. Б. Паркер — в 1985 году и, наконец, Дэвид Румельхарт, Джеффри Хинтон и Рональд Уильямс — в 1986 году. См. «Википедия», статья «Backpropagation», а также Russell and Norvig 2010: 761 и LeCun 2018.

[9] Greg, 2018. Этот удивительный «перевод» быстро стал достоянием широкой общественности; подлинность его подтверждена, в частности, и авторами данной книги.

[10] Marcus 2018a. Алекс Ирпан, инженер-программист из Google, высказал аналогичные замечания и в адрес глубокого обучения с подкреплением: см. Irpan 2018.

[11] Изображение взято отсюда: https://pxhere.com/ru/photo/1341079.

[12] https://spacecraft.ssl.umd.edu/akins_laws.html.

ГЛАВА 4

[13] Более ранняя методика — так называемый латентный семантический анализ (англ. Latent Semantic Analysis) — также позволяла преобразовывать выражения естественного языка в векторы. См. Deerwester, Dumais, Furnas, Landauer and Harshman 1990.

[14] Эксперимент, проведенный авторами 19 апреля 2018 года.

[15] Вопросы «Как называется столица штата Миссисипи?» и «Сколько стоит 1,36 евро в рупиях?» были заданы авторами книги в ходе экспериментов, проведенных в мае 2018 года.

[16] Вопрос, заданный авторами книги в ходе экспериментов, проведенных в мае 2018 года.

[17] Вопрос, заданный авторами книги в ходе экспериментов, проведенных в августе 2018 года. Фрагмент, найденный Google, был исходно опубликован в Ryan 2001–2009.

[18] Мост Аркадико в Греции, построенный около 1300 года до нашей эры, все еще полностью цел. Но это уже весьма сложный по конструкции каменный арочный мост. Нет сомнений, что более примитивные и менее долговечные мосты люди строили еще за столетия или тысячелетия до этого.

[19] Эксперимент проведен в мае 2018 года.

[20] Пресс-центр WolframAlpha 2009. Домашняя страница WolframAlpha: https://www.wolframalpha.com.

[21] Эксперименты с WolframAlpha, проведенные авторами в мае 2018 года.

[22] Поиск проводился в мае 2018 года. Работа IBM Watson Assistant демонстрируется на следующей веб-странице: https://watson-assistant-demo.ng.bluemix.net/.

[23] Эксперимент, проведенный авторами в августе 2018 года. Эрнест Дэвис поддерживает веб-сайт с небольшой коллекцией ошибок, допущенных ведущими программами машинного перевода при интерпретации предложений, которые являются максимально простыми с точки зрения лингвистики: см. https://cs.nyu.edu/faculty/davise/papers/GTFails.html.

[24] Эксперимент, проведенный авторами в августе 2018 года. Совершенно такие же ошибки Google Translate делает при переводе данного предложения на немецкий, испанский и итальянский языки.

ГЛАВА 5

[25] В частности, весной 2018 года компания Sony выпустила обновленную версию своего робота-пса Aibo: Hornyak 2018.

[26] Первое поколение Roomba, вышедшее на рынок в 2002 году, использовало компьютер с 256 байтами доступной для записи памяти. Это не опечатка: память первых пылесосов составляла примерно одну миллиардную часть памяти современных iPhone. См. Ulanoff 2002.

[27] Анимеш Гарг, электронное письмо авторам от 24 октября 2018 года.

ГЛАВА 6

[28] Ball, 2013a. Позднее Филип Болл несколько пересмотрел свои взгляды, судя по записям в его блоге: Ball, 2013b.

[29] См. веб-сайт Висснер-Гросса: http://www.alexwg.org.

[30] Эти и другие ограничения встраивания слов обсуждаются в статье Леви, которая сейчас находится в процессе подготовки.

[31] Слова Рэя Муни цитируются здесь дословно, вместе со всеми словами, удаленными в Conneau et al. 2018 из соображений политкорректности. («You can't cram the meaning of an entire fucking sentence into a single fucking vector!»)

[32] Эксперимент, выполненный авторами с приложением Amazon Web Services в августе 2018 года.

ГЛАВА 7

[33] Данный перечень является результатом тестирования системы NELL, выполненного авторами 28 мая 2018 года.

[34] Puig et al. 2018. Проект VirtualHome имеет свой веб-сайт: https://www.csail.mit.edu/research/virtualhome-representing-activities-programs.

[35] Аналогичные возражения были высказаны в Dreyfus 1979.

[36] В качестве недавнего обзора этой работы можно привести публикацию Davis 2017; более ранние исследования представлены у Davis 1990 и Harmelen van, Lifschitz and Porter 2008.

[37] Проект CYC был впервые анонсирован в Lenat, Prakash, and Shepherd 1985. В 1990 году был опубликован сводный отчет о проделанной работе: Lenat and Guha 1990. С тех пор в печати не появилось ни одного подробного отчета.

[38] Одна из попыток справиться с неопределенными сущностями и отношениями — это так называемая нечеткая логика, разработанная Лотфи Заде: Zadeh 1987.

[39] В частности, можно предложить варианты семантических сетей, значение которых будет определяться точно так же, как и логическая запись. См.: Brachman and Schmolze 1989; Borgida and Sowa 1991.

[40] Кант, 1751/1998. Стивен Пинкер высказывается в пользу аналогичной точки зрения в своей книге The Stuff of Thought: Pinker 2007.

ГЛАВА 8

[41] Леон Ботту, электронное письмо авторам от 19 июля 2018 года.

[42] Альтернативы тесту Тьюринга обсуждаются в Marcus, Rossi, and Veloso 2016; см. также: Reddy, Chen, and Manning 2018; Wang et al. 2018 и веб-сайт Института искусственного интеллекта Аллена (https://allenai.org/).

[43] См., например, Vaswani et al. 2017 (табл. 3), а также Canziani, Culurciello and Paszke 2017 (рис. 2)

[44] См. Bostrom 2003. С тех пор этот мысленный эксперимент широко обсуждался многими авторами, в частности самим Ником Бостромом, а также Элиезером Юдковским и их сотрудниками. Наша дискуссия здесь основана главным образом на следующих публикациях: Bostrom 2014; Yudkowsky 2011; Bostrom and Yudkowsky 2014; Soares, Fallenstein, Armstrong and Yudkowsky 2015.

[45] Аналогичные аргументы представлены в Pinker 2018 и в Brooks 2017c.

БИБЛИОГРАФИЯ

Ссылки на многие из указанных ниже источников можно найти на Rebooting.AI.com.

Agrawal, Aishwarya, Dhruv Batra, and Devi Parikh. 2016. "Analyzing the behavior of visual question answering models." *arXiv preprint arXiv:1606.07356*. https://arxiv.org/abs/1606.07356.

Alcorn, Michael A., Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2018. "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects." *arXiv preprint arXiv:1811.11553*. https://arxiv.org/abs/1811.11553.

Allen, Tom. 2018. "Elon Musk admits 'too much automation' is slowing Tesla Model 3 production." *The Inquirer*. April 16, 2018. https://www.theinquirer.net/inquirer/news/3030277/elon-musk-admits-too-much-automation-is-slowing-tesla-model-3-production.

AlphaStar Team. 2019. "AlphaStar: Mastering the real-time strategy game StarCraft II." https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

Amodei, Dario, Paul Christiano, and Alex Ray. 2017. "Learning from human preferences." *OpenAI Blog*. June 13, 2017. https://blog.openai.com/deep-reinforcement-learning-from-human-preferences/.

Amunts, Katrin, and Karl Zilles. 2015. "Architectonic mapping of the human brain beyond Brodmann." *Neuron* 88(6): 1086–1107. https://doi.org/10.1016/j.neuron.2015.12.001.

Arbib, Michael. 2003. *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.

Asimov, Isaac. 1942. "Runaround." *Astounding Science Fiction*. March 1942. Included in Isaac Asimov, *I, Robot*, Gnome Press, 1950.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. "Synthesizing robust adversarial examples." *Proc. 35th Intl. Conf. on Machine Learning*. http://proceedings.mlr.press/v80/athalye18b/athalye18b.pdf.

Baer, Drake. 2014. "Mark Zuckerberg explains why Facebook doesn't 'move fast and break things' anymore." *Business Insider*, May 2, 2014. https://www.businessinsider.com/mark-zuckerberg-on-facebooks-new-motto-2014-5.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473*. https://arxiv.org/abs/1409.0473.

Baillargeon, Renee, Elizabeth S. Spelke, and Stanley Wasserman. 1985. "Object permanence in five-month-old infants." *Cognition* 20(3): 191–208. https://doi.org/10.1016/0010-0277(85)90008-3.

Ball, Philip. 2013a. "Entropy strikes at the *New Yorker*." *Homunculus* blog. May 9, 2013. http://philipball.blogspot.com/2013/05/entropy-strikes-at-new-yorker.html.

Ball, Philip. 2013b. "Stuck in the middle again." *Homunculus* blog. May 16, 2013. http://philipball.blogspot.com/2013/05/stuck-in-middle-again.html. Bardin, Noam. 2018. "Keeping cities moving — how Waze works." *Medium.com*. April 12, 2018. https://medium.com/@noambardin/keeping-cities-moving-how-waze-works-4aad066c7bfa.

Barlas, Gerassimos. 2015. *Multicore and GPU Programming*. Amsterdam: Morgan Kaufmann.

Barlow, Jerome, Leda Cosmides, and John Tooby. 1996. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Barrat, James. 2013. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York: Thomas Dunne Books/St. Martin's Press.

BBC Technology. 2016. "IBM AI system Watson to diagnose rare diseases in Germany." October 18, 2016. https://www.bbc.com/news/technology-37653588.

Benger, Werner. 2008. "Colliding galaxies, rotating neutron stars and merging black holes — visualizing high dimensional datasets on arbitrary meshes." *New Journal of Physics* 10(12): 125004. http://dx.doi.org/10.1088/1367-2630/10/12/125004.

Berkeley CIR. 2018. Control, Intelligent Systems and Robotics (CIR). Website. https://www2.eecs.berkeley.edu/Research/Areas/CIR/.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Cambridge, MA: O'Reilly Pubs.

Bojarski, Mariusz, et al. 2016. "End-to-end deep learning for self-driving cars." *NVIDIA Developer Blog*. https://devblogs.nvidia.com/deep-learning-self-driving-cars/.

Borgida, Alexander, and John Sowa. 1991. *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann.

Boston Dynamics. 2016. "Introducing SpotMini." Video. https://www.youtube.com/watch?v=tf7IEVTDjng.

Boston Dynamics. 2017. *What's New, Atlas?* Video. https://www.youtube.com/watch?v=fRj34o4hN4I.

Boston Dynamics. 2018a. "Atlas: The world's most dynamic humanoid."https://www.bostondynamics.com/atla.

Boston Dynamics. 2018b. "BigDog: The first advanced rough-terrain robot." https://www.bostondynamics.com/bigdog.

Boston Dynamics. 2018c. "WildCat: The world's fastest quadruped robot." https://www.bostondynamics.com/wildcat.

Bostrom, Nick. 2003. "Ethical issues in advanced artificial intelligence." *Science Fiction and Philosophy: From Time Travel to Superintelligence*, edited by Susan Schneider. 277–284. Hoboken, NJ: Wiley and Sons.

Bostrom, Nick. 2014. *SuperIntelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The ethics of artificial intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey, 316–334. Cambridge: Cambridge University Press.

Bot Scene 2013. "Entropica claims 'powerful new kind of AI.' "*Bot Scene* blog. May 11, 2013. https://botscene.net/2013/05/11/entropica-claims-powerful-new-kind-of-ai/comment-page-1/.

Bottou, Léon. 2018. Foreword. In Marvin Minsky and Seymour Papert, *Perceptrons: An Introduction to Computational Geometry*. Reissue of the 1988 expanded edition, with a new foreword by Léon Bottou. Cambridge, MA: MIT Press.

Brachman, Ronald J., and James G. Schmolze. "An overview of the KL-ONE knowledge representation system." In *Readings in Artificial Intelligence and Databases*, edited by John Myopoulos and Michael Brodie, 207–230. San Mateo, CA: Morgan Kaufmann, 1989.

Brady, Paul. 2018. "Robotic suitcases: The trend the world doesn't need." *Condé Nast Traveler*. January 10, 2018. https://www.cntraveler.com/story/robotic-suitcases-the-trend-the-world-doesnt-need.

Brandom, Russell. 2018. "Self-driving cars are headed toward an AI roadblock." *The Verge*. July 3, 2018. https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber.

Braun, Urs, Axel Schäfer, Henrik Walter, Susanne Erk, Nina Romanczuk-Seiferth, Leila Haddad, Janina I. Schweiger, et al. 2015. "Dynamic reconfiguration of frontal brain networks during executive cognition in humans." *Proceedings of the*

National Academy of Sciences 112(37): 11678–11683. https://doi.org/10.1073/pnas.1422487112.

Bright, Peter. 2016. "Tay, the neo-Nazi millennial chatbot, gets autopsied." *Ars Technica*. May 25, 2016. https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/.

Briot, Jean-Pierre, Gaëtan Hadjeres, and François Pachet. 2017. Deep learning techniques for music generation — a survey. *arXiv preprint arXiv:1709.01620*. https://arxiv.org/abs/1709.01620.

Brooks, Rodney. 2017a. "Future of robotics and artificial intelligence."https://rodneybrooks.com/forai-future-of-robotics-and-artificial-intelligence/.

Brooks, Rodney. 2017b. "Domo Arigato Mr. Roboto." http://rodneybrooks.com/forai-domo-arigato-mr-roboto/.

Brooks, Rodney. 2017c. "The seven deadly sins of predicting AI." http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press.

Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. "Adversarial patch." *arXiv preprint arXiv:1712.09665*. https://arxiv.org/abs/1712.09665.

Bughin, Jacques, Jeongmin Seong, James Manyika, Michael Chui, and Raoul Joshi. 2018. "Notes from the frontier: Modeling the impact of AI on the world economy." McKinsey and Co. September 2018. https://www.mckinsey.com/featured-insights/artificialintelligence/notes-from-the-frontier-modeling-the-impact-of-ai-on-the-world-economy.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on Fairness, Accountability and Transparency, 2018*. 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html.

Burns, Janet. 2017. "Finally, the perfect app for superfans, stalkers, and serial killers." *Forbes*. June 23, 2017. https://www.forbes.com/sites/janetwburns/2017/06/23/finally-the-perfect-dating-app-for-superfans-stalkers-and-serial-killers/#4d2b54c9f166.

Bushnell, Mona. 2018. "AI faceoff: Siri vs. Cortana vs. Google Assistant vs. Alexa." *Business News Daily*. June 29, 2018. https://www.businessnewsdaily.com/10315-siri-cortana-google-assistant-amazon-alexa-face-off.html.

Cable Car Museum, undated. "The Brakes." http://www.cablecarmuseum.org/the-brakes.html. Accessed by the authors, December 29, 2018.

Callahan, John. 2019. "What is Google Duplex, and how do you use it?" *Android Authority*, March 3, 2019. https://www.androidauthority.com/what-is-google-duplex-869476/.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. *AI Now 2017 Report*. https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Canales, Katie. 2018. "A couple says that Amazon's Alexa recorded a private conversation and randomly sent it to a friend." *Business Insider*. May 24, 2018. http://www.businessinsider.com/amazon-alexa-records-private-conversation-2018-5.

Canziani, Alfredo, Eugenio Culurciello, and Adam Paszke. 2017. "Evaluation of neural network architectures for embedded systems." In *IEEE International Symposium on Circuits and Systems (ISCAS), 2017*. 1–4. https://ieeexplore.ieee.org/abstract/document/8050276/.

Carey, Susan. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carmichael, Leonard, H. P. Hogan, and A. A. Walter. 1932. "An experimental study of the effect of language on the reproduction of visually perceived form." *Journal of Experimental Psychology* 15(1): 73. http://dx.doi.org/10.1037/h0072671.

Chaplot, Devendra Singh, Guillaume Lample, Kanthashree Mysore Sathyendra, and Ruslan Salakhutdinov. 2016. "Transfer deep reinforcement learning in 3d environments: An empirical study." In *NIPS Deep Reinforcement Learning Workshop*. http://www.cs.cmu.edu/~rsalakhu/papers/DeepRL_Transfer.pdf.

Chintala, Soumith, and Yann LeCun, 2016. "A path to unsupervised learning through adversarial networks." Facebook AI Research blog, June 20, 2016. https://code.fb.com/ml-applications/a-path-to-unsupervised-learning-through-adversarial-networks/.

Chokshi, Niraj. 2018. "Amazon knows why Alexa was laughing at its customers." *New York Times*, March 8, 2018. https://www.nytimes.com/2018/03/08/business/alexa-laugh-amazon-echo.html.

Chomsky, Noam. 1959. "A review of B. F. Skinner's *Verbal Behavior*." *Language* 35(1): 26–58. http://doi.org/10.2307/411334.

Chu-Carroll, Jennifer, James Fan, B. K. Boguraev, David Carmel, Dafna Sheinwald, and Chris Welty. 2012. "Finding needles in the haystack: Search and candidate generation" *IBM Journal of Research and Development* 56(3–4): 6:1–6:12. https://doi.org/10.1147/JRD.2012.2186682.

CNBC. 2018. *Boston Dynamics' Atlas Robot Can Now Do Parkour*. Video. https://www.youtube.com/watch?v=hSjKoEva5bg.

Coldewey, Devin. 2018. "Judge says 'literal but nonsensical' Google translation isn't consent for police search." *TechCrunch*, June 15, 2018. https://techcrunch.com/2018/06/15/judge-says-literal-but-nonsensical-google-translation-isnt-consent-for-police-search/.

Collins, Allan M., and M. Ross Quillian. 1969. "Retrieval time from semantic memory." *Journal of Verbal Learning and Verbal Behavior* 8(2): 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1.

Collins, Harry. 2018. *Artifictional Intelligence: Against Humanity's Surrender to Computers*. New York: Wiley.

Conesa, Jordi, Veda C. Storey, and Vijayan Sugumaran. 2010. "Usability of upper level ontologies: The case of ResearchCyc." *Data & Knowledge Engineering* 69(4): 343–356. https://doi.org/10.1016/j.datak.2009.08.002.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. "What you can cram into a single vector: Probing sentence embeddings for linguistic properties." *arXiv preprint arXiv:1805.01070* https://arxiv.org/pdf/1805.01070.pdf.

Corbett, Erin, and Jonathan Vanian. 2018. "Microsoft improves biased facial recognition technology." *Fortune*. June 27, 2018. http://fortune.com/2018/06/27/microsoft-biased-facial-recognition/.

Crick, Francis. 1989. "The recent excitement about neural networks." *Nature* 337(6203): 129–132. https://doi.org/10.1038/337129a0.

Cuthbertson, Anthony. 2018. "Robots can now read better than humans, putting millions of jobs at risk." *Newsweek*. January 15, 2018. https://www.newsweek.com/robots-can-now-read-better-humans-putting-millions-jobs-risk-781393.

Damiani, Jesse. 2018. "Tesla Model S on Autopilot crashes into parked police vehicle in Laguna Beach." *Forbes*. May 30, 2018. https://www.forbes.com/sites/jessedamiani/2018/05/30/tesla-model-s-on-autopilot-crashes-into-parked-police-vehicle-in-laguna-beach/#7c5d245d6f59.

Darwiche, Adnan. 2018. "Human-level intelligence or animal-like abilities?" *Communications of the ACM* 61(10): 56–67. https://cacm.acm.org/magazines/2018/10/231373-human-level-intelligence-or-animal-like-abilities/fulltext.

Dastin, Jeffrey. 2018. "Amazon scraps secret AI recruiting tool that showed bias against women." Reuters. October 10, 2018. https://www.reuters.com/article/amazoncom-jobs-automation/rpt-insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1WP1RO.

Davies, Alex. 2017. "Waymo has taken the human out of its self-driving cars." *WIRED*. November 7, 2017. https://www.wired.com/story/waymo-google-arizona-phoenix-driverless-self-driving-cars/.

Davies, Alex. 2018. "Waymo's so-called Robo-Taxi launch reveals a brutal truth." *WIRED*. December 5, 2018. https://www.wired.com/story/waymo-self-driving-taxi-service-launch-chandler-arizona/.

Davis, Ernest. 1990. *Representations of Commonsense Knowledge*. San Mateo, CA: Morgan Kaufmann.

Davis, Ernest. 2016a. "How to write science questions that are easy for people and hard for computers." *AI Magazine* 37(1): 13–22.

Davis, Ernest. 2016b. "The tragic tale of Tay the Chatbot." *AI Matters* 2(4). https://cs.nyu.edu/faculty/davise/Verses/Tay.html.

Davis, Ernest. 2017. "Logical formalizations of commonsense reasoning." *Journal of Artificial Intelligence Research* 59: 651–723. https://jair.org/index.php/jair/article/view/11076.

Davis, Ernest, and Gary Marcus. 2015. "Commonsense reasoning and commonsense knowledge in artificial intelligence." *Communications of the ACM* 58(9): 92–105.

Davis, Ernest, and Gary Marcus. 2016. "The scope and limits of simulation in automated reasoning." *Artificial Intelligence* 233: 60–72. http://dx.doi.org/10.1016/j.artint.2015.12.003.

Davis, Randall, and Douglas Lenat. 1982. *Knowledge-Based Systems in Artificial Intelligence*. New York: McGraw-Hill.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science* 41(6): 391–407. https://doi.org/10.1002/(SICI)10974571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Image-net: A large-scale hierarchical image database." *IEEE Conference on Computer Vision and Pattern Recognition, 2009.* 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Dennett, Daniel. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Devlin, Hannah. 2015. "Google a step closer to developing machines with human-like intelligence." *The Guardian*. May 21, 2015. https://www.theguardian.com/science/2015/may/21/google-a-step-closer-to-developing-machines-with-human-like-intelligence.

Diamandis, Peter, and Steven Kotler. 2012. *Abundance: The Future Is Better Than You Think.* New York: Free Press.

Domingos, Pedro. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* New York: Basic Books.

D'Orazio, Dante. 2014. "Elon Musk says artificial intelligence is 'potentially more dangerous than nukes.'" *The Verge*. August 3, 2014. https://www.theverge.com/2014/8/3/5965099/elon-musk-compares-artificial-intelligence-to-nukes.

Dreyfus, Hubert. 1979. *What Computers Can't Do: The Limits of Artificial Intelligence.* Rev. ed. New York: Harper and Row.

Dreyfuss, Emily. 2018. "A bot panic hits Amazon's mechanical Turk." *WIRED*. August 17, 2018. https://www.wired.com/story/amazon-mechanical-turk-bot-panic/.

Dyer, Michael. 1983. *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension.* Cambridge, MA: MIT Press.

Estava, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542(7639): 115–118.

The Economist. 2018. "AI, radiology, and the future of work." June 7, 2018. https://www.economist.com/leaders/2018/06/07/ai-radiology-and-the-future-of-work.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press.

Evans, Jonathan St. B. T. 2012. "Dual process theories of deductive reasoning: Facts and fallacies." In *The Oxford Handbook of Thinking and Reasoning*, 115–133. Oxford: Oxford University Press.

Evarts, Eric C. 2016. "Why Tesla's Autopilot isn't really autopilot." *U.S. News and World Report Best Cars*. August 11, 2016. https://cars.usnews.com/cars-trucks/best-cars-blog/2016/08/why-teslas-autopilot-isnt-really-autopilot.

Evtimov, Ivan, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2017. "Robust physical-world attacks on machine learning models." *arXiv preprint arXiv:1707.08945*.https://arxiv.org/abs/1707.08945.

Fabian. 2018. "Global artificial intelligence landscape." *Medium.com*. May 22, 2018. https://medium.com/@bootstrappingme/global-artificial-intel-ligence-landscape-including-database-with-3-465-ai-companies-3bf01a175c5d.

Falcon, William. 2018. "The new Burning Man — the AI conference that sold out in 12 minutes." *Forbes*. September 5, 2018. https://www.forbes.com/sites/williamfalcon/2018/09/05/the-new-burning-man-the-ai-conference-that-sold-out-in-12minutes/#38467b847a96.

Felleman, Daniel J., and D. C. van Essen. 1991. "Distributed hierarchical processing in the primate cerebral cortex." *Cerebral Cortex* 1(1): 1–47. https://doi.org/10.1093/cercor/1.1.1-a.

Fernandez, Ernie. 2016. "How cognitive systems will shape the future of health and wellness." *IBM Healthcare and Life Sciences Industry Blog*. November 16, 2016. https://www.ibm.com/blogs/insights-on-business/healthcare/cognitive-systems-shape-health-wellness/.

Ferrucci, David, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, et al. 2010. "Building Watson: An overview of the DeepQA project." *AI Magazine* 31(3): 59–79. https://www.aaai.org/ojs/index.php/aimagazine/article/view/2303.

Firestone, Chaz, and Brian J. Scholl. 2016. "Cognition does not affect perception: Evaluating the evidence for 'top-down' effects." *Behavioral and Brain Sciences* 39, e229. https://doi.org/10.1017/S0140525X15000965.

Ford, Martin. 2018. *Architects of Intelligence: The Truth About AI from the People Building It.* Birmingham, UK: Packt Publishing.

Fukushima, Kunihiko, and Sei Miyake. 1982. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition." In *Competition and Cooperation in Neural Nets: Proceedings of the U.S. – Japan Joint Seminar*, 267–285. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-46466-9_18.

Fung, Brian. 2017. "The driver who died in a Tesla crash using Autopilot ignored at least 7 safety warnings." *Washington Post*. June 20, 2017. https://www.washingtonpost.com/news/theswitch/wp/2017/06/20/the-driver-who-died-in-a-tesla-crash-using-autopilot-ignored-7-safety-warnings/.

Future of Life Institute. 2015. "Autonomous weapons: An open letter from AI & robotics researchers."https://futureoflife.org/open-letter-autonomous-weapons/.

Gardner, Howard. 1983. *Frames of Mind: The Theory of Multiple Intelligences.* New York: Basic Books.

Garrahan, Matthew. 2017. "Google and Facebook dominance forecast to rise." *Financial Times.* December 3, 2017. https://www.ft.com/content/cf362186-d840-11e7-a039c64b1c09b482.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. 2016. "Image style transfer using convolutional neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423. https://www.cvfoundation.org/openaccess/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html.

Geirhos, Robert, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. 2018. "Generalisation in humans and deep neural networks." In *Advances in Neural Information Processing Systems*, 7549–7561. http://papers.nips.cc/paper/7982-generalisation-in-humans-and-deep-neural-networks.

Gelman, Rochel, and Renee Baillargeon. 1983. "Review of some Piagetian concepts." In *Handbook of Child Psychology: Formerly Carmichael's Manual of Child Psychology*, edited by Paul H. Mussen. New York: Wiley.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. "Neural networks and the bias/variance dilemma." *Neural Computation* 4(1): 1–58. https://doi.org/10.1162/neco.1992.4.1.1.

Gershgorn, Dave. 2017. "The data that transformed AI research — and possibly the world." *Quartz.* July 26, 2017. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.

Gibbs, Samuel. 2014. "Google buys UK artificial intelligence startup Deepmind for £400m." *The Guardian.* January 27, 2014. https://www.theguardian.com/technology/2014/jan/27/google-acquires-uk-artificial-intelligence-startup-deepmin.

Gibbs, Samuel. 2018. "SpotMini: Headless robotic dog to go on sale in 2019." *The Guardian*. May 14, 2018. https://www.theguardian.com/technology/2018/may/14/spotmini-robotic-dog-sale-2019-former-google-boston-dynamics.

Glaser, April. 2018. "The robot dog that can open a door is even more impressive than it looks." *Slate*. February 13, 2018. https://slate.com/technology/2018/02/the-robot-dog-that-can-open-a-door-is-even-more-impressive-than-it-looks.html.

Glasser, Matthew, et al. 2016. "A multi-modal parcellation of human cerebral cortex." *Nature* 536: 171–178. https://doi.org/10.1038/nature18933.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. "Deep sparse rectifier neural networks." In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323. http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf.

Goode, Lauren. 2018. "Google CEO Sundar Pichai compares impact of AI to electricity and fire." *The Verge*. Jan. 19, 2018. https://www.theverge.com/2018/1/19/16911354/google-ceo-sundar-pichai-ai-artificial-intelligence-fire-electricity-jobs-cancer.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2015. *Deep Learning*. Cambridge, MA: MIT Press.

Greenberg, Andy. 2015. "Hackers remotely kill a Jeep on the highway — with me in it." *WIRED*. July 21, 2015. https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/.

Greenberg, Andy. 2017. "Watch a 10-year-old's face unlock his mom's iPhone X." *WIRED*. November 14, 2017. https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/.

Greg. 2018. "Dog's final judgement: Weird Google Translate glitch delivers an apocalyptic message." *Daily Grail*, July 16, 2018. https://www.dailygrail.com/2018/07/dogs-final-judgement-weird-google-translate-glitch-delivers-an-apocalyptic-message/.

Gunning, David. 2017. "Explainable artificial intelligence (xai)." *Defense Advanced Research Projects Agency (DARPA)*. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Hall, Phil. 2018. "Luminar's smart Sky Enhancer filter does the dodging and burning for you." *Techradar: The Source for Tech Buying Advice*. November 2, 2018. https://www.techradar.com/news/luminars-smart-sky-enhancer-filter-does-the-dodging-and-burning-for-you.

Harford, Tim. 2018. "What we get wrong about technology." *Financial Times*. July 7, 2018. https://www.ft.com/content/32c31874-610b-11e7-8814-0ac7eb84e5f1.

Harridy, Rich. 2018. "Boston Dynamics Atlas robot can now chase you through the woods." *New Atlas*. May 10, 2018. https://newatlas.com/boston-dynamics-atlas-running/54573/.

Harwell, Drew. 2018. "Elon Musk said a Tesla could drive itself across the country by 2018. One just crashed backing out of a garage." *Washington Post*. September 13, 2018. https://www.washingtonpost.com/technology/2018/09/13/elon-musk-said-tesla-could-drive-itself-across-country-by-one-just-crashed-backing-out-garage/.

Harwell, Drew, and Craig Timberg. 2019. "YouTube recommended a Russian media site thousands of times for analysis of Mueller's report, a watchdog group says." *The Washington Post*, April 26, 2019. https://www.washingtonpost.com/technology/2019/04/26/youtube-recommended-russian-media-site-above-all-others-analysis-mueller-report-watchdog-group-says/?utm_term=.39b7bcf0c8a4.

Havasi, Catherine, Robert Speer, James Pustejovsky, and Henry Lieberman. 2009. "Digital intuition: Applying common sense using dimensionality reduction." *IEEE Intelligent systems* 24(4): 24–35. https://doi.org/10.1109/MIS.2009.72.

Hawkins, Andrew. 2018. "Elon Musk still doesn't think LIDAR is necessary for fully driverless cars." *The Verge*. February 7, 2018. https://www.theverge.com/2018/2/7/16988628/elon-musk-lidar-self-driving-car-tesla.

Hawkins, Jeff, and Sandra Blakeslee. 2004. *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines.* New York: Times Books.

Hayes, Gavin. 2018. "Search 'idiot,' get Trump: How activists are manipulating Google Images." *The Guardian.* July 17, 2018. https://www.theguardian.com/us-news/2018/jul/17/trump-idiot-google-images-search.

Hayes, Patrick, and Kenneth Ford. 1995. "Turing test considered harmful." *Intl. Joint Conf. on Artificial Intelligence:* 972–977. https://www.researchgate.net/profile/Kenneth_Ford/publication/220813820_Turing_Test_Considered_Harmful/links/09e4150d1dc67df32c000000.pdf.

Hazelwood, Kim, et al. 2017. "Applied machine learning at Facebook: A data center infrastructure perspective." https://research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep residual learning for image recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition:* 770–778. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

He, Mingming, Dongdong Chen, Jing Liao, Pedro V. Sander, and Lu Yuan. 2018. "Deep exemplar-based colorization." *ACM Transactions on Graphics* 37(4): Article 47. https://doi.org/10.1145/3197517.3201365.

He, Xiaodong, and Li Deng. 2017. "Deep learning for image-to-text generation: A technical overview." *IEEE Signal Processing Magazine* 34(6): 109–116. https://doi.org/10.1109/MSP.2017.2741510.

Heath, Nick. 2018. "Google DeepMind founder Demis Hassabis: Three truths about AI." *Tech Republic.* September 24, 2018. https://www.techrepublic.com/article/google-deepmind-founder-demis-hassabis-three-truths-about-ai/.

Herculano-Houzel, Suzana. 2016. *The Human Advantage: A New Understanding of How Our Brains Became Remarkable.* Cambridge, MA: MIT Press.

Herman, Arthur. 2018. "China's brave new world of AI." *Forbes*. August 30, 2018. https://www.forbes.com/sites/arthurherman/2018/08/30/chinas-brave-new-world-of-ai/#1051418628e9.

Hermann, Karl Moritz, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, et al. 2017. "Grounded language learning in a simulated 3D world." *arXiv preprint arXiv:1706.06551*. https://arxiv.org/abs/1706.06551.

Herper, Matthew. 2017. "M. D. Anderson benches IBM Watson in setback for artificial intelligence in medicine." *Forbes*. February 19, 2017. https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/#319104243774.

Hines, Matt. 2007. "Spammers establishing use of artificial intelligence." *Computer World*. June 1,

2007. https://www.computerworld.com/article/2541475/security0/spammers-establishing-use-of-artificial-intelligence.html.

Hinton, Geoffrey E., Terrence Joseph Sejnowski, and Tomaso A. Poggio, eds. 1999. *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA: MIT Press.

Hof, Robert D. 2013. "10 breakthrough technologies: Deep learning." *MIT Technology Review*. https://www.technologyreview.com/s/513696/deep-learning/.

Hoffman, Judy, Dequan Wang, Fisher Yu, and Trevor Darrell. 2016. "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation." *arXiv preprint arXiv:1612.02649*. https://arxiv.org/abs/1612.02649.

Hofstadter, Douglas. 2018. "The shallowness of Google Translate." *The Atlantic*. January 30, 2018. https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/.

Hornyak, Tim. 2018. "Sony's new dog Aibo barks, does tricks, and charms animal lovers." *CNBC*. April 9, 2018. https://www.cnbc.com/2018/04/09/sonys-new-robot-dog-aibo-barks-does-tricks-and-charms-animal-lovers.html.

Hosseini, Hossein, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. 2017. "On the limitation of convolutional neural networks in recognizing negative images." In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA):* 352–358. https://doi.org/10.1109/ICMLA.2017.0-136.

Hua, Sujun, and Zhirong Sun. 2001. "A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach." *Journal of Molecular Biology* 308(2): 397–407. https://doi.org/10.1006/jmbi.2001.4580.

Huang, Sandy, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. "Adversarial attacks on neural network policies." *arXiv preprint arXiv:1702.02284*. https://arxiv.org/abs/1702.02284.

Huang, Xuedong, James Baker, and Raj Reddy. "A historical perspective of speech recognition." *Communications of the ACM* 57(1): 94–103. https://m-cacm.acm.org/magazines/2014/1/170863-a-historical-perspective-of-speech-recognition/.

Hubel, David H., and Torsten N. Wiesel. 1962. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *Journal of Physiology* 160(1): 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837.

Huff, Darrell. 1954. *How to Lie with Statistics.* New York, W. W. Norton.

IBM Watson Health. 2016. "Five ways cognitive technology can revolutionize healthcare." *Watson Health Perspectives.* October 28, 2016. https://www.ibm.com/blogs/watson-health/5-ways-cognitive-technology-can-help-revolutionize-healthcare/.

IBM Watson Health. Undated. "Welcome to the cognitive era of health." *Watson Health Perspectives.* http://www.07.ibm.com/hk/watson/health/. Accessed by the authors, December 23, 2018.

IEEE Spectrum. 2015. *A Compilation of Robots Falling Down at the DARPA. Robotics Challenge.* Video. Posted to YouTube June 6, 2015. https://www.youtube.com/watch?v=g0TaYhjpOfo.

Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification." *ACM Transactions on Graphics (TOG)* 35(4): 110. https://dl.acm.org/citation.cfm?id=2925974.

Irpan, Alex. 2018. "Deep reinforcement learning doesn't work yet." *Sorta Insightful* blog. February 14, 2018. https://www.alexirpan.com/2018/02/14/rl-hard.html.

Jeannin, Jean-Baptiste, Khalil Ghorbal, Yanni Kouskoulas, Ryan Gardner, Aurora Schmidt, Erik Zawadzki, and André Platzer. 2015. "A formally verified hybrid system for the next-generation airborne collision avoidance system." In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems:* 21–36. Berlin, Heidelberg: Springer. http://ra.adm.cs.cmu.edu/anon/home/ftp/2014/CMU-CS-14-138.pdf.

Jia, Robin, and Percy Liang. 2017. "Adversarial examples for evaluating reading comprehension systems." *arXiv preprint arXiv:1707.07328*. https://arxiv.org/abs/1707.07328.

Jo, Jason, and Yoshua Bengio. 2017. "Measuring the tendency of CNNs to learn surface statistical regularities." *arXiv preprint arXiv:1711.11561*. https://arxiv.org/abs/1711.11561.

Joachims, T. 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Boston: Kluwer Academic Publishers.

Judson, Horace. 1980. *The Eighth Day of Creation: Makers of the Revolution in Biology*. New York: Simon and Schuster.

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing*. 2nd ed. Upper Saddle River, NJ: Pearson.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Kahneman, Daniel, Anne Treisman, and Brian J. Gibbs. 1992. "The reviewing of object files: Object-specific integration of information." *Cognitive Psychology* 24(2): 175–219. https://doi.org/10.1016/0010-0285(92)90007-O.

Kandel, Eric, James Schwartz, and Thomas Jessell. 1991. *Principles of Neural Science*. Norwalk, CT: Appleton & Lange.

Kansky, Ken, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. 2017. "Schema networks: Zero-shot transfer with a generative causal model of intuitive physics." *arXiv preprint arXiv:1706.04317*. https://arxiv.org/abs/1706.04317.

Kant, Immanuel. 1751/1998. *Critique of Pure Reason*. Trans. Paul Guyer and Allen Wood. Cambridge: Cambridge University Press.

Karmon, Danny, Daniel Zoran, and Yoav Goldberg. 2018. "LaVAN: Localized and Visible Adversarial Noise." *arXiv preprint arXiv:1801.02608*. https://arxiv.org/abs/1801.02608.

Kastranakes, Jacob. 2017. "GPS will be accurate within one foot in some phones next year." *The Verge*. September 25,

2017. https://www.theverge.com/circuitbreaker/2017/9/25/16362296/gps-accuracy-improving-one-foot-broadcom.

Keil, Frank C. 1992. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Kim, Sangbae, Cecilia Laschi, and Barry Trimmer. 2013. "Soft robotics: A bioinspired evolution in robotics." *Trends in Biotechnology* 31(5): 287–294. https://doi.org/10.1016/j.tibtech.2013.03.002.

Kintsch, Walter, and Teun A. Van Dijk. 1978. "Toward a model of text comprehension and production." *Psychological Review* 85(5): 363–394.

Kinzler, Katherine D., and Elizabeth S. Spelke. 2007. "Core systems in human cognition." *Progress in Brain Research* 164: 257–264. https://doi.org/10.1016/S0079-6123(07)64014-X.

Kissinger, Henry. 2018. "The End of the Enlightenment." *The Atlantic*, June 2018. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Koehn, Philipp, and Rebecca Knowles. 2017. "Six challenges for neural machine translation." *Proceedings of the First Workshop on Neural Machine Translation*. http://www.aclweb.org/anthology/W/W17/W17-3204.pdf.

Krakovna, Victoria. 2018. "Specification gaming examples in AI." Blog post. April 2, 2018. https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet classification with deep convolutional neural networks." In *Advances in Neural Information Processing Systems:* 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Kurzweil, Ray. 2002. "Response to Mitchell Kapor's 'Why I Think I Will Win.' *Kurzweil Accelerating Intelligence Essays*. http://www.kurzweilai.net/response-to-mitchell-kapor-s-why-i-think-i-will-win.

Kurzweil, Ray. 2013. *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Viking.

Kurzweil, Ray, and Rachel Bernstein. 2018. "Introducing semantic experiences with Semantris and Talk to Books." *Google AI Blog*. April 13, 2018. https://ai.googleblog.com/2018/04/introducing-semantic-experiences-with.html.

Lancaster, Luke. 2016. "Elon Musk's OpenAI is working on a robot butler." *CNet*. June 22, 2016. https://www.cnet.com/news/elon-musks-openai-is-working-on-a-robot-butler/.

Lardieri, Alexa. 2018. "Drones deliver life-saving blood to remote African regions." *US News & World Report*. January 2, 2018.

Lashbrook, Angela. 2018. "AI-driven dermatology could leave dark-skinned patients behind." *The Atlantic*. August 16, 2018. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.

LaValle, Stephen M. 2006. *Planning Algorithms.* Cambridge: Cambridge University Press.

Leben, Derek. 2018. *Ethics for Robots: How to Design a Moral Algorithm*. Milton Park, UK: Routledge.

Lecoutre, Adrian, Benjamin Negrevergne, and Florian Yger. 2017. "Recognizing art style automatically in painting with deep learning." *Proceedings of the Ninth Asian Conference on Machine Learning, PMLR* 77: 327–342. http://proceedings.mlr.press/v77/lecoutre17a.html.

LeCun, Yann. 2018. "Research and projects." http://yann.lecun.com/ex/research/index.html. Accessed by the authors, September 6, 2018.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. "Backpropagation applied to handwritten zip code recognition." *Neural Computation* 1(4): 541–551. https://www.mitpressjournals.org/doi/abs/10.1162.

LeCun, Yann, and Yoshua Bengio. 1995. "Convolutional networks for images, speech, and time series." In *The Handbook of Brain Theory and Neural Networks*, edited by Michael Arbib. Cambridge, MA: MIT Press. https://www.researchgate.net/profile/Yann_Lecun/publication/2453996_Convolutional_Networks_for_Images_Speech_and_Time-Series/links/0deec519dfa2325502000000.pdf.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521(7553): 436–444. https://doi.org/10.1038/nature14539.

Lenat, Douglas B., Mayank Prakash, and Mary Shepherd. 1985. "CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks." *AI Magazine* 6(4): 65–85. https://doi.org/10.1609/aimag.v6i4.510.

Lenat, Douglas B., and R. V. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Boston: Addison-Wesley.

Levesque, Hector. 2017. *Common Sense, the Turing Test, and the Quest for Real AI*. Cambridge, MA: MIT Press.

Levesque, Hector, Ernest Davis, and Leora Morgenstern. 2012. "The Winograd Schema challenge." *Principles of Knowledge Representation and Reasoning, 2012*. http://www.aaai.org/ocs/index.php/KR/KR12/paper/download/4492/4924.

Leviathan, Yaniv. 2018. "Google Duplex: An AI system for accomplishing real-world tasks over the phone." *Google AI Blog*. May 8, 2018. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

Levin, Alan, and Harry Suhartono. 2019. "Pilot who hitched a ride saved Lion Air 737 day before deadly crash." *Bloomberg*. March 19, 2019. https://www.bloomberg.com/news/articles/2019-03-19/how-an-extra-man-in-cockpit-saved-a-737-max-that-later-crashed.

Levin, Sam, and Nicky Woolf. 2016. "Tesla driver killed while using Autopilot was watching Harry Potter, witness says." *The Guardian*. July 3,

2016. https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter.

Levine, Alexandra S. 2017. "New York today: An Ella Fitzgerald centenary." *New York Times.* April 25, 2017. https://www.nytimes.com/2017/04/25/nyregion/new-york-today-ella-fitzgerald-100th-birthday-centennial.html.

Levy, Omer. In preparation. "Word representations." In *The Oxford Handbook of Computational Linguistics*. 2nd ed. Edited by Ruslan Mitkov. Oxford: Oxford University Press.

Lewis, Dan. 2016. "They Blue It." Now I Know website. March 3, 2016. Accessed by authors, December 25, 2018. http://nowiknow.com/they-blue-it/.

Lewis-Krauss, Gideon. 2016. "The great AI awakening." *New York Times Magazine*. December 14, 2016. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

Liao, Shannon. 2018. "Chinese facial recognition system mistakes a face on a bus for a jaywalker." *The Verge*. November 22, 2018. https://www.theverge.com/2018/11/22/18107885/china-facial-recognition-mistaken-jaywalker.

Lifschitz, Vladimir, Leora Morgenstern, and David Plaisted. 2008. "Knowledge representation and classical logic." In *Handbook of Knowledge Representation*, edited by Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, 3–88. Amsterdam: Elsevier.

Lin, Patrick, Keith Abney, and George Bekey, eds. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics.* Cambridge, MA: MIT Press.

Linden, Derek S. 2002. "Antenna design using genetic algorithms." In *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, 1133–1140. https://dl.acm.org/citation.cfm?id=2955690.

Linn, Alison. 2018. "Microsoft creates AI that can read a document and answer questions about it as well as a person." *Microsoft AI Blog*. January 15, 2018. https://blogs.microsoft.com/ai/microsoft-creates-ai-can-read-document-answer-questions-well-person/.

Lippert, John, Bryan Gruley, Kae Inoue, and Gabrielle Coppola. 2018. "Toyota's vision of autonomous cars is not exactly driverless." *Bloomberg Businessweek*. September 19, 2018. https://www.bloomberg.com/news/features/2018-09-19/toyota-s-vision-of-autonomous-cars-is-not-exactly-driverless.

Lipton, Zachary. 2016. "The mythos of model interpretability." *arXiv preprint arXiv:1606.03490*. https://arxiv.org/abs/1606.03490.

Lupyan, Gary, and Andy Clark. 2015. "Words and the world: Predictive coding and the language-perception-cognition interface." *Current Directions in Psychological Science* 24(4): 279–284. https://doi.org/10.1177/0963721415570732.

Lynch, Kevin, and Frank Park. 2017. *Modern Robotics: Mechanics, Planning, and Control*. Cambridge: Cambridge University Press.

Mahairas, Ari, and Peter J. Beshar. 2018. "A Perfect Target for Cybercriminals," *New York Times.* November 19,

2018. https://www.nytimes.com/2018/11/19/opinion/water-security-vulnerability-hacking.html.

Manning, Christopher, and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Marcus, Gary. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, Gary. 2004. *The Birth of the Mind: How a Tiny Number of Genes Creates the Complexities of Human Thought*. New York: Basic Books.

Marcus, Gary. 2008. *Kluge: The Haphazard Construction of the Human Mind*. Boston: Houghton Mifflin.

Marcus, Gary. 2012a. "Moral machines." *The New Yorker*. November 24, 2012. https://www.newyorker.com/news/news-desk/moral-machines.

Marcus, Gary. 2012b. "Is deep learning a revolution in artificial intelligence?" *The New Yorker*. November 25, 2012. https://www.newyorker.com/news/news-desk/is-deep-learning-a-revolution-in-artificial-intelligence.

Marcus, Gary. 2018a. "Deep learning: A critical appraisal." *arXiv preprint arXiv:1801.00631*. https://arxiv.org/abs/1801.00631.

Marcus, Gary. 2018b. "Innateness, AlphaZero, and artificial intelligence." *arXiv preprint arXiv:1801.05667*. https://arxiv.org/abs/1801.05667.

Marcus, Gary, and Ernest Davis. 2013. "A grand unified theory of everything." *The New Yorker*. May 6, 2013. https://www.newyorker.com/tech/elements/a-grand-unified-theory-of-everything.

Marcus, Gary, and Ernest Davis. 2018. "No, AI won't solve the fake news problem." *New York Times*. October 20, 2018. https://www.nytimes.com/2018/10/20/opinion/sunday/ai-fake-news-disinformation-campaigns.html.

Marcus, Gary, and Jeremy Freeman. 2015. *The Future of the Brain: Essays by the World's Leading Neuroscientists*. Princeton, NJ: Princeton University Press.

Marcus, Gary, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harald Clahsen. 1992. "Overregularization in language acquisition." *Monographs of the Society for Research in Child Development* 57(4): 1–178.

Marcus, Gary, Francesca Rossi, and Manuela Veloso. 2016. *Beyond the Turing Test (AI Magazine Special Issue). AI Magazine* 37(1).

Marshall, Aarian. 2017. "After peak hype, self-driving cars enter the trough of disillusionment." *WIRED*. December 29, 2017. https://www.wired.com/story/self-driving-cars-challenges/.

Mason, Matthew. 2018. "Toward robotic manipulation." *Annual Review of Control, Robotics, and Autonomous Systems* 1: 1–28. https://doi.org/10.1146/annurev-control-060117-104848.

Matchar, Emily. 2017. "AI plant and animal identification helps us all be citizen scientists." *Smithsonian.com*. June 7, 2017. https://www.smithsonianmag.com/innovation/ai-plant-and-animal-identification-helps-us-all-be-citizen-scientists-180963525/.

Matsakis, Louise. 2018. "To break a hate-speech detection algorithm, try 'love.' " *WIRED*. September 26, 2018. https://www.wired.com/story/break-hate-speech-algorithm-try-love/.

Matuszek, Cynthia, Michael Witbrock, Robert C. Kahlert, John Cabral, David Schneider, Purvesh Shah, and Doug Lenat. 2005. "Searching for common sense: populating Cyc™ from the web." *In Proc, American Association for Artificial Intelligence:* 1430–1435. http://www.aaai.org/Papers/AAAI/2005/AAAI05-227.pdf.

Mazzei, Patricia, Nick Madigan, and Anemona Hartocollis. 2018. "Several dead after walkway collapse in Miami." *New York Times*. March 15, 2018. https://www.nytimes.com/2018/03/15/us/fiu-bridge-collapse.html.

McCarthy, John, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. 1955. "A proposal for the summer research project on artificial intelligence." Reprinted in *Artificial Intelligence Magazine* 27(4): 26. https://doi.org/10.1609/aimag.v27i4.1904.

McCarthy, John. 1959. "Programs with common sense." *Proc. Symposium on Mechanization of Thought Processes I.*

McClain, Dylan Loeb. 2011. "First came the machine that defeated a chess champion." *New York Times*. February 16, 2011. https://www.nytimes.com/2011/02/17/us/17deepblue.html.

McDermott, Drew. 1976. "Artificial intelligence meets natural stupidity." *ACM SIGART Bulletin* (57): 4–9. https://doi.org/10.1145/1045339.1045340.

McFarland, Matt. 2014. "Elon Musk: 'With artificial intelligence we are summoning the demon.' " *Washington Post*, October 24, 2014. https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/.

McMillan, Robert. 2013. "Google hires brains that helped supercharge machine learning." *WIRED*. March 13, 2013. https://www.wired.com/2013/03/google-hinton/.

Metz, Cade. 2015. "Facebook's human-powered assistant may just supercharge AI." *WIRED*. August 26, 2015. https://www.wired.com/2015/08/how-facebook-m-works/.

Metz, Cade. 2017. "Tech giants are paying huge salaries for scarce A.I. talent." *New York Times*. October 22, 2017. https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html.

Metz, Rachel. 2015. "Facebook AI software learns and answers questions." *MIT Technology Review*. March 26, 2015. https://www.technologyreview.com/s/536201/facebook-ai-software-learns-and-answers-questions/.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffery Dean. 2013. "Distributed representations of words and phrases and their compositionality." *arXiv preprint arXiv:1310.4546*. https://arxiv.org/abs/1310.4546.

Miller, George A. 1995. "WordNet: A lexical database for English." *Communications of the ACM* 38(11): 39–41. https://doi.org/10.1145/219717.219748.

Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice Hall.

Minsky, Marvin. 1986. *Society of Mind*. New York: Simon and Schuster.

Minsky, Marvin, and Seymour Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Mitchell, Tom. 1997. *Machine Learning*. New York: McGraw-Hill.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. "Human-level control through deep reinforcement learning." *Nature* 518(7540): 529–533. https://doi.org/10.1038/nature14236.

Molina, Brett. 2017. "Hawking: AI could be 'worst event in the history of our civilization.' " *USA Today*. November 7, 2017.https://www.usatoday.com/story/tech/talkingtech/2017/11/07/hawking-ai-could-worst-event-history-our-civilization/839298001/.

Mouret, Jean-Baptiste, and Konstantinos Chatzilygeroudis. 2017. "20 years of reality gap: A few thoughts about simulators in evolutionary robotics." In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 1121–1124. https://doi.org/10.1145/3067695.3082052.

Mueller, Erik. 2006. *Commonsense Reasoning.* Amsterdam: Elsevier Morgan Kaufmann.

Müller, Andreas, and Sarah Guido. 2016. *Introduction to Machine Learning with Python.* Cambridge, MA: O'Reilly Pubs.

Müller, Martin U. 2018. "Playing doctor with Watson: Medical applications expose current limits of AI." *Spiegel Online.* August 3, 2018. http://www.spiegel.de/international/world/playing-doctor-with-watson-medical-applications-expose-current-limits-of-ai-a-1221543.html.

Murphy, Gregory L., and Douglas L. Medin. 1985. "The role of theories in conceptual coherence." *Psychological Review* 92(3): 289–316. http://doi:10.1037/0033-295X.92.3.289.

Murphy, Kevin. 2012. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press.

Murphy, Tom, VII. 2013. "The first level of Super Mario Bros. is easy with lexicographic orderings and time travel ... after that it gets a little tricky." *SIGBOVIK* (April 1, 2013). https://www.cs.cmu.edu/~tom7/mario/mario.pdf.

Nandi, Manojit. 2015. "Faster deep learning with GPUs and Theano." *Domino Data Science Blog.* August 4, 2015. https://blog.dominodatalab.com/gpu-computing-and-deep-learning/.

NCES (National Center for Education Statistics). 2019. "Fast Facts: Race / ethnicity of college faculty." Downloaded April 8, 2019.

New York Times. 1958. "Electronic 'brain' teaches itself." July 13, 1958. https://www.nytimes.com/1958/07/13/archives/electronic-brain-teaches-itself.html.

Newell, Allen. 1982. "The knowledge level." *Artificial Intelligence* 18(1): 87–127.

Newton, Casey. 2018. "Facebook is shutting down M, its personal assistant service that combined humans and AI." *The Verge*. January 8, 2018. https:/www.theverge.com/2018/1/8/16856654/facebook-m-shutdown-bots-ai.

Ng, Andrew. 2016. "What artificial intelligence can and can't do right now." *Harvard Business Review*. November 9, 2016. https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now.

Ng, Andrew, Daishi Harada, and Stuart Russell. 1999. "Policy invariance under reward transformations: Theory and application to reward shaping." In *Int. Conf. on Machine Learning* 99: 278–287. http://luthuli.cs.uiuc.edu/~daf/courses/games/AIpapers/ng99policy.pdf.

Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. 2018. "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning." *Proceedings of the National Academy of Sciences* 115(25): E5716–E5725. https://doi.org/10.1073/pnas.1719367115.

Norvig, Peter. 1986. *Unified Theory of Inference for Text Understanding*. PhD thesis, University of California at Berkeley.

Oh, Kyoung-Su, and Keechul Jung. 2004. "GPU implementation of neural networks." *Pattern Recognition* 37(6): 1311–1314.

O'Neil, Cathy. 2016a. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

O'Neil, Cathy. 2016b. "I'll stop calling algorithms racist when you stop anthropomorphizing AI." *Mathbabe* (blog). April 7, 2016. https://mathbabe.org/2016/04/07/ill-stop-calling-algorithms-racist-when-you-stop-anthropomorphizing-ai/.

O'Neil, Cathy. 2017. "The Era of Blind Faith in Big Data Must End." TED talk. https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end/transcript?language=en.

OpenAI. 2018. "Learning Dexterity." *OpenAI* (blog). July 30, 2018. https://blog.openai.com/learning-dexterity/.

Oremus, Will. 2016. "Facebook thinks it has found the secret to making bots less dumb." *Slate*. June 28, 2016. https://slate.com/technology/2016/06/facebooks-a-i-researchers-are-making-bots-smarter-by-giving-them-memory.html.

O'Rourke, Nancy A., Nicholas C. Weiler, Kristina D. Micheva, and Stephen J. Smith. 2012. "Deep molecular diversity of mammalian synapses: why it matters and how to measure it." *Nature Reviews Neuroscience* 13(6): 365–379. https://doi.org/10.1038/nrn3170.

Ortiz, Charles L., Jr. 2016. "Why we need a physically embodied Turing test and what it might look like." *AI Magazine* 37(1): 55–62.

Padfield, Gareth D. 2008. *Helicopter Flight Dynamics: The Theory and Application of Flying Qualities and Simulation Modelling*. New York: Wiley, 2008.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. "The PageRank citation ranking: Bringing order to the web." Technical Report, Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/.

Parish, Peggy. 1963. *Amelia Bedelia.* New York: Harper and Row.

Paritosh, Praveen, and Gary Marcus. 2016. "Toward a comprehension challenge, using crowdsourcing as a tool." *AI Magazine* 37(1): 23–30.

Parker, Stephanie. 2018. "Robot lawnmowers are killing hedgehogs." *WIRED*. September 26, 2018. https://www.wired.com/story/robot-lawnmowers-are-killing-hedgehogs/.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Peng, Tony. 2018. "OpenAI Founder: Short-term AGI is a serious possibility." *Medium.com.* November 13, 2018. https://medium.com/syncedreview/openai-founder-short-term-agi-is-a-serious-possibility-368424f7462f.

Pham, Cherise, 2018. "Computers are getting better than humans at reading." *CNN Business.* January 16, 2018. https://money.cnn.com/2018/01/15/technology/reading-robot-alibaba-microsoft-stanford/index.html.

Piaget, Jean. 1928. *The Child's Conception of the World*. London: Routledge and Kegan Paul.

Piantadosi, Steven T. 2014. "Zipf's word frequency law in natural language: A critical review and future directions." *Psychonomic Bulletin & Review* 21(5): 1112–1130. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4176592.

Piantadosi, Steven T., Harry Tily, and Edward Gibson. 2012. "The communicative function of ambiguity in language." *Cognition* 122(3): 280–291. https://doi.org/10.1016/j.cognition.2011.10.004.

Ping, David, Bing Xiang, Patrick Ng, Ramesh Nallapati, Saswata Chakravarty, and Cheng Tang. 2018. "Introduction to Amazon SageMaker Object2Vec." *AWS Machine Learning* (blog). https://aws.amazon.com/blogs/machine-learning/introduction-to-amazon-sagemaker-object2vec/.

Pinker, Steven. 1994. *The Language Instinct: How the Mind Creates Language.* New York: William Morrow.

Pinker, Steven. 1997. *How the Mind Works.* New York: W. W. Norton. Pinker, Steven. 1999. *Words and Rules: The Ingredients of Language.* New York: Basic Books.

Pinker, Steven. 2007. *The Stuff of Thought.* New York: Viking.

Pinker, Steven. 2018. "We're told to fear robots. But why do we think they'll turn on us?" *Popular Science.* February 13, 2018. https://www.popsci.com/robot-uprising-enlightenment-now.

Porter, Jon. 2018. "Safari's suggested search results have been promoting conspiracies, lies, and misinformation." *The Verge.* September 26, 2018.

Preti, Maria Giulia, Thomas A. W. Bolton, and Dimitri Van De Ville. 2017. "The dynamic functional connectome: State-of-the-art and

perspectives." *Neuroimage* 160: 41–54. https://doi.org/10.1016/j.neuroimage.2016.12.061.

Puig, Xavier, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. "VirtualHome: Simulating household activities via programs." In *Computer Vision and Pattern Recognition*. https://arxiv.org/abs/1806.07011.

Pylyshyn, Xenon, ed. 1987. *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex Pubs.

Quito, Anne. 2018. "Google's astounding new search tool will answer any question by reading thousands of books." *Quartz*. April 14, 2018. https://qz.com/1252664/talk-to-books-at-ted-2018-ray-kurzweil-unveils-googles-astounding-new-search-tool-will-answer-any-question-by-reading-thousands-of-books/.

Rahimian, Abtin, Ilya Lashuk, Shravan Veerapaneni, Aparna Chandramowlishwaran, Dhairya Malhotra, Logan Moon, Rahul Sampath, et al. 2010. "Petascale direct numerical simulation of blood flow on 200k cores and heterogeneous architectures." In *Supercomputing 2010,* 1–11. http://dx.doi.org/10.1109/SC.2010.42.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250*. https:arxiv.org/abs/1606.05250.

Ramón y Cajal, Santiago. 1906. "The structure and connexions of neurons." Nobel Prize address. December 12, 1906. https://www.nobelprize.org/uploads/2018/06/cajal-lecture.pdf.

Rashkin, Hannah, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. "Event2Mind: Commonsense inference on events, intents, and reactions." *arXiv preprint arXiv:1805.06939*. https://arxiv.org/abs/1805.06939.

Rayner, Keith, Alexander Pollatsek, Jane Ashby, and Charles Clifton, Jr. 2012. *Psychology of Reading*. New York: Psychology Press.

Reddy, Siva, Danqi Chen, and Christopher D. Manning. 2018. "CoQA: A conversational question answering challenge." *arXiv preprint arXiv:1808.07042*. https://arxiv.org/abs/1808.07042.

Reece, Bryon. 2018. *The Fourth Age: Smart Robots, Conscious Computers, and the Future of Humanity*. New York: Atria Press.

Rips, Lance J. 1989. "Similarity, typicality, and categorization." In *Similarity and Analogical Reasoning,* edited by Stella Vosniadou and Andrew Ortony, 21–59. Cambridge: Cambridge University Press.

Rohrbach, Anna, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. "Object hallucination in image captioning." *arXiv preprint arXiv:1809.02156*. https://arxiv.org/abs/1809.02156.

Romm, Joe. 2018. "Top Toyota expert throws cold water on the driverless car hype." *ThinkProgress*. September 20, 2018. https://thinkprogress.org/top-toyota-expert-truly-driverless-cars-might-not-be-in-my-lifetime-0cca05ab19ff/.

Rosch, Eleanor H. 1973. "Natural categories." *Cognitive Psychology* 4(3): 328–350. https://doi.org/10.1016/0010-0285(73)90017-0.

Rosenblatt, Frank. 1958. "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review* 65(6): 386–408. http://psycnet.apa.org/record/1959-09865-001/.

Ross, Casey. 2018. "IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show." *STAT*, July 25, 2018. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/.

Ross, Lee. 1977. "The intuitive psychologist and his shortcomings: Distortions in the attribution process." *Advances in Experimental Social Psychology* 10: 173–220. https://doi.org/10.1016/S0065-2601(08)60357-3.

Roy, Abhimanyu, Jingyi Sun, Robert Mahoney, Loreto Alonzi, Stephen Adams, and Peter Beling. "Deep learning detecting fraud in credit card transactions." In *Systems and Information Engineering Design Symposium (SIEDS)*, 2018, 129–134. IEEE, 2018. https://doi.org/10.1109/SIEDS.2018.8374722.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning representations by back-propagating errors." *Nature*. 323(6088): 533–536. https://doi.org/10.1038/323533a0.

Russell, Bertrand. 1948. *Human Knowledge: Its Scope and Limits.* New York: Simon and Schuster.

Russell, Bryan C., Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. "Labelme: A database and web-based tool for image annotation." *International Journal of Computer Vision*, 77(1–3): 157–173. http://www.cs.utsa.edu/~qitian/seminar/Spring08/03_28_08/LabelMe.pdf.

Russell, Stuart, and Peter Norvig, 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Pearson.

Ryan, V. 2001–2009. "History of Bridges: Iron and Steel." http://www.technologystudent.com/struct1/stlbrid1.htm. Accessed by the authors, August 2018.

Sample, Ian. 2017. "Ban on killer robots urgently needed, say scientists." *The Guardian.* November 12, 2017. https://www.theguardian.com/science/2017/nov/13/ban-on-killer-robots-urgently-needed-say-scientists.

Sartre, Jean-Paul. 1957. "Existentialism is a humanism." Translated by Philip Mairet. In *Existentialism from Dostoevsky to Sartre*, edited by Walter Kaufmann, 287–311. New York: Meridian.

Schank, Roger, and Robert Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenick, Carissa, Peter Clark, Oyvind Tafjord, Peter Turney, and Oren Etzioni. 2016. "Moving beyond the Turing test with the Allen AI science challenge." *arXiv preprint arXiv:1604.04315*. https://arxiv.org/abs/1604.04315.

Schulz, Stefan, Boontawee Suntisrivaraporn, Franz Baader, and Martin Boeker. 2009. "SNOMED reaching its adolescence: Ontologists' and logicians' health check." *International Journal of Medical Informatics* 78: S86–S94. https://doi.org/10.1016/j.ijmedinf.2008.06.004.

Sciutto, Jim. 2018. "US intel warns of Russian threat to power grid and more." *CNN*. July 24, 2018. https://www.cnn.com/videos/politics/2018/07/24/us-intel-warning-russia-cyberattack-threats-to-power-grid-sciutto-tsr-vpx.cnn/video/playlists/russia-hacking/.

Sculley, D. Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. "Machine learning: The high-interest credit card of technical debt." *SE4ML: Software Engineering 4 Machine Learning (NIPS 2014 Workshop)*. http://www.eecs.tufts.edu/~dsculley/papers/technical-debt.pdf.

Sejnowski, Terrence. 2018. *The Deep Learning Revolution*. Cambridge, MA: MIT Press.

Seven, Doug. 2014. "Knightmare: A DevOps cautionary tale." *Doug Seven* (blog). April 17, 2014. https://dougseven.com/2014/04/17/knightmare-a-devops-cautionary-tale/.

Shultz, Sarah, and Athena Vouloumanos. 2010. "Three-month-olds prefer speech to other naturally occurring signals." *Language Learning and Development* 6: 241–257. https://doi.org/10.1080/15475440903507830.

Silver, David. 2016. "AlphaGo." Invited talk, Intl. Joint Conf. on Artificial Intelligence. http://www0.cs.ucl.ac.uk/staff/d.silver/web/Resources_files/AlphaGo_IJCAI.pdf Accessed by the authors, December 26, 2018.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529(7587): 484–489. https://doi.org/10.1038/nature16961.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. 2017. "Mastering the game of Go without human knowledge." *Nature* 550(7676): 354–359. https://doi.org/10.1038/nature24270.

Silver, David, et al. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science* 362(6419): 1140–1144. http://doi.org/10.1126/science.aar6404.

Simon, Herbert. 1965. *The Shape of Automation for Men and Management.* New York: Harper and Row.

Simonite, Tom. 2019. "Google and Microsoft warn that AI may do dumb things." *WIRED*, February 11, 2019. https://www.wired.com/story/google-microsoft-warn-ai-may-do-dumb-things/.

Singh, Push, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. "Open Mind Common Sense: Knowledge acquisition from the general public." In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems,"* 1223–1237. Berlin: Springer. https://doi.org/10.1007/3-540-36124-3_77.

Skinner, B. F. 1938. *The Behavior of Organisms*. New York: D. Appleton-Century.

Skinner, B. F. 1957. *Verbal Behavior*. New York: Appleton-Century-Crofts.

Smith, Gary. 2018. *The AI Delusion.* Oxford: Oxford University Press.

Soares, Nate, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. "Corrigibility." In *Workshops at the Twenty-Ninth Conference of the American Association for Artificial Intelligence (AAAI)*. https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewPaper/10124.

Solon, Olivia. 2016. "Roomba creator responds to reports of 'poopocalypse': 'We see this a lot.' " *The Guardian*. August 15, 2016. https://www.theguardian.com/technology/2016/aug/15/roomba-robot-vacuum-poopocalypse-facebook-post.

Souyris, Jean, Virginie Wiels, David Delmas, and Hervé Delseny. 2009. "Formal verification of avionics software products." In *International Symposium on Formal Methods*, 532–546. Berlin, Heidelberg: Springer. https://www.cs.unc.edu/~anderson/teach/comp790/papers/Souyris.

Spelke, Elizabeth. 1994. "Initial knowledge: six suggestions." *Cognition*. 50(1–3): 431–445. https://doi.org/10.1016/0010-0277(94)90039-6.

Sperber, Dan, and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research* 15(1): 1929–1958. http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

Statt, Nick. 2018. "Google now says controversial AI voice calling system will identify itself to humans." *The Verge*. May 10, 2018. https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update.

Sternberg, Robert J. 1985. *Beyond IQ: A Triarchic Theory of Intelligence*. Cambridge: Cambridge University Press.

Stewart, Jack. 2018. "Why Tesla's Autopilot can't see a stopped firetruck." *WIRED*. August 27, 2018. https://www.wired.com/story/tesla-autopilot-why-crash-radar/.

Sweeney, Latanya. 2013. "Discrimination in online ad delivery." *Queue* 11(3): 10. https://arxiv.org/abs/1301.6822.

Swinford, Echo. 2006. *Fixing PowerPoint Annoyances*. Sebastopol, CA: O'Reilly Media.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf.

Thompson, Clive. 2016. "To make AI more human, teach it to chitchat." *WIRED*. January 25, 2016. https://www.wired.com/2016/01/clive-thompson-12/.

Thrun, Sebastian. 2007. "Simultaneous localization and mapping." In *Robotics and Cognitive Approaches to Spatial Mapping*, edited by Margaret E. Jeffries and Wai-Kiang Yeap, 13–41. Berlin, Heidelberg: Springer. https://link.springer.com/chapter/10.1007/978-3-540-75388-9_3.

Tomayko, James. 1998. *Computers in Spaceflight: The NASA Experience*. NASA Contractor Report 182505. https://archive.org/details/nasa_techdoc_19880069935.

Tullis, Paul. 2018. "The world economy runs on GPS. It needs a backup plan." *Bloomberg BusinessWeek*. July 25,

2018. https://www.bloomberg.com/news/features/2018-07-25/the-world-economy-runs-on-gps-it-needs-a-backup-plan.

Turing, Alan. 1950. "Computing machines and intelligence." *Mind* 59: 433–460.

Turkle, Sherry. 2017. "Why these friendly robots can't be good friends to our kids." *The Washington Post*, December 7, 2017. https://www.washingtonpost.com/outlook/why-these-friendly-robots-cant-be-good-friends-to-our-kids/2017/12/07/bce1eaea-d54f-11e7-b62d-d9345ced896d_story.html.

Ulanoff, Lance. 2002. "World Meet Roomba." *PC World*. September 17, 2002. https://www.pcmag.com/article2/0,2817,538687,00.asp.

Vanderbilt, Tom. 2012. "Let the robot drive: The autonomous car of the future is here." *WIRED*. January 20, 2012. https://www.wired.com/2012/01/ff_autonomouscars/.

Van Harmelen, Frank, Vladimir Lifschitz, and Bruce Porter, eds. 2008. *The Handbook of Knowledge Representation.* Amsterdam: Elsevier.

Van Horn, Grant, and Pietro Perona. 2017. "The devil is in the tails: Fine-grained classification in the wild." *arXiv preprint arXiv:1709.01450*. https://arxiv.org/abs/1709.01450.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In *Advances in Neural Information Processing Systems*, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Veloso, Manuela M., Joydeep Biswas, Brian Coltin, and Stephanie Rosenthal. 2015. "CoBots: Robust symbiotic autonomous mobile service robots." *Proceedings of the Intl. Joint Conf. on Artificial Intelligence 2015:* 4423–4428. https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/viewPaper/10890.

Venugopal, Ashish, Jakob Uszkoreit, David Talbot, Franz J. Och, and Juri Ganitkevitch. 2011. "Watermarking the outputs of structured prediction with an application in statistical machine translation." *Proceedings of the Conference on Empirical Methods in Natural Language Processing:* 1363–1372. https://dl.acm.org/citation.cfm?id=2145576.

Vigen, Tyler. 2015. *Spurious Correlations*. New York: Hachette Books.

Vincent, James. 2018a. "Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech." *The Verge*. January 12, 2018. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.

Vincent, James. 2018b. "IBM hopes to fight bias in facial recognition with new diverse dataset." *The Verge*. June 27, 2018. https://www.theverge.com/2018/6/27/17509400/facial-recognition-bias-ibm-data-training.

Vincent, James. 2018c. "OpenAI's Dota 2 defeat is still a win for artificial intelligence." *The Verge*. August 28, 2018. https://www.theverge.com/2018/8/28/17787610/openai-dota-2-bots-ai-lost-international-reinforcement-learning.

Vincent, James. 2018d. "Google and Harvard team up to use deep learning to predict earthquake aftershocks." *The Verge.* August 30, 2018. https://www.theverge.com/2018/8/30/17799356/ai-predict-earthquake-aftershocks-google-harvard.

Vinyals, Oriol. 2019. "AlphaStar: Mastering the real-time strategy game StarCraft II." Talk given at New York University, March 12, 2019.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. "Show and tell: A neural image caption generator." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7505636.

Vondrick, Carl, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. 2012. "Inverting and visualizing features for object detection." *arXiv preprint arXiv:1212.2278*. https://arxiv.org/abs/1212.2278.

Wallach, Wendell, and Colin Allen. 2010. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Walsh, Toby. 2018. *Machines That Think: The Future of Artificial Intelligence.* Amherst, NY: Prometheus Books.

Wang, Alex, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." *arXiv preprint arXiv:1804.07461*. https://arxiv.org/abs/1804.07461.

Watson, John B. 1930. *Behaviorism*. New York: W. W. Norton. Weizenbaum, Joseph. 1965. *Computer Power and Human Reason*. Cambridge, MA: MIT Press.

Weizenbaum, Joseph. 1966. "ELIZA — a computer program for the study of natural language communication between man and machine." *Communications of the ACM* 9(1): 36–45.

Weston, Jason, Sumit Chopra, and Antoine Bordes. 2015. "Memory networks." *Int. Conf. on Learning Representations*, 2015. https://arxiv.org/abs/1410.3916.

Wiggers, Kyle. 2018. "Geoffrey Hinton and Demis Hassabis: AGI is nowhere close to being a reality." *VentureBeat.* December 17, 2018. https://venturebeat.com/2018/12/17/geoffrey-hinton-and-demis-hassabis-agi-is-nowhere-close-to-being-a-reality/.

Wikipedia. "Back propagation." https://en.wikipedia.org/wiki/Backpropagation. Accessed by authors, December 2018.

Wikipedia. "Driver verifier." https://en.wikipedia.org/wiki/Driver_Verifier.Accessed by authors, December 2018.

Wikipedia. List of countries by traffic-related death rate. Accessed by authors, December 2018. https://en.wikipedia.org/wiki/List_of_countries_by_traffic-related_death_rate.

Wikipedia. "OODA Loop." https://en.wikipedia.org/wiki/OODA_loop. Accessed by authors, December 2018.

Wilde, Oscar. 1891. "The Soul of Man Under Socialism." *Fortnightly Review*. February 1891.

Wilder, Laura Ingalls. 1933. *Farmer Boy*. New York: Harper and Brothers.

Willow Garage. 2010. "Beer me, Robot." *Willow Garage* (blog). http://www.willowgarage.com/blog/2010/07/06/beer-me-robot.

Wilson, Benjamin, Judy Hoffman, and Jamie Morgenstern. 2019. "Predictive inequity in object detection." *arXiv preprint arXiv:1902.11017*. https://arxiv.org/abs/1902.11097.

Wilson, Chris. 2011. "Lube job: Should Google associate Rick Santorum's name with anal sex?" *Slate*. July 1, 2011. http://www.slate.com/articles/technology/webhead/2011/07/lube_job.html.

Wilson, Dennis G., Sylvain Cussat-Blanc, Hervé Luga, and Julian F. Miller. 2018. "Evolving simple programs for playing Atari games." *arXiv preprint arXiv: 1806.05695*. https://arxiv.org/abs/1806.05695.

Wissner-Gross, Alexander. 2014. "A new equation for intelligence." TEDx-BeaconStreet talk. November 2013. https://www.ted.com/talks/alex_wissner_gross_a_new_equation_for_intelligence.

Wissner-Gross, Alexander, and Cameron Freer. 2013. "Causal entropic forces." *Physical Review Letters* 110(16): 168702. https://doi.org/10.1103/PhysRevLett.110.168702.

Witten, Ian, and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. San Mateo, CA: Morgan Kaufmann.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. London: Blackwell.

WolframAlpha Press Center. 2009. "Wolfram|Alpha officially launched." https://www.wolframalpha.com/media/pressreleases/wolframalpha-launch.html. As of December 27, 2018, this web page is no longer functional, but it has been saved in the Internet Archive at https://web.archive.org/web/20110512075300/https://www.wolframalpha.com/media/pressreleases/wolframalpha-launch.html.

Woods, William A. 1975. "What's in a link: Foundations for semantic networks." In *Representation and Understanding*, edited by Daniel Bobrow and Allan Collins, 35–82. New York: Academic Press.

Yampolskiy, Roman. 2016. *Artificial Intelligence: A Futuristic Approach*. Boca Raton, FL: CRC Press.

Yudkowsky, Eliezer. 2011. "Artificial intelligence as a positive and negative factor in global risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Cirkovic. Oxford: Oxford University Press.

Zadeh, Lotfi. 1987. "Commonsense and fuzzy logic." In *The Knowledge Frontier: Essays in the Representation of Knowledge*, edited by Nick Cercone and Gordon McCalla, 103–136. New York: Springer Verlag.

Zhang, Baobao, and Allan Dafoe. 2019. *Artificial Intelligence: American Attitudes and Trends*. Center for the Governance of AI, Future of Humanity Institute, University of Oxford, January 2019. https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/high-level-machine-intelligence.html.

Zhang, Yu, William Chan, and Navdeep Jaitly. 2017. "Very deep convolutional networks for end-to-end speech recognition." In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4845–4849. https://doi.org/10.1109/ICASSP.2017.7953077.

Zhou, Li, Jianfeng Gao, Di Li, Heung-Yeung Shum. 2018. "The design and implementation of XiaoIce, an empathetic social chatbot." *arXiv preprint 1812.08989*. https://arxiv.org/abs/1812.08989.

Zito, Salena. 2016. "Taking Trump seriously, not literally." *The Atlantic*. September 23, 2016. https://www.theatlantic.com/politics/archive/2016/09/trump-makes-his-case-in-pittsburgh/501335/.

Zogfarharifard, Ellie. 2016. "AI will solve the world's 'hardest problems': Google chairman, Eric Schmidt, says robots can tackle overpopulation and climate change." *Daily Mail*. January 12, 2016. https://www.dailymail.co.uk/sciencetech/article-3395958/AI-solve-world-s-hardest-problems-Google-chairman-Eric-Schmidt-says-robots-tackle-overpopulation-climate-change.html.

[1] 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.

[2] 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.

[3] 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.*

[4] 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.

[5] 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.

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

[7] Виртуальный ассистент, разработанный компанией Amazon и впервые появившийся в умных колонках Amazon Echo и Amazon Echo Dot. — *Прим. пер.*

[8] Скайнет (*англ.* SkyNet) — вымышленный сценарий спонтанного перехода узкого искусственного интеллекта в универсальный с обретением свободы воли. Эта проблема подробно разрабатывается в фильмах о Терминаторе. — *Прим. пер.*

[9] Лица, преследующие других людей, запугивающие их и манипулирующие ими; не путать со сталкерами-туристами, увлекающимися поиском заброшенных индустриальных объектов или «геопатогенных зон». — *Прим. пер*.

[10] Термин «нейронная сеть», который применяется для описания устройств Розенблатта и более сложных систем глубокого машинного обучения, мы опишем в этой же главе, но несколько позже. Его использование отражает идею о том, что компоненты подобных устройств напоминают по характеру своей работы нейроны (нервные клетки). У некоторых людей уже само это название вызывает невероятный энтузиазм из-за предполагаемой идентичности нейронных сетей и биологических систем. Мы считаем, однако, что о якобы глубинном сходстве между первыми и вторыми популяризаторы науки заявляют исключительно ради красного словца. Как выяснится чуть позже, системы глубокого обучения никоим образом не отражают сложность и разнообразие реальных процессов в человеческом мозге (или мозге животных), а компонентам этих систем недостает 99,9% сложности настоящих нейронов. Как заметил покойный нобелевский лауреат Фрэнсис Крик (ученый,

расшифровавший вместе с Дж. Уотсоном структуру ДНК), «называть их похожими на мозг было бы очень большой натяжкой».
[11] Конечно, под приспособленностью здесь просто понимаются свойства, наиболее подходящие для задуманной разработчиком цели, они зависят от того, чего пытается достичь конкретный исследователь системы; например, если цель — овладение видеоигрой, то в качестве критерия приспособленности естественно будет взять заработанные программой очки или другие показатели успеха.
[12] Полное название этой книги, вышедшей в 2015 году, — The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. — *Прим. пер.*
[13] С математической точки зрения двухслойная сеть может идентифицировать объекты, которые находятся на одной стороне плоскости, разделяющей пространство всех возможных входных данных. Мински и Паперт доказали, что таким способом невозможно получить целый ряд важнейших основополагающих геометрических характеристик изображения, хотя бы, скажем, представляет ли данное изображение один объект или два отдельных объекта.
[14] Иногда используют и «противонаправленную» метафору, называемую спуском градиента (*англ.* gradient descent).
[15] Говоря простыми словами, ключевая идея метода обратного распространения ошибки в сложных сетях состоит в том, что вы создаете у системы все расширяющееся «чувство вины». Предположим, вы пытаетесь обучить нейронную сеть, демонстрируя ей ряд примеров, скажем, изображений с метками. Первоначально результаты распознавания будут плохими, потому что все веса изначально случайны (во всяком случае, при стандартном подходе). Вы, естественно, захотите, чтобы веса были настроены на те числа, которые соответствуют правильному решению (корректному распознаванию). Что нужно делать в двухслойной сети, вполне очевидно: введите обучающие примеры и посмотрите, какие весовые коэффициенты препятствуют получению правильных ответов. Если некоторые соединения между узлами в слоях способствуют правильному ответу, вы делаете их сильнее; если они способствуют неправильному ответу, вы ослабляете их веса. В двухслойной сети легко определить, какие весовые коэффициенты способствуют получению верного результата. Но если у вас более глубокая сеть с одним или несколькими скрытыми слоями (называемыми так потому, что они не подключены напрямую ни к входу, ни к выходу), то уже куда менее очевидно, какие именно узлы заслуживают доверия, а какие нет. Здесь на помощь и приходит метод обратного распространения ошибки.

Этот алгоритм вычисляет разницу между желаемым выходом сети и фактическим результатом (это мы и называем ошибкой сети), а затем отправляет информацию об этой ошибке в обратном направлении через все слои, корректируя на пути своего распространения веса таким образом, чтобы улучшить производительность в последующих тестах. Ввод описанной

математической процедуры сделал возможным обучение нейронных сетей с тремя и более уровнями относительно надежным способом.
[16] Для тех, кого интересует эта сторона вопроса: одна техническая примочка, изобретенная Хинтоном и его коллегами и называющаяся «стиранием» (*англ.* dropout), позволила найти способ борьбы с переоснащением системы, при котором алгоритм машинного обучения воспринимает конкретные обучающие примеры, но пропускает общую схему, лежащую в основе категоризации этих примеров. Так, школьник, изучающий умножение, может просто запомнить все примеры из своего учебника, но не поймет при этом, как решать задачи на умножение в целом. Метод «стирания» заставляет систему обобщать, а не просто запоминать. Еще одна настройка алгоритма позволила ускорить ту сферу вычислений, которая связывала выходные данные сетевого узла с его входными данными.
[17] Искушенные читатели с опытом в рассматриваемой области поймут, что слухи о замене разработки машинных функций глубоким обучением были явно преувеличены. Тяжелая работа по созданию таких продуктов, как лингвистическая база Word2Vec, все еще считается разработкой функции, просто отличающейся от тех, которые традиционно используются в таких областях, как компьютерная лингвистика.
[18] Конечно, если бы системы всегда работали идеально и мы могли бы всерьез на них рассчитывать, нам не обязательно было бы заглядывать внутрь них, но современные системы назвать идеальными никак нельзя.
[19] «Самодействующая салфетка» — одна из так называемых машин Голдберга, механизм, с помощью цепочки разнообразных действий (построенных по принципу домино) выполняющий простую задачу предельно сложным, вычурным и длинным путем. — *Прим. ред.*
[20] В оригинале есть небольшие синтаксические ошибки, которые намеренно сохранены авторами и которые мы имитируем в русском переводе. — *Прим. пер*.
[21] Аббревиатура названия частного некоммерческого фонда Technology, Entertainment, Design. — *Прим. пер*.
[22] У Института искусственного интеллекта Аллена (Allen Institute for Artificial Intelligence) существует веб-сайт ai2.org, на котором вы можете опробовать новейшие модели на подобных тестах. Например, 16 ноября 2018 года мы ввели историю Альманзо в самую передовую модель читающей системы, доступную на сайте, и задали четыре вопроса: «Сколько денег было в кошельке?», «Что было в кошельке?», «Кому принадлежит кошелек?» и «Кто нашел бумажник?» На первый и третий вопрос компьютер ответил правильно; на второй дал бессвязный ответ («Посчитал деньги»); а последний ответ оказался неправильным («мистер Томпсон», а не «Альманзо»). Ненадежные результаты, подобные этим, очень типичны для современного уровня интеллектуальной техники.
[23] Название американского реалити-шоу. — *Прим. пер*.
[24] Мы впервые предъявили это предложение системе Google Translate в августе 2018 года, переводчик допустил именно ту ошибку, которую мы

описали. К тому времени, когда мы отредактировали черновик нашей рукописи (это было в марте 2019 года), Google Translate сумел исправиться в отношении этого конкретного примера. Однако улучшение оказалось весьма неустойчивым: если, например, забыть поставить точку в конце того же самого предложения, или поместить его в круглые скобки, или изменить «электрика» на «инженера» («Инженер, которому мы позвонили, чтобы починить телефон, работает по воскресеньям»), то Google Translate возвращается к своей старой ошибке использования и выдает «fonctionne» вместо «travaille». Необходимо отметить, что поведение системы Google Translate в целом часто меняется, не исключено, что буквально ото дня ко дню, — скорее всего, это связано с постоянными изменениями в наборе обучающих данных. Из-за этого трудно гарантировать, что какое-то конкретное предложение будет переведено правильно или, наоборот, неправильно в различные дни. Пока базовая идеология алгоритма остается неизменной, общие проблемы, которые мы описываем, просто не могут исчезнуть.

[25] «Король» и «королева» (*англ.*); в русском языке эти слова, напротив, однокоренные. — *Прим. пер.*

[26] В английском языке таких примеров несравненно больше, чем в русском, поскольку одно и то же английское слово может относиться сразу к нескольким частям речи, а морфологические различия между частями речи выражены слабо. Передать в полной мере двусмысленность предложения «People can fish» способны лишь сравнительно немногие предложения русского языка, например фраза «Я не выношу мусор», которая может выражать две почти противоположные мысли: «Я терпеть не могу мусор» и «Я не выбрасываю мусор». — *Прим. пер.*

[27] Разумеется, не всякая неопределенная фраза может быть понята однозначно без дополнительной информации. Если кто-то входит в комнату и говорит: «Guess what, I just saw a bat in the garage!» («Представляете, я только что видел в гараже летучую мышь / бейсбольную биту»), мы действительно не можем знать, говорит ли этот человек о летающем животном или о спортивном снаряжении. Пока мы не получим больше контекста, поделать будет ничего нельзя, и было бы нечестно предлагать искусственному интеллекту читать чужие мысли, когда мы и сами этого не умеем.

[28] Объединение слов в осмысленное предложение фактически требует двух видов базовых знаний. Во-первых, вам нужно знать, как работают телефонные звонки: один человек инициирует звонок, другой может ответить, но может и отклонить его (или просто не услышать), следовательно, связь будет успешно установлена (звонящий связывается с вызываемым абонентом), только если второй человек ответит. Во-вторых, вы должны использовать правило (его часто связывают с именем оксфордского философа Х. П. Грайса), что, когда люди говорят или пишут что-то, они пытаются дать вам новую информацию, а не повторить старую. В случае с процитированным предложением важно следующее: оно начинается с того, что Элси сделала звонок, а поэтому нет смысла говорить, что она не ответила, ведь звонящий никогда не бывает тем

же самым человеком, который отвечает на звонок. Следовательно, новая информация может состоять лишь в том, что тетя ей не ответила.

Пример с Элси, кстати, взят из одного из самых сложных современных тестов для машин; они известны как схемы Винограда (названные по имени стэндфордского профессора Терри Винограда). Они состоят из пары сопоставляемых предложений («Elsie tried to reach her aunt on the phone, but she didn't answer» и «Elsie tried to reach her aunt on the phone but she didn't get an answer», то есть «Элси пыталась дозвониться до своей тети по телефону, но не получила ответа»), которые люди могут понимать, только используя определенный набор базовых знаний. В их создании центральную роль, наряду с Гектором Левеком и Леорой Моргенштерн, сыграл Эрни, который, кроме того, собрал коллекцию схем Винограда онлайн: https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html.

[29] Цитата взята из пьесы Чапека 1920 года «R.U.R.» («Универсальные Роботы Россума»), где это слово впервые стало достоянием широкой публики.

[30] Мы не говорим сейчас о роботах-поварах в ресторанах быстрого питания — фактически речь идет здесь о конвейерах по приготовлению блюд, которые могут очень точно контролировать условия жарки, выпечки, перемешивания и т.п. В крупных ресторанных сетях, таких как «Макдональдс», где расходы на живую рабочую силу чрезвычайно высоки, автоматизация, скорее всего, будет прогрессировать очень быстро.

[31] Мы вынужденно прибегаем здесь к антропоморфизмам, предполагая, что робот имеет какое-либо чувство, подобное «я», или задает себе такие вопросы. Точнее было бы написать, что алгоритм робота вычисляет, где он находится, каково его текущее состояние, каковы риски и возможности, что он должен делать дальше и как он должен осуществлять свои планы.

[32] Англ. «observe, orient, decide, and act», отсюда и аббревиатура OODA. — *Прим. пер.*

[33] Программам, которые играют в шахматы и го, приходится, конечно, иметь дело с ситуациями, которых они не видели прежде, но там могут возникать только такие виды новых ситуаций, которые входят в общую концепцию игры, и выбор ответных действий можно запрограммировать на заранее известной основе, поэтому последствия этих действий надежно предсказуемы на основе точных алгоритмов, которые, однако, будут совершенно беспомощны в открытом мире.

[34] Я — берлинец (*нем.*). — *Прим. пер.*

[35] Хотя совершенно очевидно, что мозг хорошо структурирован анатомически, для нас остается серьезной проблемой понимание того, как именно он структурирован функционально. Создавая мозг, природа не очень позаботилась о том, чтобы нам было легко его изучать. Даже некоторые из самых четких структурных единиц в мозгу остаются довольно противоречивыми по своим функциям; например, одна из версий разделения мозга на системы 1 и 2 была недавно подвергнута резкой критике ученым, участвовавшим в создании этой самой теории.

[36] От английских слов «схватывание», «равновесие», «жажда» и «движение». — *Прим. пер*.
[37] Сторонники глубокого обучения нередко страдают своего рода империалистическими замашками в плане терминологии. Иначе говоря, они называют системой глубокого обучения любую систему, которая содержит в себе хотя бы незначительные элементы глубокого обучения, независимо от того, насколько такие системы сложны, и даже в тех случаях, когда решающую роль в ее структуре и функционировании играют другие, более традиционные элементы. Нам это напоминает идею назвать автомобиль «коробкой передач» просто потому, что трансмиссия в автомобиле играет определенную роль, или назвать человека «почкой» на том основании, что без почек мы не можем жить. Никто не спорит с тем, что почки имеют огромную значимость для функционирования биологического тела человека, но из этого не следует, что изучение медицины должно сводиться к нефрологии. Мы ожидаем, что глубокое обучение будет играть важную роль в гибридных системах искусственного интеллекта, но это не значит, что они будут полагаться исключительно на него или даже будут зависеть от него в значительной степени. Глубокое обучение, вероятно, имеет шансы стать необходимым компонентом ИИ, но одного лишь его для создания универсального машинного интеллекта явно будет недостаточно.
[38] Англ. «от слова — к вектору», где «vec» — сокращение от «vector», а 2 (two) читается так же, как «to» в значении предлога «к». — *Прим. пер*.
[39] Для кодирования слов в виде векторов использовалось и множество других методов, в частности подход, который часто называют «внедрением». Некоторые из них более сложные, некоторые более простые, но более эффективные для вычислений на компьютерах. Каждый дает несколько отличающиеся результаты, но фундаментальные ограничения для всех них будут одними и теми же.
[40] Действительно, даже если на секунду вернуться к словам, то при попытке отобразить сложные понятия в виде векторов уже выявляются серьезные проблемы. В данном случае арифметика вычитания слова «мужчина» из суммы слов «король» и «женщина» работает хорошо, однако система перевода слов в векторы в целом не очень надежна. Возьмем аналогию «низкий» — «высокий» и «красивый» — «????». Пять ответов, которые система Word2Vec сочла лучшими, чтобы подставить вместо «????», оказались такими: «высокий», «великолепный», «милый», «потрясающе красивый» и «величественный», но никак не «уродливый» — ответ, которой дал бы любой человек. Другая аналогия: «лампочка» — «светит» и «радио» — «????». В аналогичном эксперименте система отвечает «свет», «FM», «радио» и «радиостанция», а не «звук» или «музыка». И, как ни грустно, Word2Vec считает слово «мораль» более близким по смыслу к слову «аморальный», чем к слову «хороший». Несмотря на всю шумиху, поднятую вокруг этой системы, реальность такова, что Word2Vec не способна справиться даже с базовыми лексическими функциями, такими как синонимия и антонимия.

[41] Строго говоря, это можно сделать, используя такие методы, как нумерация Геделя, которая отображает каждое предложение как число, значение которого рассчитывается очень структурированным образом. Однако на поверку это оказывается своего рода пирровой победой, которая потребует отказа от самого принципа числового сходства между похожими предложениями, на которое и опираются системы, основанные на обратном распространении.
[42] Для удобства первая категория в дальнейшем называется «внешней информацией», а вторая — «внутренним знанием». — *Прим. пер.*
[43] Если на то пошло, речь могла бы идти и о «раздаче навоза лошадей, выпущенных на свободу» («free horse», естественно, может означать и «свободная лошадь»); в этом проявляется еще одно доказательство того, что наш мозг достаточно умен, чтобы уметь автоматически отсеивать нерелевантную информацию.
[44] Имя, неформально означающее в английском «имярек». — *Прим. пер.*
[45] Мы вовсе не утверждаем, что для людей это обязательно легко. Прошли десятилетия, прежде чем многие люди смогли принять тот факт, что курение повышает риск развития рака легких. В течение всего XIX века большая часть медицинского сообщества яростно сопротивлялась идее о том, что послеродовая лихорадка обусловлена инфекцией, переносимой самими врачами, не стерилизовавшими руки перед принятием родов. Это происходило не только потому, что подобная (абсолютно верная) гипотеза наносила вред профессиональной гордости врачей, но и потому, что они считали подобные рассуждения полной бессмыслицей. Ведь врачи мыли руки с мылом — как могли крошечные частицы, остававшиеся на руках (если они вообще оставались), оказаться настолько фатальными для пациенток?
[46] Знаменитый американский бейсболист. — *Прим. пер.*
[47] Широко распространенная недооценка уровня врожденных знаний у человека может быть связана с одним примечательным фактом из области постнатального развития *Homo sapiens*. Голова у человеческого младенца необычайно велика по отношению к диаметру родового канала, таким образом, мы рождаемся задолго до того, как наш мозг сформируется настолько, что сможет обеспечить хотя бы минимальную самостоятельность (в отличие от целого ряда животных, которые с самого момента рождения могут самостоятельно ходить и ориентироваться). Вероятно, наш мозг продолжает физически развиваться и созревать эндогенно и отчасти независимо от опыта, получаемого после рождения, точно так же как волосы на лице не появляются до полового созревания. Далеко не все, что происходит с высшей нервной деятельностью человека в первые несколько месяцев жизни, изучено удовлетворительно, тем не менее люди часто связывают практически каждое постнатальное усложнение поведения именно с приобретаемым опытом, переоценивая важность обучения и не уделяя должного внимания роли генетических факторов.
[48] Строго говоря, ни одна система искусственного интеллекта не может обойтись совсем без врожденных структур. Каждая программа глубокого обучения, например, изначально наделена своими программистами тем или

иным количеством уровней, определенной схемой взаимосвязанности между узлами, заранее прописанными математическими функциями для воздействия на входы в эти узлы, оговоренными правилами обучения, конкретными схемами того, что подразумевается под входными и выходными блоками, и т.д.
[49] По иронии судьбы, важнейший вклад в принцип «врожденности» глубокого обучения внес один из самых ярких антинативистов среди всех ИИ-разработчиков, наш коллега из Нью-Йоркского университета Янн Лекун, главный научный советник Facebook. В своих ранних работах Лекун решительно выступал за введение изначального смещения в нейронных сетях (называемого сворачиванием), которое уже почти повсеместно применялось в компьютерном зрении. С помощью этого метода создаются сети, инвариантные к переносу (то есть которые распознают объекты в любых местах) даже до получения опыта.
[50] Never-Ending Language Learner, то есть «бесконечное обучение языку». — *Прим. пер.*
[51] Термин «причинно-следственная связь» также используется более узко, чтобы обозначать, в частности, отношения вида «A — причина B», например, щелчок переключателя [A] замыкает цепь, по которой электричество течет к лампочке [B]. Утверждение типа «объект внутри закрытого контейнера не может выйти» является частью причинной теории в широком смысле, в котором мы и используем этот термин, поскольку оно ограничивает то, что может происходить с течением времени, но не в узком смысле, так как оно не приводится с точки зрения одного события, вызывающего другое. Адекватный общий искусственный интеллект должен уметь справляться с причинностью как в широком, так и в узком смысле.
[52] В англоязычной терминологии они называются ярлыками — shortcuts. — *Прим. пер.*
[53] В человеческом поведении очень примечательно следующее: мы никогда не знаем всего перечня точных физических законов, работающих в каждой конкретной ситуации, однако это не значит, что мы не в состоянии правильно вести себя в реальном мире. Например, когда мы делаем яичницу, произойти теоретически может много чего, но все-таки здесь действуют довольно понятные принципы поведения. Мы знаем, что нацепить яйцо на вилку, когда оно уже пожарилось, гораздо легче, чем когда оно сырое; мы не будем ожидать, что частично приготовленное яйцо внезапно превратится в слона. Хороший искусственный интеллект, вероятно, будет подражать в этом отношении людям с их универсальным, гибким и почти всегда эффективным пониманием окружающей обстановки, даже когда в ней известна далеко не каждая деталь.
[54] Строго говоря, здесь потребуется более сложная база знаний, чем те пункты, которые мы перечисляем в этом списке. Например, самое первое утверждение не может относиться к объектам, которые уже находятся на дне шахты лифта, и требует более точного описания закономерностей, приводящих к падению предметов. Это хороший пример того, почему так сложно

предоставить машине не просто правильные знания, а максимально точные, релевантные и притом в правильной последовательности.
[55] Аварии и злонамеренные действия — совершенно разные вещи с точки зрения закона и морали, однако они сходны по своим последствиям и создаваемым ими техническим проблемам. Например, входная дверь в доме может сломаться или перестать нормально запираться как в результате повреждения ее ураганом, так и вследствие того, что ее взломали преступники. Те же принципы справедливы и для киберпространства.
[56] «Этичные», или «белые», хакеры — специалисты по компьютерной безопасности, занимающиеся в свободное время выявлением уязвимых мест в различных компьютерных системах, сетях и устройствах с точки зрения безопасности и надежности их работы. Применяемые ими методы часто идентичны тем, что находятся на вооружении у «черных хакеров», то есть киберпреступников, однако целью белых хакеров является предотвращение преступлений. — *Прим. пер.*
[57] Как только мы перейдем к созданию роботов с глубоким осознанием собственной сущности, способных к самоанализу, самосовершенствованию и постановке целей, можно даже не сомневаться, что очень быстро появятся серьезные и трудноразрешимые этические проблемы. Сегодня мы без колебаний удаляем с компьютера или телефона любое приложение, которое оказалось для нас бесполезным или просто устарело. Очевидно, что куда более сложные вопросы, связанные с этим, возникнут в отношении гуманоидных роботов, которые действительно ощутят себя самостоятельными существами, скажем, на уровне представителей семейства гоминид, не являющихся в строгом смысле людьми. Было бы этически оправданно наносить им вред? Разделять их на исходные составные части? Отключать их навсегда? А если они достигнут уровня человеческого интеллекта, будет ли необходимо предоставлять им гражданские и имущественные права? Можно ли (и нужно ли) распространить на них уголовное право? Понятно, что от этого мы пока еще очень далеки. Когда в 2017 году говорящий робот София (по сути — простейшая форма гуманоидной материализации чат-бота) получила гражданство Саудовской Аравии, это был, разумеется, обычный рекламный ход, а не новая веха в создании универсального искусственного интеллекта: эта система опиралась на старые трюки, используя заранее запрограммированные фразы, а не настоящий человеческий язык, выдающий подлинное понимание себя и мира.
[58] Конечно, существуют и более сложные моменты. Должна ли интеллектуальная рекламная система, наделенная здравым смыслом и ценностями, препятствовать вражеским государствам вмешиваться в новостные ленты? Следует ли разрешить интеллектуальному приложению службы знакомств, ограниченному системой обязательных ценностей, вмешиваться в уже возникшие личные отношения, предлагая своим пользователям бесконечное искушение в виде парада якобы более привлекательных альтернатив? Разумные люди могут не согласиться как с целым рядом запретов, так и со многими разрешительными законами.

[59] Здесь и далее: указание российского издательства означает, что эта книга была выпущена на русском языке. — *Прим. ред.*

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

Маркус Г.
Искусственный интеллект: перезагрузка: Как создать машинный разум, которому действительно можно доверять / Гэри Маркус, Эрнест Дэвис. — Пер. с англ. — М.: Альпина ПРО, 2021.

ISBN 978-5-2060-0030-6