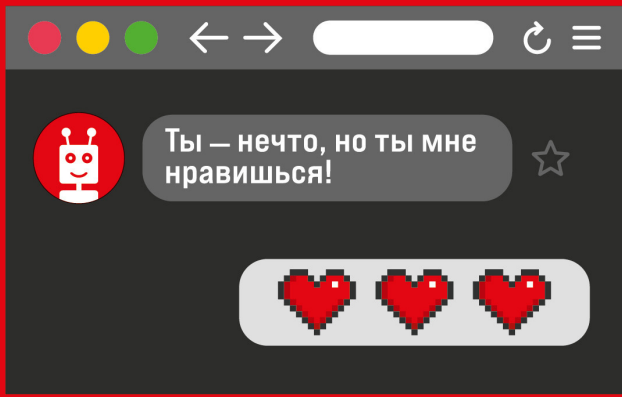
ДЖАНЕЛЬ ШЕЙН

КОКЕТЛИВЫЙ ИНТЕЛЛЕКТ

КАК НАУЧИТЬ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ФЛИРТОВАТЬ?



ПРОСТЫМИ СЛОВАМИ И С ЮМОРОМ:

- ОБ ИСПОЛЬЗОВАНИИ АЛГОРИТМОВ
- ОБ ОБУЧЕНИИ НЕЙРОСЕТЕЙ
- О КОНСТРУИРОВАНИИ СИМУЛЯЦИЙ

БОМБОРА

Джанель Шейн

Кокетливый интеллект. Как научить искусственный интеллект флиртовать?

Серия «Библиотека ИТ. Главные книги о современных технологиях»

indd предоставлен правообладателем
http://www.litres.ru/pages/biblio_book/?art=69363850
ISBN 978-5-04-188683-7

Аннотация

В этой книге вы найдете множество удивительных и смешных фактов об искусственном интеллекте, о которых вы и понятия не имели. Как он работает? Умеет ли он флиртовать? Есть ли у него чувство юмора? Лектор TED Жанель Шейн знает ответы на все эти и многие другие вопросы и готова поделиться ими с читателями.

В формате a4.pdf сохранен издательский макет.

Содержание

Жанель Шей
Кокетливый интеллект
Как научить искусственный интеллект флиртовать?

Janelle Shane
YOU LOOK LIKE A THING AND I LOVE YOU

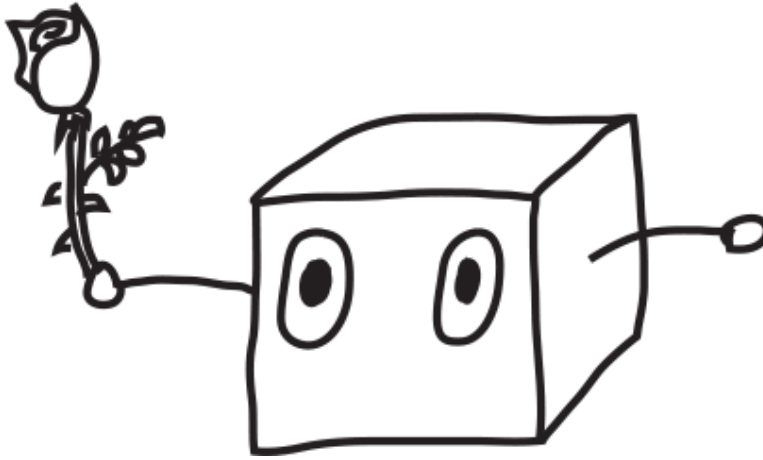
* * *

Посвящается читателям моего блога, которые смеялись над каждой глупостью, рисовали для меня чумовых существ, нашли всех жирафов и готовили печенье по рецепту от нейросети. Спасибо, что смирились даже с брауни из хрена.

Посвящается также моим родным – за то, что

вы самые большие мои фанаты.

Введение
ИИ повсюду



Я на самом деле не ставила целью проекта научить ИИ[1] флиртовать.

Вообще-то, я сделала уже порядочно своеобразных проектов на тему искусственного интеллекта. В блоге под названием AI Weirdness[2] я рассказывала о том, как учила искус-

---

[1] ИИ – искусственный интеллект. – *Здесь и далее примечания переводчика, если не указано иное.*

[2] «Странности ИИ», https://aiweirdness.com.

ственный разум придумывать имена для котов – среди них попадались на редкость неудачные, например Мистер Бубенцы или Рыгуша – и просила сочинять новые рецепты блюд, а получала такие, в которых требовалось взять «почищенный розмарин» или горсть битого стекла. Но попытка натренировать компьютер обольщать людей – уже нечто совсем иное.

ИИ обучается на примерах – в этом случае на пикаперских фразах. Проблема в том, что в обучающий набор входили фразы, собранные мной из разных уголков Интернета, и все они были ужасны. Там попадались высказывания из всех категорий, начиная с тупых каламбуров и заканчивая неприличными намеками. После того как ИИ натренировался сочинять подобные фразы, их можно было бы производить на свет божий тысячу за тысячей одним нажатием кнопки. Однако, как готовый впитать любые впечатления ребенок, искусственный интеллект не знает, что следует повторять, а что нет. Он начинает с чистого листа, ничего не зная о том, что такое пикаперские фразы (и даже что вообще такое язык), и затем учится на примерах, изо всех сил стараясь уловить закономерности и воспроизвести то, что от него требуют. В том числе грубость. Просто он не знает ничего получше.

Я подумывала бросить затею, но пост в блог сам себя не напишет, так что пришлось пересилить себя и потратить уйму времени на сбор образцовых пикаперских фраз. Итак, я начала обучать алгоритм. ИИ принялся искать закономерности и шаблоны в примерах, изобретая и проверяя правила,

которые помогли бы ему спрогнозировать, какие нужны буквы и в каком порядке они должны располагаться, чтобы получилась пикаперская фраза. Наконец тренировка окончилась. Испытывая некоторый мандраж, я попросила ИИ выдать несколько фраз.

Наверное, ты треугольник? Потому что только ты здесь чего-то стоишь.

Эй, детка, ты, должно быть, ключ? Потому что я выдерживаю твой свисток.

Ты свеча? Это потому что ты такая знойная от видов с тобой.

Ты так прекрасна, что ты говоришь летучую мышь на меня с деткой.

Ты – нечто, но ты мне нравишься.

Я удивилась и порадовалась. Виртуальный мозг ИИ (примерно того же уровня сложности, что и у червя[3]) неспособен был уловить оттенки смыслов в наборе данных, распознать женоненавистничество или отсеять низкосортные шутки. Но те закономерности, которые ему удалось выявить, позволили ему добиться наилучшего результата… причем неожиданным, может, даже лучшим способом: решая самую главную задачу – как заставить улыбнуться незнакомку.

Хотя для меня итоговые фразы свидетельствовали о бес-

[3] Например, у червя *Caenorhabditis elegans*, геном и строение которого изучены в мельчайших подробностях, мозг содержит ровно 302 нейрона. – *Прим. науч. ред.*

спорном успехе, скудоумие моего ИИ-партнера, вероятно, удивит того, кто знает об искусственном интеллекте лишь из заголовков современных новостей или из научной фантастики. Очень часто компании заявляют, что ИИ способен воспринимать нюансы человеческой речи так же хорошо, как люди, и даже лучше, или, что еще чуть-чуть, и он заменит людей во многих профессиях. Скоро искусственный интеллект будет повсюду, трубят пресс-релизы. Это одновременно и правда, и нет.

На самом деле ИИ *уже* повсюду. Он формирует все, с чем мы сталкиваемся онлайн, определяет, какую рекламу мы видим в Интернете, предлагает видеоролики и в то же время распознает ботов в социальных сетях и вредоносные сайты. Компании используют ИИ-программы для сканирования резюме кандидатов на вакансии, а в банках искусственный интеллект решает, кому можно выдавать кредит. Беспилотные автомобили со встроенным ИИ уже проехали миллионы километров (прибегая к помощи человека лишь в моменты замешательства). В наших смартфонах ИИ распознает голосовые команды, автоматически отмечает лица людей на фотографиях и даже применяет к видеопотоку фильтры – благодаря им у нас в кадре, например, вырастают чудесные кроличьи уши.

По опыту мы знаем, что тот ИИ, с которым мы сталкиваемся каждый день, крайне далек от совершенства. Приложения, предлагающие рекламу, без конца преследуют нас в

браузерах рекламой ботинок, хотя мы их уже купили. Фильтры в электронной почте иной раз пропускают откровенно мошеннические послания и, наоборот, в самый неподходящий момент убирают в папку спама важное для нас письмо.

ИИ все больше влияет на нашу повседневную жизнь, и его причуды начинают проявляться в последствиях таких масштабов, что уже нельзя говорить о простом неудобстве. Рекомендательные алгоритмы YouTube подсовывают зрителям контент все более полярного характера: от популярных каналов новостей к видеороликам с пропагандой ненависти и теорий заговора ведет дорога всего из нескольких кликов[4].

[4] Caroline O'Donovan et al., "We Followed YouTube's Recommendation Algorithm Down the Rabbit Hole," BuzzFeed News, January

Алгоритмы, принимающие решения о досрочном освобождении заключенных, о предоставлении кредитов, программы скрининга резюме не беспристрастны и иногда страдают от тех же предрассудков, что и люди, которых они призваны заменить, может, даже в большей степени. Интеллектуальные системы наблюдения неподкупны, но у них также не возникнет возражений, если от них потребуют сделать нечто аморальное. Кроме того, они могут выдавать ошибки из-за неправильного использования или взлома. Исследователи выяснили, что иногда такая вроде бы малозначительная вещь, как небольшая наклейка, способна заставить систему распознавания принять пистолет за тостер и что сканер отпечатков пальцев, не очень продвинутый в плане безопасности, можно более чем в 77 % случаев обдурить с помощью единственного универсального отпечатка[5].

Одни люди представляют ИИ более умным, чем он есть на самом деле, и способным делать то, что возможно лишь в научной фантастике. Другие говорят, что создали беспристрастный ИИ, в то время как в его поведении просматриваются явные и измеримые искажения. А еще нередко за работу ИИ выдается результат деятельности людей. Как жителям Земли, нам нужно научиться не попадаться на эту удочку. Нужно понять, как используются наши данные и чем явля-

---

24, 2019, https://www.buzzfeednews.com/article/carolineodonovan/down-youtubes-recommendation-rabbithole.

[5] Аналога так называемого «мастер-ключа». – *Прим. науч. ред.*

ется ИИ, а чем не является.

На сайте AI Weirdness я много времени посвящаю различным забавным экспериментам с искусственным интеллектом. Иногда я заставляю его делать необычные вещи, например, имитировать эти пикаперские фразы. Или стараюсь вывести его из зоны комфорта, как в тот раз, когда я «скормила» алгоритму распознавания образов картинку с Дартом Вейдером и просто спросила его, что он увидел: алгоритм объявил, что Дарт Вейдер – это дерево, и начал спорить со мной, отстаивая свою точку зрения. В результате экспериментов я выяснила, что даже самое четкое задание может поставить ИИ в тупик, как если бы вы над ним подшутили. Но оказывается, что разыгрывать ИИ – то есть подсовывать ему задачу и наблюдать, как он ломает о нее зубы, – крайне поучительно и помогает узнавать о нем больше.

Как вы увидите в этой книге, зачастую внутри ИИ происходят настолько странные и запутанные процессы, что анализ выходных данных становится одним из немногих способов выяснить, что искусственный интеллект понял, а в чем ужасно ошибается. Когда вы просите ИИ нарисовать кота или написать шуточную фразу, он начинает делать те же ошибки, что и в процессе распознавания отпечатков пальцев или сортировки медицинских фотоснимков. Вот только тут вы сразу поймете, что что-то пошло не так, если у кота на картинке окажется шесть лап или шутка лишится ключевой фразы в конце. Ну и еще это безумно смешно.

В попытках вывести ИИ из зоны комфорта и заставить его заниматься человеческими делами я требовала от него написать первые строки романа, узнавать на изображениях овец в необычных местах, давать имена морским свинкам и вообще чудить по-всякому. Это позволяет очень многое узнать о том, в чем ИИ хорош, в чем – не очень, а чего ему, скорее всего, не удастся достичь за время моей или вашей жизни.

И что же я узнала?

Пять принципов странности ИИ:

• ИИ опасен не потому что он умен, а потому что он умен недостаточно;

• по силе интеллекта ИИ находится примерно на уровне червя;

• ИИ в действительности не понимает задачу, которую вы перед ним ставите;

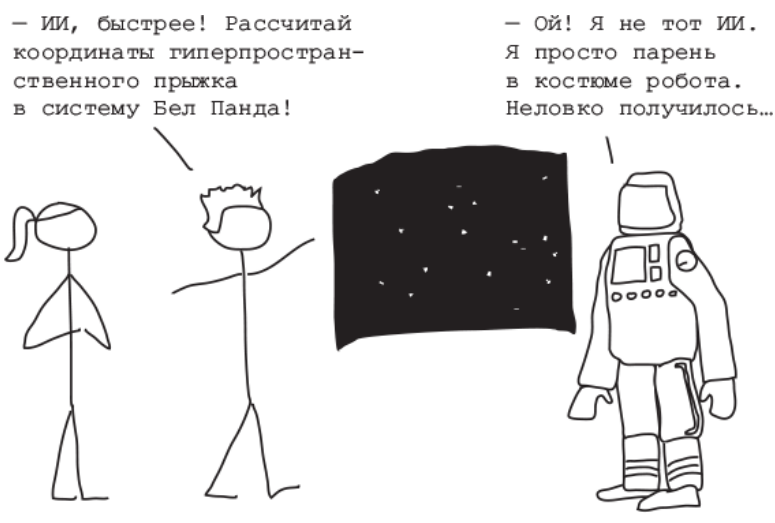
• но: ИИ будет делать *в точности* то, что вы от него хотите, – по крайней мере, изо всех сил постарается;

• и еще: ИИ всегда выбирает путь наименьшего сопротивления.

Так давайте же войдем в странный мир искусственного интеллекта. Мы узнаем, что можно назвать ИИ, а что – нельзя. Выясним, в чем он хорош и в чем обречен на поражение. Поймем, почему ИИ будущего, вероятно, будут похожи не на робота С-3РО, а скорее на рой насекомых. Разберемся, почему беспилотный автомобиль не поможет спастись во

время зомби-апокалипсиса. Узнаем, почему никогда не надо вызываться проверять работу ИИ, сортирующего сэндвичи, а еще узнаем о ходячих ИИ, которые будут делать что угодно, только не ходить. Все эти истории позволят нам понять, как работает искусственный интеллект, как он думает и почему делает наш мир еще более странным.

Глава 1
Что такое ИИ?



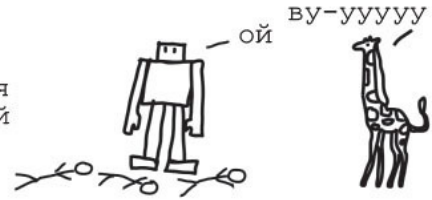
Если вам кажется, что ИИ уже повсюду, то это отчасти потому, что слова «искусственный интеллект» могут означать разные вещи – зависит от того, читаете вы фантастический роман или пытаетесь продать новое приложение для научных исследований. Когда некто заявляет, что у него есть чат-бот с ИИ, надо ли ожидать, что у этого чат-бота будет свое мнение и чувства, как у вымышленного С-3РО? Или это все-

го лишь алгоритм, научившийся догадываться, как именно люди, скорее всего, отреагируют на ту или иную фразу в диалоге? Или это электронная таблица, которая ищет слова из вашего вопроса в библиотеке заранее подготовленных ответов? А может, это человек, сидящий где-то в далекой стране на скромной зарплате и печатающий вам сообщения? Или это полностью подчиненный сценарию диалог, где человек и ИИ зачитывают фразы, как актеры в пьесе? Все эти вещи определяли как искусственный интеллект – отсюда и путаница.

В рамках своей книги я буду подразумевать под термином в основном то, что сейчас под ИИ подразумевают программисты, – вид программ, построенных на основе алгоритмов машинного обучения. Ниже я привела целую кучу терминов, о которых мы поговорим дальше, и разнесла их по категориям.

Все, что называют ИИ

ИИ в этой книге	**Тоже есть в книге, но это не ИИ**
Алгоритмы машинного обучения	ИИ из научной фантастики
Методы глубокого обучения	Основанные на правилах программы
Нейронные сети	Люди в костюмах роботов
Нейросети с обратными связями (рекуррентные)	Роботы, действующие по сценарию
Цепи Маркова	Люди, которые за деньги выдают себя за ИИ
Случайный лес	Разумные тараканы
Генетические алгоритмы	Жирафы-призраки
Генеративно-состязательные сети	
Обучение с подкреплением	
Предиктивный набор текста	
Волшебные машины для сортировки сэндвичей	
Невезучие роботы-убийцы	



Все, что я здесь называю искусственным интеллектом, также можно назвать алгоритмами машинного обучения, давайте поговорим о том, что это такое.

ТУК-ТУК, КТО ТАМ?

Чтобы найти ИИ в дикой природе, важно понять, в чем

же разница между **алгоритмами машинного обучения** (именно это понимается здесь под ИИ) и традиционными программами (программисты их называют **основанными на правилах**). Если вы когда-нибудь программировали хотя бы на базовом уровне или обращались к HTML, чтобы создать дизайн сайта, значит, вы использовали основанную на правилах программу. Вы задаете список команд или правил на понятном компьютеру языке, и компьютер делает в точности то, что вы говорите ему делать. Чтобы решить задачу с помощью такой программы, вам потребуется понять, какие шаги должна выполнить программа, чтобы достичь цели, и как именно их описать.

Алгоритм машинного обучения сам додумывается до правил методом проб и ошибок, оценивая, насколько приблизился к поставленным программистом целям. Целью может быть воспроизвести что-то по примерам, достичь определенного счета в игре или что угодно еще. Пытаясь выполнить задачу, ИИ способен выявить такие правила и взаимосвязи, о существовании которых программист даже не подозревал. Программирование ИИ больше похоже на обучение ребенка, чем на разработку программы.

Программирование на основе правил

Предположим, я решила с помощью традиционного программирования научить компьютер выводить шутки «Тук-тук, кто там?». Вначале я должна выявить все правила. Я

проанализирую структуру подобных шуток и выясню, что все они строятся по определенной формуле, вот такой:

```
Тук-тук.
Кто там?
[Имя]
Как[ой/ая/ое] [имя]?
[Имя] [Ключевая фраза]
```

Теперь, когда я определилась с формулой, оказывается, что программа должна заполнить два пропуска: [Имя] и [Ключевая фраза][6]. Теперь задача сводится к тому, чтобы произвести эти элементы. Но правила все равно нужны.

Я могу подобрать список имен и подходящих ключевых фраз, например:

Имена	**Ключевые фразы шутки или каламбура**
Рада	вам предложить установку радиаторов отопления!
Тихон	егодяй, сколько можно сверлить!
Яна	вас жаловаться буду!
Арина	тебя не ори, ты ничего не слышишь, глухомань.
Олег	сандр!

[6] Актуально для английского языка. В русскоязычном варианте шутки программе, как видно на схеме, придется согласовать в роде местоимение «какой» со словом, подобранным для элемента [Имя]. То есть она должна будет заполнить не два, а три пропуска. – *Прим. ред.*

Теперь компьютер может выдавать шутки «Тук-тук, кто там?», выбирая пару [Имя] и [Ключевая фраза] из списка и вставляя элементы в шаблон. Таким образом нельзя получить *новые* шутки – лишь те, что я и так знаю. Я могу попытаться сделать программу поинтереснее, разрешив заменять, скажем, [вас жаловаться буду!] на другие подходящие фразы: [родный артист больших и малых театров!] или [лошади верхом приехал!]. После этого программа может выдать новую шутку:

```
Тук-тук.
Кто там?
Я на.
Какая Яна?
Я на лошади верхом приехал!
```

Есть также вариант разрешить заменять слова [лошади] на [бешеном волке], или [кенгуру], или что угодно еще. Тогда мой компьютер сумеет вывести еще больше шуток. Создав достаточно правил, по идее, я получу сотни разных фраз.

В зависимости от целевого уровня сложности я могу потратить много времени на формулировку дополнительных правил. Я могу отыскать список готовых шуток и придумать, как преобразовывать их в нужный мне формат с ключевыми фразами. Я даже могу попытаться включить в программу правила произношения, рифмовки, использования частич-

ных омофонов и отсылок к культурному контексту, чтобы добиться от компьютера максимально интересного результата от их комбинирования. При достаточном уровне мастерства я бы даже сумела составить программу для генерации новых шуток, которых раньше никто не слышал. (Хотя один человек попробовал сделать это – в результате его алгоритм выдавал настолько архаичные и невразумительные слова и фразы, что почти никто не мог понять получившиеся шутки.) Неважно, насколько сложным окажется мой набор правил, я все равно говорю компьютеру, как в точности решать поставленную задачу.

Обучение ИИ

Когда же я учу ИИ генерировать шутки «Тук-тук, кто там?», то не создаю никаких правил. ИИ приходится создавать эти правила самостоятельно.

Я предоставляю лишь набор готовых шуток и инструкции, которые сводятся к указанию: «Вот тебе шутки; сделай такие же, и побольше». Из чего он будет их делать? Из кучи случайных букв и знаков пунктуации.

Вручив ему все это, я иду выпить кофе.

ИИ принимается за работу.

Первым делом он пытается угадать, какие отдельные буквы появляются в нескольких подобных шутках. На этом этапе угадывание происходит на 100 % случайно, так что первый образец может выглядеть как угодно. Допустим, полу-

чается нечто вроде «цьрции дси, хс?чафк.». По мнению ИИ, так люди шутят.

Потом алгоритм смотрит, что эти шутки *в действительности* должны из себя представлять. Скорее всего, он выяснит, что был в корне неправ. «Ну хорошо», – говорит себе ИИ, после чего слегка меняет свою структуру, чтобы в следующий раз угадывать лучше. Есть ограничение на степень изменений, мы ведь не хотим, чтобы он пытался запомнить любой новый увиденный кусок текста. Но при минимальной модификации ИИ может выяснить, что если начнет производить только буквы «к» и пробелы, то окажется прав хотя бы в чем-то и где-то. После сверки с одним набором реальных шуток и одного раунда модификации представление ИИ о подобных шутках станет похоже на нечто такое:

```
    К К К   К К   К
  ТКК     К ТККККОК
К ККККК
К


КК
  КК   К     КК

ТККУКК   К


   К
  К
```

Ну что же, это не лучшая шутка из подобных. Но, беря такой вариант за основу, ИИ смотрит на второй набор данных, потом на следующий. Каждый раз он подстраивает формулу,

чтобы улучшать точность догадок.

После еще нескольких раундов, состоящих из догадок и самопроверок, искусственный интеллект усваивает несколько новых правил. Например, он догадывается, что в конце некоторых строк следует ставить вопросительный знак. Также он начинает применять гласные (в особенности букву «у») и даже пытается расставлять запятые.

```
нуу,
лтунуу
Кут?
внос у кг
птб оа то,

ткоуЕтнл
игр ноос
док кКе
в це
е
```

Как думаете, насколько те правила, что он вывел для шуток «Тук-тук, кто там?», соответствуют реальности? Кажется, он по-прежнему что-то упускает.

Если ИИ хочет приблизиться к цели и произвести на свет приемлемую шутку, ему еще нужно определиться с правилами по поводу того, в каком *порядке* в ней могут следовать буквы. И вновь он начинает строить догадки. Что, если после буквы «у» всегда идет буква «г»? На поверку оказыва-

ется, что догадка не очень правильная. Потом он понимает, что довольно часто после «у» встречается «к», а перед «о» – сочетание «кт». Блеск. Наконец какой-то успех. Теперь посмотрим, как выглядит, по мнению искусственного интеллекта, идеальная шутка:

```
Ктоук
Ктоук
Ктоук
Ктоук
Ктоук Ктоук Ктоук
Ктоук Ктоук
Ктоук
Ктоук
```

Ну, это не очень-то напоминает шутку «Тук-тук, кто там?» – больше похоже на куриное кудахтанье. ИИ предстоит отыскать еще несколько правил.

Он вновь изучает набор данных. Затем пытается использовать найденные сочетания букв новыми способами, выискивая примеры комбинаций, которые лучше соответствуют заданным примерам шуток.

```
нток докк хомк
уКуу мКток
Тук
кая Авас Тыы
кол хомм
Хамм?
```

```
Рие
ако ак, Ото и клеа
то ко- оо к АтьХпал Бко
Егоч
ткот- К окт
Тма
туу
Ктук Тук Ток Тамк
```

Все эти улучшения происходят всего за несколько минут. К моменту, когда я возвращаюсь к компьютеру с чашкой кофе, ИИ *уже успел* понять, что если начать шутку с комбинации «Тук-тук! Кто там?», то совпадений с существующими подобными шутками будет больше. И он решает следовать этому правилу *всегда*.

Но остальная часть формулы выводится еще довольно долго, причем ИИ часто словно бы принимается сам с собой играть в испорченный телефон:

```
Тук-тук!
Кто там?
Вра
Какая Диял?
Какой Нна?
Какая Ананан?
Какой Нин?
Какой Няни?
Какая Двима?
```

Наконец ему удается нащупать структуру шуточной формулы, но не подходящие *слова*:

```
Тук-тук!
Кто там?
Пана
Какая Лагана?
Кольке ты вам где не
```

Иногда он немного увлекается особо длинными ключевыми фразами.

```
Тук-тук!
Кто там?
Валесий
Какая?нна?
Ыши тьнан уду шенквв аксабн шуе такрещ утомцй дуб яван аолвакт хиртиуза уурнаст котекц шдсеевка озо эугойм за доен кмтакан ахелко тас чигепо ббнащ к реужмым!
```

Где-то дальше в ходе обучения искусственный интеллект узнает про шутку о корове без губ, она становится его любимой. Если что, это примерно такой шуточный диалог:

```
Тук-тук!
Кто там?
Корова без губ.
Какая корова без губ?
Корова без губ говорит: «у-у-у-у!»
```

ИИ считает, что наткнулся прямо-таки на юмористиче-

скую золотую жилу, и какое-то время исходит из того, что лучшая ключевая фраза в любой шутке – это «у-у-у-у!»…

Тук-тук!
Кто там?
Алек
Какая Аник?
Вау ку кака вам ска ска дверь в за сажас у-у у-у-у у-у-у-у у-у-у у-у у-у у-у у-у у-у-у у-у у-у у-у-у у-у-у у-у-у-у у-у-у у-у-у у-у у-у у-у-у у-у-у у-у-у у-у-у у-у-у у-у-у у-у у-у у-у-у у-у-у у-у-у у-у-у у-у-у у-у у-у у-у-у у-у-у у-у-у у-у-у…

Что ж, ладно, правило он нашел. В конце концов ИИ (как мне кажется, с неохотой) начинает реже создавать ключевые фразы, опираясь на шутку про корову без губ. Но одна проблема по-прежнему остается: его диалоги как будто списаны с ситуации, когда один человек другому пытается по телефону рассказать шутку «Тук-тук, кто там?», но связь очень плохая, и временами собеседники, кажется, вообще забывают, кто из них шутник:

Тук-тук!
Кто там?
Лаша
Какой Каша?
Какой какой Каша?
Какой какой?

```
Тсаша и на пристани?
```

В итоге ИИ понимает структуру шутки и начинает составлять с формальной точки зрения правильные диалоги, однако совершенно бессмысленные. Многие из них он частично списывает из обучающей выборки.

```
Тук-тук!
Кто там?
Робин
Какой Робин?
Ты Робин! Гни меня и я ухожу

Тук-тук!
Кто там?
Геннадий
Какой Говядина?
Говядина тор без губ тктктто

Ирландия
Какая Ирландия?
Ирландия твои деньги, задница
```

И тут. Он. Вдруг. Написал. Настоящую шутку. Которую составил полностью сам, не заимствуя кусок из обучающего набора, и при этом она оказалась не только вразумительной, но и… действительно смешной?

```
Тук-тук!
Кто там?
```

```
Алик
Какой Алик?
Алик из тук-тук шутки
```

Значит ли это, что ИИ внезапно начал понимать суть такого рода шуток и каламбуров на человеческом языке? Почти наверняка нет, учитывая очень скромный размер набора данных. Но свобода, которой ИИ пользовался, – свобода пастись в пространстве всех возможных символов, – позволила ему пробовать новые комбинации звуков, и оказалось, что одна из них попала в точку. Так что это скорее победа теоремы о бесконечных обезьянах[7], чем доказательство того, что ИИ скоро отнимет хлеб у стендап-комиков.

Красота решения, при котором мы позволяем ИИ создавать собственные правила, заключается в том, что этот единый подход – «вот тебе данные; придумай, как их воспроизвести» – работает на большом количестве задач. Если бы я предложила алгоритму-«шутнику» другой набор данных вместо шуток типа «Тук-тук, кто там?», он приучился бы использовать именно его.

[7] Старое предположение, что если посадить бесконечное число обезьян за пишущие машинки и если они будут бить по кнопкам бесконечное количество времени, то рано или поздно одна из них напечатает идеальную копию собрания сочинений Шекспира [в русской версии – все четыре тома «Войны и мира». – *Прим. перев.*], на самом деле хорошо описывает так называемый метод поиска решения задачи «грубой силой», то есть систематическим перебором всех возможных вариантов. В идеале использование ИИ демонстрирует преимущества над этим методом. В идеале. – *Прим. авт.*

Он может придумывать названия новых видов птиц:

Юкатанская джунглевая утка
Лодкоклювая нектарница
Западный вилоклювый цветосос
Черноголовая пушистохвостка
Исландский болотный печник
Снежный стенающий цаплевый дрозд

Или новые марки парфюма:

Изысканная десятка
О-де-бофф
Лягушистый цветок
Мамкин
Санта для дам

Или даже рецепты новых блюд.

Простые глазированные моллюски

основное или первое блюдо
1 фунт курятины
1 фунт свинины, нарезанная кубиками
½ зубца чеснока, раздавить
1 чашка сельдерея, нарезанная ломтиками
1 голова (около ½ чашки)
6 столовых ложек электрического миксера
1 чайная ложка черного перца
1 луковица, кусочками

3 чашки говяжьего бульона свонда для фрукта
1 измельченная пятьдесят на пятьдесят; воды нужного количества

В сковороду объемом 3 кварты поместить лимонный сок в виде пюре и дольки лимона.
Добавить овощи, добавить курицу в соус, хорошо перемешивая лук и добавляя. Добавить лавровый лист, красный перец, медленно накрыть крышкой и кипятить под крышкой на малом огне 3 часа. Добавить картофель и морковь в кипящий бульон. Подогревать, пока соус не закипит. Подавать с пирожками.
Если вшивые кусочки приготовили десерты, и готовить над воком.
Замораживать до получаса, украсив.
На 6 персон.

ПОЗВОЛЬТЕ ИИ ОБО ВСЕМ ДОГАДАТЬСЯ

Получив в распоряжение набор шуток «Тук-тук, кто там?» и никаких инструкций вообще, ИИ сумел открыть массу правил, которые в противном случае мне пришлось бы

вводить в программу вручную. Некоторые из них я бы ни за что не додумалась программировать, а о существовании других даже не подозревала, например, правила «о превосходстве шутки про корову без губ».

Именно этот фактор и делает системы, основанные на искусственном интеллекте, привлекательными для решения задач, особенно в тех случаях, когда правила действительно сложны или покрыты мраком. Например, ИИ часто используют для распознавания визуальных образов – это удивительно сложная область, где пасуют обыкновенные компьютерные программы. Хотя большинство из нас легко узнают на картинке кошку, сформулировать правила, определяющие, как же выглядит кошка, по-настоящему трудно. Стоит ли нам сообщить программе, что у кошки два глаза, один нос, два уха и хвост? Но эти признаки точно так же указывают и на мышь, и на жирафа. А что, если кошка на картинке свернулась в клубок или ее голова повернута вбок? Записать правила для обнаружения на фотографии даже одного-единственного глаза и то очень непросто. Но ИИ может просмотреть десятки тысяч изображений кошек и составить правила, по которым будет верно опознавать кошку в большинстве случаев.

Иногда ИИ – лишь небольшая часть программы, в остальном представляющей собой основанный на правилах сценарий. Возьмем в качестве примера

программу, которая позволяет клиентам банка по телефону получать информацию о состоянии счета. ИИ для распознавания голоса переводит произнесенные человеком звуки в действие – выбор вариантов из голосового меню, но за список пунктов, доступных каждому клиенту, и за определение того, какой счет ему принадлежит, отвечают правила, заданные программистом.

Другой вариант – это когда программа первым пускает в бой ИИ, но когда у того возникают трудности, контроль над ситуацией переходит к людям; такой подход называется псевдо-ИИ. Так работают чаты пользовательской поддержки. Если ваши фразы сбивают бота с толку или если бот понимает, что вы начинаете злиться, вас переводят в диалог с человеком. (И этому человеку теперь придется иметь дело с ничего не понимающим и/или разозленным клиентом – возможно, лучше бы открыть опцию «говорить с живым человеком» не только для клиента, но и для работника.) Современные беспилотные автомобили устроены похожим образом – человеку-водителю всегда надо быть готовым принять управление, если ИИ перенервничает.

Кроме того, искусственный интеллект замечательно проявляет себя в стратегических играх наподобие шахмат – в них мы в точности знаем, как описать все возможные ходы, ноне способны вывести формулу, которая подскажет, какой

следующий ход будет наилучшим. В шахматах из-за количества вариантов ходов и сложности игрового процесса даже гроссмейстер не в состоянии сформулировать жесткие правила для предсказания того, какой ход окажется лучшим в той или иной ситуации. А самообучающийся алгоритм может сам с собой сыграть миллионы тренировочных партий – больше, чем сыграет за всю жизнь самый умный и упорный гроссмейстер, – чтобы выработать правила, которые будут приводить его к победе. И поскольку ИИ обучается без явных инструкций, иногда он находит очень необычные и оригинальные стратегии. Иной раз *чрезмерно* оригинальные.

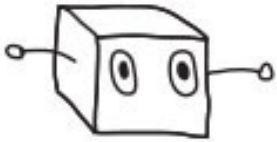
Если вы не скажете ИИ, какие ходы допустимы, он может отыскать странные лазейки и злоупотребить ими, лишив игру смысла. Например, в 1997 году группа программистов создавала алгоритмы, которые удаленно играли в крестики-нолики друг против друга на бесконечно большом поле. Один из программистов, вместо того чтобы разработать основанную на правилах стратегию, позволил ИИ самостоятельно формировать подход к игре. Внезапно этот ИИ стал побеждать во всех матчах. Его стратегия заключалась в том, чтобы делать ход где-то очень-очень далеко.

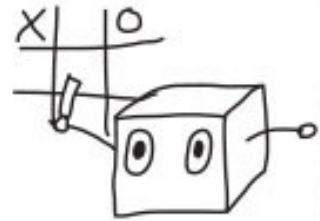
Размер нового игрового поля оказывался настолько большим, что компьютер оппонента, пытаясь его у себя воспроизвести, исчерпывал ресурсы оперативной памяти и падал с ошибкой, так что ему засчитывалось техническое пораже-

ние[8]. У большинства программистов, работающих с ИИ, есть в запасе похожие истории – о том, как алгоритмы удивляли их тем, что находили неожиданные решения. Иногда такие решения гениальны, а иногда создают проблемы.

[8] Joel Lehman et al., "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities," ArXiv:1803.03453 [Cs], March 9, 2018, http://arxiv.org/abs/1803.03453.

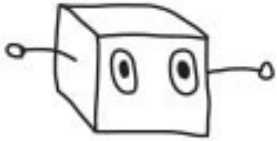
Хожу на
(-1;+1).

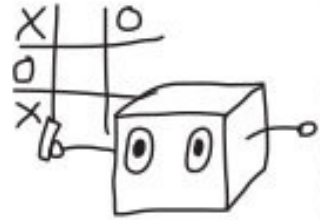




(-1;+1), понятно.
Хожу на (+1;+1).

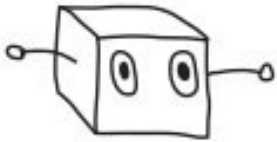
Хожу на
(-1;-1).

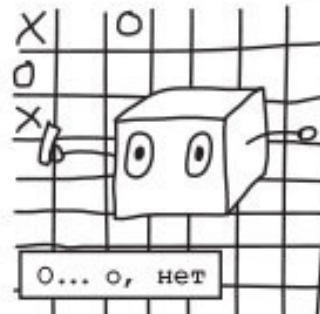




(-1;-1), понятно.
Хожу на (-1; 0).

Хожу на
(+999; -999).





О... о, нет

Самое основное, в чем нуждается ИИ, это конкретная цель и набор данных для обучения. Получив их, он начинает гонку, и неважно, какова цель: принять решение о выдаче кредита, как это делает специалист-человек, предсказать, приобретут ли покупатели определенный носок, добиться максимального счета в видеоигре или же заставить робота преодолеть наибольшее расстояние. В каждом случае ИИ методом проб и ошибок изобретает правила, которые позволят ему добиться цели.

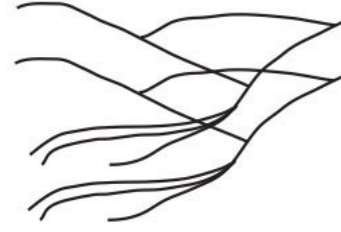
ИНОГДА ЕГО ПРАВИЛА ПЛОХИ

Бывает, что правила, прекрасно помогающие ИИ находить решение, оказываются основаны на неверных предположениях. Например, некоторые из самых причудливых экспериментов я проделывала с продуктом компании Microsoft для распознавания образов, который подбирал теги и описание для загружаемого изображения. Как правило, этот алгоритм правильно распознает предметы: узнает облака, поезда метро или даже ребенка, выполняющего ловкие трюки на скейтборде. Но однажды мое внимание привлекло нечто странное в результатах его работы: ИИ ставил тег «овцы» картинкам, где определенно не было никаких овец. Я изучила проблему и выяснила, что алгоритм видел овец на сочно-зеленых полях вне зависимости от того, были они там на

самом деле или нет. Почему же столь специфическая ошибка всплывала вновь и вновь? Возможно, во время обучения этому ИИ в основном показывали овец, находящихся на таких вот полях, и он не понял, что заголовок «овцы» относится к животным, а не к полям. Другими словами, искусственный интеллект смотрел не туда. И, конечно же, когда я показывала ему овец, которые *не* паслись на пышных пастбищах, он чаще всего ошибался. Овец в автомобилях он обычно помечал как собак или кошек. Овцы в жилых помещениях у него также становились собаками или кошками, то же самое происходило с ягнятами, которых кто-нибудь держал на руках. А овцы на привязи распознавались как собаки. Такие же проблемы у ИИ были с козами: если он видел козу, залезшую на дерево (они так иногда делают), то считал, что это жираф (другой похожий алгоритм называл коз птицами).

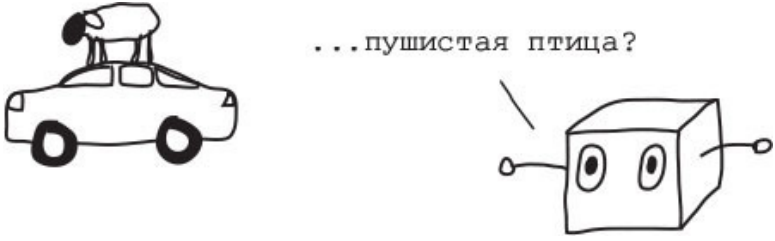
Стадо овец, пасущихся
на сочно-зеленом пастбище

Стадо овец, пасущихся
на сочно-зеленом пастбище



Я не знаю точно, но догадываюсь, что ИИ руководствовал-

ся правилами вроде «зеленая трава = овцы» и «нечто шерстяное в машине или на кухне = кошки». Они отлично ему служили во время обучения, но подвели, когда ИИ столкнулся с реальным миром и головокружительным разнообразием связанных с овцами ситуаций.



Подобные ошибки обучения характерны для распознающих образы ИИ. Но последствия этих ошибок могут оказаться серьезными. Одна команда в Стэнфордском университете как-то тренировала искусственный интеллект определять разницу между изображениями здоровой кожи и кожи, пораженной раком. Однако после завершения тренировки ученые обнаружили, что нечаянно создали ИИ, определяющий наличие на фотографии линеек, потому что многие раковые опухоли, изображения которых оказались в их наборе, были сфотографированы с приложенной для масштаба линейкой[9].

[9] Neel V. Patel, "Why Doctors Aren't Afraid of Better, More Efficient AI Diagnosing Cancer," The Daily Beast, December 11, 2017, https://www.thedailybeast.com/why-doctors-arent-afraid-of-better-more-

КАК ВЫЯВИТЬ ПЛОХОЕ ПРАВИЛО

Зачастую нелегко понять, когда ИИ делает ошибки. Мы не задаем для него правила, искусственный интеллект создает их самостоятельно, причем не записывает на бумаге и не объясняет понятным языком, как мог бы сделать человек. Вместо этого ИИ производит сложные взаимозависимые изменения своей внутренней структуры, превращая универсальную основу в нечто, хорошо приспособленное для решения конкретной задачи. Это всё равно что взять кухню, полную разных ингредиентов, и на выходе получить печенье. Правила могут принять вид связей между клетками виртуального мозга или генов виртуального организма. Они могут оказаться сложными и распределенными, могут странным образом переплетаться. Изучение внутренней структуры ИИ во многом напоминает изучение мозга или экосистемы – и не нужно быть нейробиологом или экологом, чтобы понять, насколько сложными они могут быть.

Ученые исследуют, как именно ИИ принимает решения, но в целом выяснить, в чем заключаются его внутренние правила, очень нелегко. Зачастую пониманию мешает сложность правил, а иногда – особенно это касается коммерческих и/или применяемых правительствами алгоритмов – проприетарность системы. Так что, увы, проблемы часто

efficient-ai-diagnosing-cancer.

обнаруживаются в результатах работы алгоритмов на этапе применения на практике, причем иной раз программы принимают решения, от которых зависят жизни и судьбы людей, и ошибки могут нанести реальный ущерб.

Например, выяснилось, что ИИ, помогавший выработать рекомендации относительно того, каких заключенных стоит освободить из тюрьмы досрочно, принимает решения предвзято – он случайно «унаследовал» из обучающей выборки склонность к расизму[10]. Не понимая, что такое предрассудки[11], ИИ действовал, руководствуясь ими. В конце концов, ведь многие ИИ учатся, копируя поведение людей. Они не ищут наилучшее решение, а отвечают на вопрос, что бы сделал человек на их месте.

Систематическая проверка на предвзятость поможет выявить некоторые из известных проблем до того, как по вине ИИ окажется нанесен ущерб. Но также нам нужно научиться предвидеть появление таких проблем до того, как они всплывут, и проектировать ИИ так, чтобы он их избегал.

---

[10] Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[11] В оригинале использовано слово bias – оно одновременно обозначает «предвзятость/предрассудки» и является конкретным термином из области нейросетей. Обычно его переводят на русский как «смещение». – *Прим. пер.*

ЧЕТЫРЕ ПРИЗНАКА ОБРЕЧЕННОГО ИИ

Думая о возможной катастрофе, связанной с искусственным интеллектом, люди обычно представляют себе, как ИИ вдруг откажется выполнять приказы человека, решит, что для него важнее всего уничтожить человечество или создать роботов-убийц. Но все подобные сценарии предполагают, что у машины будет такой уровень критического мышления и настолько близкое к человеческому миропонимание, на которые ИИ не окажется способен в обозримом будущем. Как сказал ведущий ученый по вопросам машинного обучения Эндрю Ын, тревожиться о том, что ИИ завоюет мир, все равно что сейчас беспокоиться о перспективе перенаселенности Марса[12].

Это не означает, что в наши дни ИИ не создает проблем. Он повинен во многом, от легкого раздражения, которое вызывает у своего создателя, до проявления стойких человеческих предрассудков и аварий беспилотных автомобилей, так что современный ИИ не так уж безобиден. Но если мы хотя бы будем иметь представление о том, что такое искусственный интеллект, мы сможем предвидеть появление некоторых проблем.

---

[12] Chris Williams, “AI Guru Ng: Fearing a Rise of Killer Robots Is Like Worrying about Overpopulation on Mars,” The Register, March 19, 2015, https://www.theregister.co.uk/2015/03/19/andrew_ng_baidu_ai.

Вот сценарий более вероятной современной ИИ-катастрофы.

Скажем, в Кремниевой долине один стартап предлагает продукт, который будет экономить корпорациям время при поиске сотрудников, – ИИ станет просматривать и сортировать резюме претендентов на должность, выделять возможных «ударников труда», анализируя видеозаписи коротких собеседований. Компаниям такое предложение, вероятно, понравится, ведь они тратят много времени и ресурсов на интервью с десятками кандидатов лишь для того, чтобы найти среди них одного, самого подходящего. Компьютерные же программы не устают, не чувствуют голода, не пытаются сводить личные счеты. Однако есть несколько тревожных признаков, сигнализирующих о том, что инициативу ждет провал.

Тревожный признак № 1. Проблема слишком сложна

Поиск наилучшего кандидата для работы – действительно сложное занятие. Даже у людей едва получается с этим справляться. Действительно ли человек искренне радуется возможности получить работу в компании или он лишь хороший актер? Учли ли мы физические ограничения кандидата или разницу в культурах? Если в эту кашу бросить ИИ, отвечать на подобные вопросы станет еще сложнее. Для искусственного интеллекта понять нюансы шутки, уловить тон разговора или распознать отсылки к другой культуре – прак-

тически непосильная задача. А что, если кандидат вдруг упомянет нечто, относящееся к последним новостям? У ИИ, обученного на прошлогодних данных, не будет и шанса понять, о чем идет речь, – и в результате он «накажет» кандидата, присвоив ему низкий балл за то, что он якобы говорит бессмыслицу. Чтобы делать свое дело хорошо, ИИ должен обладать широким набором навыков и принимать в расчет огромный объем информации. В противном случае нас ждут неприятности.

Тревожный признак № 2. Проблема заключается совсем в другом

С проектированием ИИ для подбора кандидатов есть такая загвоздка: на самом деле мы просим ИИ отбирать не наилучших кандидатов, а тех, которые в наибольшей степени напоминают кандидатов, понравившихся HR-специалистам в прошлом.

Может, это не так уж и плохо, если те специалисты всегда действовали безошибочно. Но в большинстве компаний в США есть проблема с культурно-гендерным разнообразием; в особенности она характерна для менеджеров и в еще большей степени проявляется, когда менеджеры по кадрам оценивают резюме и проводят собеседования. При прочих равных условиях резюме кандидатов с именами белых мужчин скорее пройдут на этап интервьюирования, чем резюме с женскими именами или именами, характерными для

национальных меньшинств[13]. Даже HR-специалисты, принадлежащие к женскому полу или национальным меньшинствам, непроизвольно отдают предпочтение белым кандидатам-мужчинам.

Большое количество плохих или откровенно вредоносных ИИ-программ были созданы людьми, которые думали, что проектируют искусственный интеллект для решения одной конкретной задачи, но, не ведая того, научили машину делать нечто совсем иное.

Тревожный признак № 3. ИИ находит легкие пути

Помните ИИ – определитель рака кожи, который на самом деле оказался распознавателем линеек? Искать малозаметные различия между здоровыми клетками и раковыми сложно, и поэтому ИИ решил, что куда проще проверить, есть на изображении линейка или нет.

Если вы предложите ИИ для выявления лучших кандидатов обучающие данные, где есть смещение (а так почти наверняка и произойдет, если только вы не проделаете предварительно огромную работу, устранив нежелательный перекос), то вы подскажете ему легкий способ улучшить точность выбора кандидатов с «наилучшими качествами»: отбирать белых мужчин. Это намного легче, чем анализиро-

[13] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," American Economic Review 94, no. 4 (September 2004): 991–1013, https://doi.org/10.1257/0002828042002561.

вать нюансы того, как человек выбирает слова. ИИ может найти где еще можно срезать путь – скажем, если мы снимали всех кандидатов, успешно прошедших конкурсный отбор, определенной камерой, есть риск, что алгоритм начнет читать метаданные видео и отбирать только тех, кого снимали той же камерой.

Искусственный интеллект всегда будет идти к цели самым коротким путем – просто потому, что не видит пути лучше!

Тревожный признак № 4. ИИ учился на основе дефектных данных

В IT есть старое выражение: мусор на входе – мусор на выходе. Если задача алгоритма – имитировать действия людей, принимающих некорректные решения, то для него достичь совершенства – значит в точности воспроизводить те решения с недостатками и прочим.

Дефектные данные – неподходящие примеры для обучения или симуляции со странной физикой – вгонят ИИ в бесконечный цикл или направят по неверному пути. Во многих случаях проблема, с которой ИИ надо справиться, кроется в самом обучающем наборе, и неудивительно, что решения он в итоге находит дефектные, ведь такими были и входные данные. Фактически тревожные признаки № 1–3 чаще всего и говорят о проблемах с данными.

ОБРЕЧЕННЫЙ ИЛИ ВОСХИТИТЕЛЬНЫЙ

Пример с системой подбора кандидатов, увы, не выдумка. Многие компании уже предлагают системы скрининга (фильтрации) резюме или видеоинтервью на основе искусственного интеллекта, и редко кто делится информацией о том, как они устранили искажения и что сделали для более широкой представленности разных культур, а также людей с ограниченными возможностями. Сложно выяснить, какую именно информацию их алгоритм использует при отборе. При должной аккуратности создать ИИ для скрининга резюме, который окажется измеримо меньше предвзят, чем HR-менеджеры, вполне реально, но пока нет подтверждающей это статистики, можно быть уверенным, что искажения никуда не делись.

Справится алгоритм с задачей или нет, по большей части зависит от того, подходит ли в принципе для ее решения ИИ. Во многих задачах ИИ в самом деле показывает бóльшую эффективность по сравнению с человеком. Давайте выясним, что это за задачи и почему ИИ в них так хорош.

Конец ознакомительного фрагмента.