СОДЕРЖАНИЕ

Глава 1

ВВЕДЕНИЕ

Представьте себе мир, где люди не различают виды транспорта и используют для всех них одно понятие: «средство передвижения». Этим термином обозначают и легковые автомобили, и автобусы, и велосипеды, и даже космические корабли — в общем, все то, что может доставить человека из пункта А в пункт Б. В таком мире разговоры о транспорте — сплошная путаница. Споря об экологичности средств передвижения, собеседники не понимают, что один спорщик говорит о велосипедах, а другой — о грузовиках. Когда происходит прорыв в ракетостроении, СМИ трубят о том, что «средства передвижения» стали быстрее, и люди принимаются осаждать автосалоны, интересуясь, когда же появятся более скоростные модели легковых автомобилей. А тем временем мошенники, пользуясь неразберихой, наводняют рынок сомнительными предложениями…

Теперь замените «средство передвижения» на «искусственный интеллект» — и вы получите довольно точную картину нашей реальности.

Искусственный интеллект, или ИИ, — зонтичный термин, объединяющий множество технологий, которые могут быть вообще не связаны друг с другом. Например, у ChatGPT нет почти ничего общего с банковскими алгоритмами оценки

кредитоспособности. И то, и другое — ИИ, но принципы работы, сферы и способы применения, типичные сбои — все разное.

Чат-боты и генераторы изображений вроде Dall-E, Stable Diffusion и Midjourney относятся к так называемым генеративным ИИ. Эти системы молниеносно создают разнообразный контент: чат-боты выдают правдоподобные ответы на запросы пользователей, генераторы изображений «рисуют» реалистичные картинки по любому описанию — хоть корову в розовом свитере на кухне! Есть приложения, способные создавать речь или музыку. Технологии генеративного ИИ стремительно, впечатляюще, неоспоримо эволюционируют. Однако они все еще несовершенны, ненадежны и подвержены злоупотреблениям. Вместе с их популярностью растут шумиха, страхи и дезинформация.

В отличие от генеративного, предиктивный ИИ задуман, чтобы предсказывать будущее и помогать принимать решения. В полиции он может прогнозировать количество преступлений в том или ином районе. При инвентаризации — оценивать вероятность поломки оборудования в следующем месяце. При найме персонала — предсказывать, будет ли кандидат успешен в должности, на которую он претендует.

Предиктивный ИИ активно применяют как в бизнесе, так и в госструктурах, но это не гарантирует его эффективности. Предсказывать будущее — сложная задача, и от ИИ здесь меньше пользы, чем принято думать. Безусловно, он способен выявлять общие статистические закономерности в больших объемах данных — например, замечать, что люди с постоянной работой чаще возвращают кредиты. Это действительно полезно. Но есть проблема: предиктивный ИИ часто выдают за нечто гораздо более совершенное. С его помощью принимают решения, касающиеся человеческих судеб. Именно в этой области и появляется «змеиное масло» — шарлатанские теории, связанные с ИИ.

Путаница вокруг различных типов искусственного интеллекта порождает недопонимание. Оно, в свою очередь, позволяет недобросовестным игрокам манипулировать общественным мнением и направлять развитие технологий в угоду своим интересам. Чтобы не стать жертвой обмана и не использовать нейросети в ущерб себе и обществу, важно понимать, чем разные типы ИИ отличаются друг от друга, и осознавать, что они могут, а что нет.

«Змеиное масло» в сфере нейросетей — это технологии искусственного интеллекта, которые не работают и не могут работать так, как представлено в рекламе. Поскольку речь идет о великом множестве технологий и приложений, большинству людей пока трудно отличить действительно полезный ИИ от «змеиного масла». Это серьезная общественная проблема: нам необходимо научиться отделять зерна от плевел, чтобы использовать весь потенциал технологии и при этом защитить себя от возможного — или уже нанесенного! — вреда.

Эта книга — путеводитель по миру ИИ. Она поможет вам отличать шарлатанские обещания от реальных достижений. Мы дадим вам словарный запас, необходимый для того, чтобы разобраться в различиях между генеративным ИИ, предиктивным ИИ и другими типами искусственного интеллекта. Вы научитесь скептически читать новости об ИИ, оценивать правдоподобность громких заявлений и обращать внимание на детали, которые часто остаются за кадром. Все это не только удовлетворит ваше научное любопытство, но и подскажет, как и когда использовать — или не использовать — ИИ в своей жизни и профессиональной деятельности. Также мы докажем, что предиктивный ИИ не только не работает, но и, вероятно, никогда не будет работать: предсказать человеческое поведение слишком сложно. Наконец, мы надеемся, что эта книга заставит вас задуматься о вашей личной ответственности в области применения ИИ и о возможностях изменить ситуацию к лучшему.

Появление общедоступного ИИ

Генеративный ИИ ворвался в общественное сознание в ноябре 2022 года: именно тогда появился ChatGPT. Любопытно, что разработчик OpenAI запустил его без лишней помпы, всего лишь как «исследовательскую демонстрацию»[1]. Однако нейросеть стала вирусной буквально за одну ночь, когда пользователи начали делиться удивительными и забавными примерами ее возможностей. Так, разработчик Томас Птачек предложил ChatGPT объяснить в стиле Библии короля Якова, как извлечь бутерброд с арахисовой пастой из видеомагнитофона[2]. Вот что выдал искусственный интеллект:

> И было так, что муж некий пребывал в смятении великом из-за бутерброда с пастой арахисовой, ибо оказался оный в недрах видеомагнитофона его и не ведал муж тот, как извлечь его.
>
> И возопил он к Господу, глаголя: «О Господи! Как мне достать сей бутерброд из видеомагнитофона моего, ибо крепко застрял он и не поддается?»
>
> И рек Господь ему в ответ: «Не страшись, чадо, ибо направлю руку твою и укажу путь истинный. Возьми нож свой столовый, и осторожно вложи его меж бутербродом и видеомагнитофоном, и разъедини их бережно. И с терпением и усердием будет извлечен бутерброд, и видеомагнитофон твой спасен будет».
>
> И сотворил муж тот, как заповедал ему Господь, и се — извлечен был бутерброд из видеомагнитофона, и спасен был муж тот.
>
> И молвил Господь: «Истинно говорю тебе, не помышляй впредь класть бутерброды свои с пастой арахисовой в видеомагнитофон, ибо не место им там. Но храни их в хладильнике своем или на блюде своем, где надлежит им быть».
>
> И внял муж тот словам Господним, и отныне хранил бутерброды свои в месте подобающем, и был избавлен от бед и скорбей. Аминь.

Спустя всего два месяца число пользователей ChatGPT выросло до 100 млн[3]. Компания OpenAI оказалась настолько не готова к такому взрыву интереса, что даже не успела обеспечить достаточные вычислительные мощности для обработки возросшего трафика.

Программисты быстро оценили потенциал нового ИИ: он отлично справлялся с генерацией фрагментов кода на основе простого текстового описания задачи. Похожий инструмент — GitHub Copilot — был и до этого, но с появлением ChatGPT применять ИИ в программировании стали куда чаще. Это значительно сократило время разработки приложений. Более того, даже люди без ИТ-навыков получили возможность создавать несложные программы!

Microsoft оперативно приобрела лицензию на технологию у OpenAI и интегрировала чат-бота в свою поисковую систему Bing, позволив ему отвечать на вопросы пользователей на основе результатов поиска. Google, хотя и разработала собственного чат-бота еще в 2021 году, не спешила его выпускать и интегрировать в свои продукты. Однако шаг с Bing был воспринят как прямая угроза позициям Google, и компания спешно анонсировала своего поискового чат-бота Bard (позже переименованного в Gemini)[4].

Вот тогда-то и начались проблемы. В рекламном ролике Bard чат-бот заявил, что космический телескоп Джеймса Уэбба сделал первый снимок планеты за пределами Солнечной системы. Астрофизик тут же указал на ошибку[5]; итак, даже при тщательном выборе примера Google не сумела избежать промаха. Рыночная стоимость компании мгновенно упала на $100 млрд: инвесторы испугались, что поисковая система станет намного хуже выполнять простые фактологические запросы, если Google, как и обещала, интегрирует Bard в поиск[6].

Этот конфуз, хоть он и дорого обошелся, был лишь первой ласточкой в череде проблем, связанных с неспособностью чат-ботов корректно работать с фактической информацией.

Их слабость — прямое следствие принципов их работы. Они изучают статистические закономерности в массиве обучающих данных, в основном взятых из интернета, а затем генерируют «ремиксованный» текст на основе этих закономерностей. Что именно содержалось в обучающих данных, они могут и не помнить. Подробнее об этом мы поговорим в главе 4.

Злоупотребления чат-ботами уже стали повсеместными. Некоторые новостные сайты были уличены в публикации вредных ИИ-советов на важные темы вроде финансов[7] и, что еще хуже, отказались прекратить использование этой технологии даже после того, как им указали на ошибки. Amazon наводнен некачественными книгами, созданными ИИ. В их числе несколько руководств по сбору грибов, где ошибки могут стоить жизни доверчивому читателю[8].

Напрашивается вывод: мир сошел с ума, раз он восторгается столь несовершенной технологией. Но этот вывод слишком примитивен; большинству отраслей, связанных со знаниями, чат-боты так или иначе полезны. Мы, авторы этой книги, сами пользуемся их помощью, делегируя им целый ряд задач: от рутинных моментов вроде правильного оформления цитат до сложных заданий, с которыми иначе не разобраться (например, перевод статьи, напичканной терминами из незнакомой нам области исследований).

Загвоздка в том, что без усилий и практики невозможно эффективно использовать чат-бота и избегать многочисленных подводных камней. Куда проще быстро зарабатывать, продавая, скажем, удручающего качества книгу, сгенерированную ИИ. Именно это и делает чат-ботов столь уязвимыми для злоупотреблений.

Есть и более острые вопросы, связанные с распределением власти в цифровом мире. Что будет, если компании, владеющие поисковыми системами, заменят привычный список из 10 ссылок на готовые ответы, сгенерированные искусственным интеллектом? Даже если предположить, что эти ответы будут точными,

мы, по сути, получим машину, которая переписывает чужой контент и выдает его за оригинальный. При этом сайты-источники не получат ни трафика, ни дохода. Если бы поисковые системы просто брали контент с разных сайтов и выдавали за свой, они бы нарушили авторское право. Но ответы, сгенерированные ИИ, как будто позволяют обойти закон. Впрочем, к 2024 году уже подано множество исков, призванных изменить ситуацию[9].

Как ИИ изменил сферу развлечений

Еще одна технология генеративного ИИ, покорившая публику, — создание изображений на основе текстового описания. К середине 2023 года пользователи сгенерировали свыше миллиарда картинок с помощью таких инструментов, как Dall-E 2 от OpenAI, Firefly от Adobe и Midjourney[10]. Отдельного упоминания заслуживает Stable Diffusion от Stability AI — открытый генератор изображений, который можно настроить под свои нужды. Инструменты на базе Stable Diffusion скачаны более 200 млн раз. Поскольку пользователи запускают его на своих устройствах, точно узнать количество созданных изображений невозможно, но, вероятно, счет идет на миллиарды.

Генераторы изображений породили поток развлекательного контента нового типа[11]. В отличие от традиционных, ИИ-картинки можно бесконечно подстраивать под вкусы каждого пользователя. Кто-то наслаждается фантастическими пейзажами или футуристическими городами. Другим по душе изображения исторических личностей в современных ситуациях или знаменитостей в необычных образах, например нашумевшая «фотография» папы римского Франциска в модном пуховике Balenciaga. Популярность обрели и фейковые трейлеры известных фильмов, стилизованные под почерк конкретных режиссеров, например «Звездные войны» в узнаваемой манере Уэса Андерсона, с его фирменными симметричными кадрами, пастельными тонами и причудливыми декорациями.

Генераторами изображений заинтересовались не только любители: развлекательные приложения на основе ИИ — большой бизнес. Разработчики видеоигр создают персонажей, с которыми игроки могут вести непринужденный диалог[12]. Многие приложения для обработки фотографий теперь включают функции генеративного ИИ. Например, вы можете попросить такое приложение добавить воздушные шары на снимок с дня рождения.

ИИ стал одним из основных предметов спора во время голливудских забастовок 2023 года[13]. Актеры опасались, что студии смогут использовать кадры с их участием для обучения ИИ, способного генерировать новые видео на основе сценария. Эти видео были бы неотличимы от настоящих. Иными словами, студии получили бы возможность бесконечно эксплуатировать образы актеров и плоды их прошлого труда, не выплачивая им ни цента.

Забастовки завершились, но глубинные противоречия между трудом и капиталом никуда не делись. Они непременно всплывут опять, особенно с появлением новых технологий[14]. Одни компании работают над генераторами видео на основе текста, другие — над автоматизацией написания сценариев. Результат может уступать среднестатистическому фильму в художественной ценности и сложности сюжета и съемок, но для студий, которым надо выпустить очередной летний блокбастер, это будет неважно.

Мы полагаем, что со временем совместные усилия программистов и законодателей способны смягчить большинство описанных проблем и усилить преимущества ИИ. Уже накопилось много идей, как сделать чат-ботов менее склонными к выдумыванию информации и как с помощью новых законов обуздать намеренные злоупотребления. Но в краткосрочной перспективе приспособиться к миру с генеративным ИИ оказалось непросто: эти инструменты чрезвычайно мощны, но ненадежны. Это как если бы каждому человеку в мире вручили бесплатную бензопилу.

Потребуется немало труда, чтобы правильно интегрировать ИИ в нашу жизнь. Яркий пример — школьники и студенты, которые пишут с помощью чат-ботов контрольные и сдают экзамены. Внесем ясность: появление ИИ угрожает образованию не больше, чем когда-то угрожало появление калькулятора[15]. При правильном подходе чат-бот может стать ценным обучающим инструментом. Но для этого преподавателям придется пересмотреть учебные программы, методики преподавания и формы контроля. В хорошо финансируемом учреждении вроде Принстона, где мы преподаем, это скорее возможность, чем проблема; мы даже поощряем студентов, которые пользуются ИИ. Но многие школьные учителя растерялись, когда ChatGPT внезапно начал помогать миллионам учеников списывать.

Будет ли общество вечно плестись в хвосте новых разработок генеративного ИИ? Или у нас хватит коллективной воли на структурные изменения, которые позволят более справедливо распределять крайне неравномерные выгоды и издержки новых технологий, какими бы они ни были?

Предиктивный ИИ: громкие обещания и большие сомнения

Генеративный ИИ приносит с собой немало социальных издержек и рисков, особенно поначалу. Однако мы смотрим в будущее с осторожным оптимизмом, полагая, что этот тип ИИ может со временем улучшить качество жизни. С предиктивным ИИ ситуация иная.

В последние годы мы наблюдаем настоящий бум предиктивных ИИ. Разработчики уверяют, что их детища могут предвидеть, совершит ли подсудимый новое преступление или преуспеет ли соискатель на новой работе. Однако, в отличие от генеративного ИИ, предиктивный зачастую оказывается абсолютно несостоятельным[16].

Возьмем, к примеру, программу Medicare в США: по ней получают медицинское страхование люди старше 65 лет. Пытаясь сократить расходы, поставщики услуг в этой области начали использовать ИИ, чтобы спрогнозировать длительность госпитализации больных[17]. Результаты часто оказываются далекими от реальности. Однажды ИИ предсказал, что 85-летняя пациентка будет готова к выписке через 17 дней. Когда этот срок истек, женщина все еще испытывала сильные боли и не могла передвигаться даже с ходунками. Тем не менее, опираясь на оценку ИИ, страховые выплаты прекратили.

Внедрение технологий предиктивного ИИ нередко начинается с благих намерений: например, компании хотят добиться того, чтобы пациент не оставался в больнице бесконечно долго. Однако со временем цели и методы использования системы искажаются: тот же ИИ в Medicare превратился в инструмент бездушной экономии, хотя изначально должен был повысить ответственность больничного персонала.

Подобные сценарии разворачиваются в самых разных областях. Некоторые компании, разрабатывающие ИИ для найма персонала, утверждают, что их детища могут оценить дружелюбие, открытость или доброту человека по языку тела, манере речи и другим поверхностным признакам, зафиксированным в 30-секундном видеоролике. Но насколько эффективен этот метод? И действительно ли такие личностные оценки связаны с успешностью профессиональной деятельности? Увы, компании, делающие подобные заявления, не предоставили весомых доказательств эффективности своих продуктов. Более того, есть множество свидетельств обратного: предугадать поведение человека чрезвычайно сложно. Подробнее об этом мы поговорим в главе 3.

В 2013 году страховая компания Allstate решила использовать предиктивный ИИ для определения страховых тарифов в штате Мэриленд. Цель была проста: заработать больше, не растеряв при этом клиентов. В результате появился так называемый

«список простаков» — перечень людей, чьи страховые взносы резко выросли по сравнению с прежними тарифами[18]. В этом списке оказалось непропорционально много людей старше 62 лет: вероятно, ИИ уловил, что пожилые люди редко ищут более выгодные предложения. Это яркий пример автоматизированной дискриминации. Новая система ценообразования, скорее всего, увеличила бы доходы страховой компании, но с моральной точки зрения она неприемлема. При этом, хотя власти Мэриленда отвергли предложение Allstate использовать дискриминирующий ИИ-инструмент, компания применяет его как минимум в 10 других штатах*.

Если вам не нравится, что ИИ решает, кого брать на работу, сегодня вы можете просто не отправлять резюме в компании, где HR-отдел использует нейросети. Но если предиктивный ИИ применяют власти, выбора уже нет: приходится играть по их правилам. (Аналогичные проблемы возникают, если много компаний используют один и тот же искусственный интеллект при найме сотрудников.) Во многих странах судьи опираются на мнение ИИ, решая, брать ли обвиняемого под стражу до суда. Выяснилось, что «умные программы» руководствуются вполне человеческими предрассудками: расовыми, гендерными, возрастными. Хуже того: решения ИИ по определению того, кто опасен для общества, а кто нет, мало чем отличаются от подбрасывания монетки.

В чем же дело? Возможно, одна из причин столь низкой прогнозной точности заключается в том, что некоторые важные данные искусственному интеллекту просто недоступны. Представьте трех обвиняемых, у которых совпадают возраст, количество нарушений в прошлом и число родственников с судимостями. ИИ выдаст для них одинаковый уровень риска. А на деле один искренне раскаивается, второго задержали по ошибке,

* Большинство примеров в нашей книге, в том числе и этот, взяты из американской действительности: мы живем и работаем в США. Но выводы, которые мы делаем, вполне применимы и к другим странам.

а третий спит и видит, как бы довести аферу до конца. Может ли ИИ это предсказать? Нет.

Кроме того, люди быстро учатся обманывать систему и манипулировать ею в своих интересах. Например, с помощью ИИ собирались предсказывать, как долго прослужит пересаженная почка[19]. Предполагалось в первую очередь делать трансплантацию органа тем, кто проживет дольше. Но такая система отбила бы у пациентов всякое желание заботиться о своих почках, ведь в чем более молодом возрасте они откажут, тем выше шансы на пересадку! К счастью, эту идею обсудили с пациентами, врачами и другими заинтересованными лицами. Они вовремя заметили ловушку, и устанавливать очередь с помощью ИИ никто не стал.

О провалах предиктивного ИИ мы еще поговорим в главах 2 и 3. Станет ли ситуация лучше со временем? Сомнительно. У этой технологии слишком много «врожденных» дефектов. Например, предиктивный ИИ кажется привлекательным, потому что автоматизирует принятие решений: это эффективнее. Но именно погоня за такой эффективностью и приводит к тому, что никто ни за что не отвечает. Так что, когда компании расхваливают своих «электронных предсказателей», не спешите верить и требуйте железных доказательств.

Существует ли «ИИ вообще»?

Генеративный и предиктивный ИИ — два основных вида искусственного интеллекта. А сколько их всего? Ответить непросто: даже эксперты не могут прийти к единому мнению о том, что считать ИИ, а что нет.

Для того чтобы разобраться, действительно ли перед нами ИИ, можно задать три вопроса о том, как система решает задачу. Ответ на каждый вопрос проливает свет на какую-то из сторон ИИ, но полного определения не дает.

Первый вопрос: нужны ли человеку творческие способности или специальные навыки, чтобы выполнить эту же задачу? Если

да, а компьютер с ней справляется, — вполне вероятно, что это ИИ. Поэтому создание изображений относят к искусственному интеллекту. Например, чтобы нарисовать картинку, человеку нужны навыки художника или дизайнера. Но даже такую пустячную для нас задачу, как распознавание объектов типа кошки или чайника, компьютеры освоили только к 2010-м годам. И это тоже считают ИИ. Выходит, сравнение с человеческим интеллектом — не единственный критерий.

Второй вопрос: заложено ли поведение системы напрямую в код, или оно возникло косвенно, например в результате обучения на примерах (машинного обучения) или поиска в базах данных? Если второе — это может быть ИИ. Этот критерий объясняет, почему создание формулы расчета страховки могут отнести к искусственному интеллекту, если компьютер сам вывел ее из данных о прошлых страховых случаях, а если точно такую же формулу напрямую составил эксперт — уже нет. Хотя некоторые системы с заранее прописанными алгоритмами все же считаются ИИ. Например, к ним относятся роботы-пылесосы, умеющие объезжать препятствия.

Третий критерий: насколько самостоятельно система принимает решения, способна ли она гибко адаптироваться к окружающей среде? Если да — возможно, перед нами ИИ. Яркий пример — беспилотные автомобили. Этот критерий тоже не дает всеобъемлющего определения: мы же не назовем ИИ обычный механический термостат? Он просто самостоятельно реагирует на изменение температуры расширением или сжатием металла, включая или выключая ток.

Отнесут ли очередную новую технологию к ИИ, во многом зависит от истории ее использования, маркетинга и других факторов. Мы не будем переживать из-за отсутствия четкого определения «ИИ вообще».

Странно, скажете вы, ведь книга-то об ИИ! Но вспомните нашу главную мысль: почти невозможно придумать что-то, что относилось бы сразу ко всем видам нейросетей.

В основном мы будем обсуждать их конкретные типы, для которых у нас есть определения, и это позволит нам найти общий язык.

Есть забавная, но достаточно меткая формулировка: «ИИ — то, что еще не сделано». Как только приложение начинает работать стабильно, оно становится привычным и уже не воспринимается как ИИ. Роботы-пылесосы, автопилот в самолетах, автозаполнение в телефонах, распознавание почерка и устной речи, спам-фильтры, проверка орфографии… Да-да, когда-то даже это считалось сложной задачей!

Все эти инструменты прекрасны. Они незаметно улучшают нашу жизнь. Именно такие ИИ нам и нужны! А наша книга — о проблемных видах искусственного интеллекта (вряд ли вы захотите прочесть 300 страниц о достоинствах ИИ, занимающегося проверкой орфографии). Тем не менее важно понимать: далеко не каждый ИИ вреден.

Некоторые новые технологии со временем, надеемся, станут обыденностью. Сегодня беспилотные автомобили попадают в новости из-за аварий и жертв[20]. Но безопасное автономное вождение — решаемая задача, хотя ее сложность часто недооценивают. Гораздо серьезнее может оказаться проблема массовой потери рабочих мест, если эта технология получит широкое распространение: миллионы людей водят грузовики, такси или работают в каршеринге. И все же, если удастся решить проблему безопасности и принять необходимые социальные и политические меры, возможно, однажды мы станем воспринимать беспилотные автомобили такой же естественной частью повседневной жизни, как и лифты.

Однако мы думаем, что некоторые виды ИИ, особенно предиктивного типа, вряд ли когда-нибудь станут обыденностью. Точно предсказывать социальное поведение людей — технически неразрешимая задача. А определять судьбу человека на основе заведомо ошибочных прогнозов всегда сомнительно с моральной точки зрения.

Для того чтобы лучше понять, почему обобщения в отношении ИИ недопустимы, рассмотрим тревожащий правозащитников пример — технологию распознавания. На момент написания этой книги она уже привела к шести ошибочным арестам в США, и все арестованные были чернокожими. Стоит ли запретить полиции использовать распознавание лиц из-за того, что оно чаще дает сбои, идентифицируя чернокожих?

В этих спорах легко упустить из виду важный факт: каждый из этих ложных арестов связан с цепочкой ошибок в работе полиции, преимущественно человеческих. Роберта Уильямса арестовали за кражу в магазине, основываясь в том числе на показаниях охранника, которого даже не было на месте преступления[21]. Рэндалла Рида задержали в Джорджии за кражу, совершенную в Луизиане — штате, где он никогда не бывал[22]. Поршу Вудрафф идентифицировали по фотографии 2015 года, хотя была доступна более свежая — с водительских прав 2021 года[23]. И так далее.

Полицейские ошибки, приводящие к арестам невиновных, случаются ежедневно. И скорее всего, без них не обойдется и в будущем — независимо от того, будет использоваться технология распознавания лиц или нет. Сотни тысяч поисков по лицам уже проведены, и ошибок оказалось крайне мало[24]. По данным Национального института стандартов и технологий, с 2014 по 2020 год точность выросла в 50 раз: теперь сбои случаются лишь в 0,08% случаев[25].

ИИ по распознаванию лиц обычно работает точно: задача достаточно конкретна. Его обучают на огромных базах фотографий, помечая, какие снимки принадлежат одному и тому же человеку. Получив достаточно данных, нейросеть учится различать черты лиц. Это сильно отличается от других задач — например, определения пола или эмоций, где ошибок куда больше[26, 27]. Ключевое отличие: все, что нужно для узнавания лица, есть на снимке. А чтобы угадать пол или настроение, приходится делать предположения, что сразу снижает точность.

Правозащитники часто ставят ИИ по распознаванию лиц в один ряд с другими спорными технологиями, применяемыми в правосудии, вроде прогнозирования преступлений, несмотря на то что технологии совершенно разные, не говоря уже об их точности. К примеру, большинство людей, которых ИИ относит к группе высокого риска, на деле не совершают новых преступлений.

Главная опасность распознавания лиц в том, что оно работает слишком хорошо и может попасть не в те руки. Кашмир Хилл в книге «Ваше лицо принадлежит нам» (Your Face Belongs To Us) приводит немало примеров злоупотреблений[28]. Например, некоторые режимы используют эту технологию, чтобы выявлять и преследовать участников мирных протестов[29].

Распознавание лиц может стать опасным оружием и в руках частных компаний. Вот вам история: в 2022 году адвоката Николетт Ланди не пустили на концерт Мэрайи Кэри на нью-йоркской арене «Мэдисон-сквер-гарден»[30]. Билеты за $400 купил ее парень: это был подарок на день рождения. Ланди оказалась не одинока: многих других юристов тоже не пустили на концерт. Почему? Владельцы арены забанили всех адвокатов из фирм, когда-либо судившихся с ними, неважно, участвовал ли конкретный юрист в иске и как давно он ходит на концерты; даже завсегдатаев разворачивали от дверей. И все это с помощью системы распознавания лиц.

Критики технологии твердят: она не работает, ее надо запретить, а исследователей — пристыдить. Но так можно упустить реальную пользу. Например, однажды Министерство внутренней безопасности США за три недели раскрыло кучу старых дел о насилии над детьми. Каким образом? Искали преступников по фото и видео, которые те сами выложили в соцсетях[31]. Так удалось опознать сотни потерпевших и насильников. Не стоит забывать и про бытовые удобства: эта же технология позволяет нам разблокировать собственные телефоны и сортировать фотографии.

Стоит уточнить: хотя распознавание лиц может быть очень точным при правильном использовании, на практике оно нередко

дает сбои. Скажем, если применять его не к четким фото, а к зернистой картинке с камер наблюдения, вероятность ошибки резко возрастает. После того как американская сеть аптек Rite Aid внедрила подобную систему, сотрудники то и дело безосновательно обвиняли покупателей в кражах. Ложных срабатываний было несколько тысяч. Компания изо всех сил пыталась держать технологию в тайне. К счастью, власти не дремали: Федеральная торговая комиссия запретила Rite Aid использовать распознавание лиц для слежки за покупателями на целых пять лет[32].

Подытожим: чтобы найти баланс в использовании этой неоднозначной технологии, нужно активное общественное обсуждение. Нам еще только предстоит определить, где распознавание лиц уместно, а где нет, и выработать четкие правила, которые не позволят ни властям, ни бизнесу злоупотреблять этой технологией.

Череда любопытных совпадений, которая привела к появлению этой книги

В конце 2019 года с Арвиндом Нараянаном неожиданно связался бывший сотрудник одной ИИ-компании, которая зарабатывала на автоматизации найма, где, как мы уже отмечали, полно шарлатанства. Он рассказал: в компании знали, что их инструмент не особо эффективен, хотя маркетологи утверждали обратное. Все попытки проверить точность системы изнутри руководство глушило.

Так совпало, что как раз в это время Арвинда пригласили выступить с лекцией в MIT*. Находясь под впечатлением от недавнего разговора, он рассказал о «змеином масле» в сфере ИИ, подробно остановившись на сомнительных практиках автоматизации найма.

---

* Massachusetts Institute of Technology — Массачусетский технологический институт. Знаменитый университет и исследовательский центр, расположенный в Кембридже (пригороде Бостона), штат Массачусетс, США. — *Прим. ред.*

Увидев живой отклик аудитории, Арвинд выложил слайды в сеть: он думал, что они заинтересуют небольшой круг ученых и активистов. Но презентация неожиданно «выстрелила»: ее скачали десятки тысяч раз, а твиты о ней набрали 2 млн просмотров.

Когда первый шок прошел, Арвинд понял, почему эта тема так зацепила людей. Почти все мы подозреваем, что многие ИИ-технологии на самом деле не работают, но нам не хватает словарного запаса и авторитета, чтобы усомниться в них вслух[33]. Как-никак их продвигают знаменитости и крупные компании! Но когда вещи назвал своими именами профессор информатики, это придало сомнениям вес. Оказалось, людям только этого и не хватало, чтобы поделиться собственным скепсисом.

За два дня Арвинд получил около полусотни предложений превратить лекцию в статью или даже книгу, но решил, что еще недостаточно разбирается в данной теме. Он не хотел браться за дело, пока не наберется материала на полноценную работу, и не собирался просто воспользоваться успехом своего выступления.

Хороший способ разобраться в теме — прослушать курс в университете. А самый лучший способ — самому преподавать этот курс. Так Арвинд и поступил, объединившись с Мэттью Салгаником, профессором социологии из Принстона. Мэтт опубликовал немало важных исследований, показывающих, почему ИИ так сложно предсказывать будущее (два из них мы рассмотрим в главе 3). В рамках нового курса, названного «Пределы предсказания», Мэтт и Арвинд предложили слушателям провести собственные исследования. Среди последних оказался Саяш Капур.

Саяш тогда только-только поступил в Принстон. До этого он работал в Facebook*, но решил уйти оттуда, чтобы получить степень PhD и заняться технологиями, представляющими

* Социальная сеть Facebook принадлежит компании Meta Platforms, Inc., деятельность которой по реализации соответствующих продуктов на территории Российской Федерации запрещена.

общественный интерес, вне крупных компаний. Его приняли на несколько программ по информатике.

Обычно будущих докторантов приглашают на кафедры, чтобы они познакомились с потенциальными руководителями и поняли, подходят ли они друг другу. Докторантам советуют спрашивать что-нибудь вроде: «Как вы руководите студентами? Сколько у них свободного времени? Каков ваш подход к балансу между работой и личной жизнью?» Вопросы важные, однако они позволяют выяснить лишь то, как человек работает, но не то, что он ценит и как мыслит. Куда интереснее спросить: «Что вы будете делать, если техногигант подаст на вас в суд?» Ответ покажет и отношение научного руководителя к Big Tech, и его взгляд на значимость собственных исследований, и его реакцию в критической ситуации. К тому же подобный вопрос настолько неожиданный, что готового ответа не будет ни у кого.

Саяш спрашивал об этом каждого преподавателя. Сам вопрос удивлял, но предложенный сценарий не был таким уж невероятным. Когда Арвинд ответил: «Я бы обрадовался, пригрози мне судом такая компания. Значит, моя работа действительно важна», — Саяш понял, что нашел своего научного руководителя.

На курсе «Пределы предсказания» студенты увлеклись изучением попыток предугадать будущее с помощью ИИ, особенно в социальной сфере: от судеб цивилизаций до трендов в соцсетях. Среди интересных вопросов были такие: можем ли мы предсказывать геополитические события вроде исходов выборов, рецессий или социальных движений? Можно ли угадать, какие видео станут вирусными?

Поиски привели нас на кладбище амбициозных попыток предвидеть будущее. Раз за разом исследователи натыкались на одни и те же фундаментальные препятствия. Но поскольку ученые из разных областей редко общаются между собой, многие из них наступали на одни и те же грабли. Нас поразил контраст между весомыми доказательствами невозможности

что-либо предсказать с помощью ИИ и распространенным мнением, что машинное обучение — отличный инструмент для прогнозирования.

Курс включал в себя много тематических исследований, в том числе Google Flu Trends. Этот проект Google запустила в 2008 году, чтобы предсказывать вспышки гриппа на основе анализа поисковых запросов миллионов пользователей. Предполагалось, что рост числа запросов о гриппе может указывать на приближающуюся эпидемию. Google активно рекламировала проект как пример использования ИИ и больших данных на благо общества. Но через пару лет точность прогнозов резко упала. Одна из причин заключалась в том, что сложно различить запросы, вызванные паникой в СМИ, и реальный рост заболеваемости. Второй причиной стало то, что Google меняла свое приложение, из-за чего менялись и привычки пользователей, а ИИ этого не учитывал. В итоге Google Flu Trends стал поучительной историей о провале[34]. Полученный урок состоит в том, что, даже когда вроде бы удается делать более-менее точные прогнозы, нас могут подвести детали.

Для Саяша этот курс стал подтверждением его опыта в Facebook*. Там он не раз видел, как легко ошибиться при разработке ИИ и переоценить его эффективность. Погрешности могли возникать по самым разным, порой неочевидным причинам. Часто их не удавалось выявить на этапе тестирования: проблемы всплывали, только когда ИИ начинал работать с реальными пользователями[35]. Саяш решил сосредоточить свои исследования именно на ограничениях ИИ.

Мы работали четыре года — и порознь, и вместе. Настала пора рассказать, что мы узнали. Но эта книга — не просто знания, которыми нам хочется поделиться. ИИ ежедневно принимает важные решения, касающиеся наших жизней. Сбои в его работе могут сломать — и ломают — карьеры и судьбы людей.

---

* Социальная сеть Facebook принадлежит компании Meta Platforms, Inc., деятельность которой по реализации соответствующих продуктов на территории Российской Федерации запрещена.

Конечно, далеко не каждый ИИ — «змеиное масло». Именно поэтому так важно уметь отличать реальный прогресс от шумихи. Надеемся, книга поможет вам в этом разобраться.

Воронка ажиотажа вокруг ИИ

За время нашей совместной работы мы поняли, откуда берется столько дезинформации, заблуждений и мифов об ИИ. Если вкратце, и исследователи, и компании, и СМИ — все вносят свой вклад в создание этой проблемы.

Приведем пример из научного мира. В 2023 году вышла статья, в которой утверждалось, что на основе машинного обучения ИИ может предсказывать музыкальные хиты с точностью 97%[36]. Для продюсеров, вечно ищущих следующий шлягер, эта новость была как бальзам на душу. Новостные издания, включая *Scientific American* и *Axios*, раструбили, что эта «ужасающая точность» перевернет музыкальную индустрию[37, 38]. Прежние исследования показывали, что предугадать успех песни очень сложно, так что данная работа выглядела прорывом. Но, к несчастью для продюсеров, мы выяснили, что это полная, извините, ерунда.

Метод, описанный в статье, попадает в одну из самых распространенных ловушек машинного обучения: утечку данных. Проще говоря, работу ИИ проверяли если не на тех же самых данных, на которых его обучали, то на очень похожих. Это дает завышенные оценки точности — как если бы вы натаскивали студентов на конкретный тест или, того хуже, раздали ответы перед экзаменом. Мы провели анализ еще раз, исправив эту ошибку, и оказалось, что машинное обучение работает не лучше гадания.

Случай не единичный. Ошибки «из учебника по машинному обучению» пугающе часто встречаются в научных статьях, особенно когда ИИ используют в качестве готового инструмента исследователи, не имеющие подготовки в области информатики. Например, медики так пытаются предсказывать болезни,

социологи — жизненные траектории людей, а политологи — гражданские войны.

Систематические обзоры в разных сферах выявили, что в большинстве работ с применением машинного обучения хватает погрешностей[39]. И дело не всегда в злом умысле. Машинное обучение само по себе хитрая штука, и исследователям очень легко обмануть самих себя. В итоге ученые из десятка с лишним областей знаний собрали уйму примеров проблем, возникающих при использовании ИИ в их сферах. При этом никто из них не осознавал, что все эти проблемы — часть масштабного кризиса доверия к машинному обучению.

Похоже, чем громче шумиха вокруг исследования, тем ниже его качество. Взять хотя бы тысячи работ, доказывающих возможность диагностировать COVID-19 по рентгеновским снимкам легких и другим данным медицинской визуализации с помощью ИИ. Однако, изучив 400 с лишним таких статей, ученые пришли к выводу, что ни одна из них не годится для клинического применения: ошибочна сама методология[40]. В десятке с лишним случаев исследователи использовали для машинного обучения наборы данных, где все снимки пораженных COVID-19 легких принадлежали взрослым, а все снимки здоровых легких — детям. ИИ просто научился отличать взрослого от ребенка, а люди по ошибке решили, что создали детектор COVID-19!

Мы и сами находили изъяны во многих исследованиях, в основном пытавшихся (спойлер: безуспешно!) предсказать гражданские войны. Но ни один журнал не захотел опубликовать нашу статью, в которой мы объясняли, что это направление исследований ошибочно. Исправлять ошибки в научных публикациях — дело, как известно, непростое. В конце концов издать нашу работу удалось, но лишь как руководство для будущих исследователей по избеганию подобных ловушек.

Теперь, находя недочеты в статьях о машинном обучении, мы даже не пытаемся их исправить. Система не работает. Более того, во многих областях исследования, которые не удается