СОДЕРЖАНИЕ

1. Введение 9

Появление общедоступного ИИ 12

Как ИИ изменил сферу развлечений 15

Предиктивный ИИ: громкие обещания и большие сомнения 17

Существует ли «ИИ вообще»? 20

Череда любопытных совпадений, которая привела к появлению этой книги 25

Воронка ажиотажа вокруг ИИ 29

Что такое «змеиное масло»? 34

Для кого эта книга 41

2. Как ошибается предиктивный ИИ 44

Предиктивный ИИ управляет судьбами 46

Хорошее предсказание еще не гарантирует хорошее решение 51

Непрозрачный ИИ — нечестная игра 53

Чрезмерная автоматизация 56

Прогнозы не о тех людях 58

Предиктивный ИИ усиливает неравенство 61

Мир без предсказаний 64

Подведем итоги 66

3. Почему ИИ не может предсказать будущее? 68

Краткая история предсказаний будущего с помощью компьютера 70

Давайте разберемся 74

Конкурс Fragile Families Challenge 77

Почему эксперимент завершился разочарованием? 81

Прогнозирование в уголовном правосудии 86

Провал предсказать непросто. А успех? 88

Лотерея мемов 93

От частного к коллективному 97

Подведем итоги. Причины ограничений прогнозирования 104

4. Долгий путь к генеративному ИИ 106

80 лет назад 112

Неудача и возрождение 114

Машины учатся видеть 117

Техническое и культурное значение ImageNet 120

Классификация и генерация изображений 124

Как ИИ присваивает чужое творчество 129

ИИ для распознавания изображений легко может превратиться в инструмент слежения 133

От изображений к тексту 135

От моделей к чат-ботам 139

Электронные пустобрехи 145

Дипфейки, мошенничество и тому подобное 147

Какой ценой даются улучшения 149

Подведем итоги 152

5. Продвинутый ИИ: угроза человечеству? 156

Что думают эксперты? 157

Лестница универсальности 162

А что на верхних ступеньках лестницы? 169

Прогресс ускоряется? 171

Мятежный ИИ? 174

Запрещать ли мощный ИИ? 178

Лучший подход — защита от конкретных угроз 181

Подведем итоги 184

6. Почему ИИ не наведет порядок в соцсетях? 186

Когда все вырвано из контекста 190

Культурная некомпетентность 195

ИИ отлично предсказывает… прошлое 201

Человеческая изобретательность против ИИ 204

Вопрос жизни и смерти 208

Поговорим о регулировании 212

Самое сложное — провести черту 217

Семь ограничений ИИ в сфере модерации контента 224

Проблема, которую они сами создали 227

Будущее модерации контента 232

7. Почему мифы об ИИ так живучи? 236

Чем шумиха вокруг ИИ отличается от ажиотажа вокруг других технологий 241

Суета вокруг ИИ: история и культура 244

Прозрачность? А зачем? 248

Кризис воспроизводимости в исследованиях ИИ 251

Как СМИ вводят нас в заблуждение 256
Авторитет — двигатель шумихи 261
Когнитивные искажения: как (не) сбиться с пути 265

8. Что дальше? 268

Кого привлекает «змеиное масло» 271
Как принять непредсказуемость 276
Регулирование: преодоление ложной дихотомии 279
Ограничения регулирования 284
ИИ и будущее профессий 286
Взросление с ИИ: мир Кая 291
Взросление с ИИ: мир Майи 296

Благодарности 302
Источники 304

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