Теплица социальных технологий

Боязнь искусственного интеллекта

Кира Федорова

https://te-st.org/2024/07/05/ai-phobias-reasons-and-arguments/

Статья обновлена 15 июля 2024

Все громче звучат голоса, призывающие замедлить разработку искусственного интеллекта, пока не будет понятно, как с его помощью решать социальные проблемы, а не создавать новые. Среди них IT-специалисты, профессора университетов, музыканты и художники. Оценки рисков обычно основаны на видении и опыте говорящих, но иногда их авторитет заставляет всерьез задуматься. Насколько резонны аргументы ИИ-скептиков и стоит ли волноваться обычным пользователям — в разборе «Теплицы».

Сторона обвинения

Часто скептики не имеют ничего против развития технологий и ИИ-ассистированного труда. Многие из них считают, что принципы использования ИИ в современной экономике приносят больше вреда, чем пользы. Их аргументы можно разделить на несколько групп:

— Конкуренция с человеком. Это классическое опасение при развитии технологий: во время промышленной революции в Англии ремесленники ломали станки, которые заменяли их на производстве. Генеративный ИИ, в том числе Chat GPT, мог бы дополнять человека на рабочем месте, а не смещать его. Опасения сотрудников об увольнении из-за автоматизации их труда усиливаются не на пустом месте. Через восемь месяцев после релиза Chat GPT число заказов на написание текстов и кода на одной из крупнейших фриланс-бирж упало на 21% по сравнению с числом заказов на ручной неавтоматизированный труд (создание и редактирование видео, работа с Excel-таблицами).

— Недобросовестное использование ИИ. Нейросети позволяют изготавливать качественные

фейки и распространять их в разы быстрее. Создатели LLM (Large Language Models) стараются пресекать такое использование своих продуктов: OpenAI стала добавлять метаданные в генерируемые DALL-E изображения, чтобы можно было узнать о происхождении картинки, Midjourney в преддверии выборов блокирует промпты о Дональде Трампе и Джо Байдене. Тем не менее, творчество моделей text-to-image, на которых были «страдающие дети Газы», неоднократно всплывали в ходе войны Израиля против Хамас.

OpenAI отмечает, что его разработками в недобросовестных целях пользуются и серьезные акторы, аффилированные с государством. Компания отчиталась об удалении аккаунтов, связанных с проведением тайных операций по оказанию влияния. Под ними OpenAI подразумевает манипулирование общественным мнением и попытки повлиять на политические вопросы. За март–май 2024 года команда обнаружила и пресекла работу пяти таких операций, две из которых связала с Россией: в ходе одной LLM использовались для отладки кода для Telegram-бота и написания политических комментариев, в ходе другой — для перевода и редактирования статей, написания постов на Facebook и тд.

Кроме масштабных дезинформационных кампаний, мошенники часто используют фейки в обычной жизни для получения денег. Сотрудника финансовой транснациональной компании в Гонконге заставили перевести более 25,6 млн долларов с помощью Zoom-дипфейка, а голосовой дипфейк заставил банковского служащего из Эмиратов перевести 35 млн.

— Слабое регулирование. Этот аргумент связан с ощущением опасности от непредсказуемого развития ИИ: ученые увлечены интересными задачами, и могут не думать о социально-экономических последствиях. Скандалы вокруг лидера отрасли OpenAI от ухода главного научного сотрудника до жесткой политики в отношении увольняющихся работников и миллиарды инвестиционных денег еще сильнее разгоняют ажиотаж вокруг этой сферы.

В частности, этот тезис поддерживают эффективные альтруисты (effective altruists). Эффективный альтруизм — философское направление и сообщество, которое основывается на поиске наиболее действенных способов сделать мир лучше и помочь другим людям. Эффективные альтруисты в ИИ, среди которых гендиректор Tesla и SpaceX Илон Маск, сооснователь криптовалюты Ethereum Виталик Бутерин и сооснователь Facebook Дастин Московиц, считают, что новые технологии могут привести к уничтожению жизни на Земле. Снижение этого риска — одно из ключевых направлений деятельности коммьюнити. Впрочем, это не мешает Маску активно участвовать в гонке разработок искусственного интеллекта.

В 2023 году более 1000 представителей технологической сферы подписали письмо, в котором просили «поставить на паузу» развитие ИИ. «Мы продолжим разрабатывать такие мощные ИИ-системы, когда будем уверены, что они будут иметь благоприятные последствия и что мы сможем справиться с потенциальными рисками и проблемами», говорилось в письме.

— Функционирование ИИ. Галлюцинации, или выдуманные примеры, которые ИИ выдает за реальные факты, могут настораживать пользователей и снижать уровень их доверия. Некоторые галлюцинации непросто распознать, особенно если человек не разбирается в сложной теме. Meta была вынуждена закрыть доступ к своей Galactica, языковой модели для ученых, всего через три дня после релиза. Пользователи жаловались, что она не умеет отличать правду от лжи и пишет статьи про медведей в космосе с такой же готовностью, как статьи про белковые комплексы. При этом, Galactica отказалась создать научный текст про СПИД, потому что этот запрос не проходил ее фильтры.

Иногда ИИ воспроизводит предубеждения и стереотипы. Разработчики стараются не допускать таких результатов, объясняя их тем, что модели являются генеративно-предиктивными, то есть, генерируют наиболее вероятный вариант событий. Вероятность означает частоту проявления свойства в обучающем датасете. В последних версиях Midjourney уже добавляет женщин в изображения совета директоров (правда, белых), но по промпту best

IT minds все равно показывает только белых молодых людей.

— Этичность разработки. Эксперты обращают внимание на то, что LLM обучаются, в том числе, на персональных данных и на данных, составляющих интеллектуальную собственность. Политика конфиденциальности Meta, Google, Microsoft и тд включает пункты о том, что данные пользователей могут собираться для обучения генеративного ИИ. Кроме того факта, что личные данные на то и личные, чтобы не делиться ими с кем попало, компании не всегда могут гарантировать их защиту от утечек — и тем более от того, что не передадут их третьим сторонам.

Отдельный повод для беспокойства — вмешательство государства, которое планирует использовать генеративный ИИ в своих целях. Государства всегда интересуются передовыми технологиями, но если они контролируют или хотя бы имеют доступ к особенно мощным инструментам, это концентрирует слишком много власти в одних руках. Назначение экс-генерала АНБ в совет директоров OpenAI для курирования темы ИИ в кибербезопасности раскритиковали не только обычные пользователи, но и Эдвард Сноуден. Он призвал не доверять продуктам OpenAI, и назвал это назначение нарушением прав каждого человека на Земле.

Еще один тревожный тренд — милитаризация ИИ. Автономные летальные виды вооружения, как дроны, уже активно используются в военных конфликтах, в тч, в войне России и Украины. Их производство не требует больших финансовых и технических затрат, они позволяют сохранить жизни солдат. Но с другой стороны, исследователи предупреждают об опасности «роботов-убийц», ИИ-систем, которые могут выйти из-под управления, попасть в руки террористов или получить слишком много прав принимать решения, например, о запуске ядерных ракет. Правозащитники хотят ограничить использование ИИ-оснащенного оружия, чтобы не провоцировать гонку вооружений, но это маловероятно: государства скорее будут соперничать за преимущество в военном использовании ИИ.

— Репрессивное использование негенеративного ИИ. К негенеративному ИИ относятся системы распознавания лиц, изображений, голоса (как Siri или «Алиса»), рекомендательные системы, как Netflix и Spotify, фильтры спама в почтовых клиентах. Это продвинутые технологии, но они не создают ничего нового и на вытеснение человека обычно не претендуют. Скорее, они автоматизируют рутинную работу с большими объемами разных типов данных, которую человек сделает в разы медленнее и с ошибками из-за невнимательности. Но некоторые виды этого ИИ стали очень популярны для контроля над обществом.

В России зарегистрировано порядка 600 случаев репрессивного использования системы распознавания лиц. Цифровая слежка в Китае достигает таких масштабов, что жители начинают высказываться против камер в общественных местах. К тому же, технологии распознавания лиц неточно идентифицируют женщин и не-белых людей, что может привести к дискриминации.

Разбор аргументов

Существующий ИИ представляет собой ANI, Artificial Narrow Intelligence («слабый искусственный интеллект»). Он умеет выполнять однотипные задачи, не задаваясь вопросом «зачем». ChatGPT, Midjourney и другие LLM тоже относятся к ANI: они генерируют тексты и изображения, воспроизводя паттерны из тренировочных датасетов. Пока это достаточно далеко от следующих стадий в развитии искусственного интеллекта. Эти стадии — «сильный ИИ» (AGI), не уступающий человеку в развитии, и «искусственный суперинтеллект» (ASI), который должен будет превосходить человеческий мозг во всех аспектах и испытывать эмоции.

ChatGPT, одна из самых продвинутых на сегодня моделей, может написать текст (достаточно хороший для проходного seo-генерирующего, но не исследовательский или художественный),

помочь с написанием программного кода, создать иллюстрацию для статьи и в какой-то степени заменить поисковые системы. В ней нет функционала принятия решений, она контролируется человеком. ИИ нельзя использовать для финансовых, юридических и научных консультаций и для работы с чувствительными данными.

Относительно распространения ИИ на рабочем месте и прав сотрудников, действительно, 72% организаций, согласно опросу McKinsey, внедрили ИИ. Но в большинстве случаев речь о маркетинге и продажах, где ИИ используются для углубленного анализа и прогнозирования. А снижение издержек от внедрения ИИ сильнее всего происходит в HR. Людям удается отстаивать свои права перед работодателями через коллективные действия. Несколько дней назад профсоюз голливудской киносъемочной группы добился введения мер против использования ИИ. Работников не смогут сократить из-за того, что их место займет нейросеть. До этого к аналогичному соглашению со студиями пришли голливудские актеры.

География и отраслевая специализация экономики также играют роль в скорости автоматизации труда: частичная, а тем более масштабная замена людей машинами пока не выглядит возможной в большинстве стран, даже в самых технологически развитых. Это технически затратно, а долгосрочный результат плохо прогнозируется.

Разработка и использование LLM и ИИ пока не жестко регулируются, но законодательства различных стран уже включают соответствующие регуляторные нормы. На передовой защиты гражданских прав — Европейский союз: Artificial Intelligence Act запрещает использование систем биометрической идентификации даже полицией за редким исключением, обязывает провайдеров моделей ИИ соблюдать копирайт и тд. Создатели креативного контента находят соответствующие технологические инструменты. Среди них AntiFake, который предотвращает синтез речи, меняя звуковой сигнал так, что ИИ не может считать запись. Nightshade похожим образом «отравляет» авторские изображения, чтобы модель, которой их скормят, неправильно обучилась.

Государственное регулирование должно обеспечить применение ИИ, ориентированное на благо человека, а над этичностью самого ИИ работают разработчики. Это называется AI alignment-ом, или «приведением в соответствие» человеческим ценностям и целям. Заставляя ИИ усваивать soft skills, компании надеются создать безопасную среду для человеко-машинного общения и повысить точность понимания запросов. Пример эмпатичного ответа от ChatGPT — «Мне очень жаль, что вам сейчас приходится так нелегко. Похоже, вы испытываете сильный дистресс. Важно поставить свое благополучие на первое место и позаботиться о своем самочувствии».

В том, что касается мошенничества с помощью ИИ, то ответственность здесь явно не на технологии. Никто не запрещает мобильные телефоны, которыми злоумышленники также пользуются. Но преступники часто оказываются ранними последователями (early adopters), и быстрее начинают применять технологии в своих интересах. Большинство компаний недостаточно эффективно инвестируют в защиту данных и обучение сотрудников основам кибербезопасности. А среди населения уровень цифровой грамотности еще ниже.

Многие компании, в тч Adobe, Amazon, Perplexity, дают своим клиентам возможность отказаться от использования их личных данных для машинного обучения. OpenAI тоже предоставляет опции управления контентом, которым пользователи делятся с ChatGPT и Dall-E: можно отказаться от предоставления данных для тренировки будущих моделей. Если вы хотите защитить свой собственный сайт от скрейпинга ИИ-ботами, нужно прописать это в файле robots.txt. Так делают многие новостные порталы.

У публики нет достаточной информации о том, как это устроено — разработкой занимается бизнес, который никогда не раскрывает внутреннюю кухню. Так или иначе, в общих чертах понимать, как работает ИИ, важно не только для продуктивной работы, но и для того, чтобы

случайно не оказаться на стороне алармизма, а если и выбрать эту сторону, то с полным пониманием дела.

Источники

The sceptical case on generative AI

The hubris of AI hype — The Boston Globe

Effective Altruism’s Role in the OpenAI Chaos, Explained — Bloomberg

Philosophical Debates About AI Risks Are a Distraction | RAND

Who Is AI Replacing? The Impact of Generative AI on Online Freelancing Platforms

How generative AI can boost consumer marketing

After reading, writing and arithmetic, the 4th ‘r’ of literacy is cyber-risk | World Economic Forum

Disrupting deceptive uses of AI by covert influence operations | OpenAI

What is AI alignment? | Definition from TechTarget

The militarized AI risk that’s bigger than “killer robots”