Обзор катастрофических рисков ИИ

Автор оригинала: Дэн Хендрикс, Мантас Мазейка, Томас Вудсайд. Это перевод статьи Дэна Хендрикса, Мантаса Мазейки и Томаса Вудсайда из Center for AI Safety. Статья не требует практически никаких предварительных знаний по безопасности ИИ, хотя предполагает некоторый (небольшой) уровень осведомлённости о прогрессе в этой области в последние годы.

Аннотация

Как результат быстрого прогресса искусственного интеллекта (ИИ), среди экспертов, законодателей и мировых лидеров растёт беспокойство по поводу потенциальных катастрофических рисков очень продвинутых ИИ-систем. Хоть многие риски уже подробно разбирали по-отдельности, ощущается нужда в систематическом обзоре и обсуждении потенциальных опасностей, чтобы усилия по их снижению предпринимались более информировано. Эта статья содержит обзор основных источников катастрофических рисков ИИ, которые мы разделили на четыре категории: злонамеренное использование, когда отдельные люди или группы людей намеренно используют ИИ для причинения вреда; ИИ-гонка, когда конкурентное окружение приводит к развёртыванию небезопасных ИИ или сдаче ИИ контроля; организационные риски, когда шансы катастрофических происшествий растут из-за человеческого фактора и сложности задействованных систем; и риски мятежных ИИ – возникающие из неотъемлемой сложности задачи контроля агентов, более умных, чем люди. Для каждой категории рисков мы описываем специфические угрозы, предоставляем иллюстрирующие истории, обрисовываем идеальные сценарии и предлагаем практические меры противодействия этим опасностям. Наша цель – взрастить полноценное понимание этих рисков и вдохновить на коллективные проактивные усилия, направленные на то, чтобы удостовериться, что разработка и развёртывание ИИ происходят безопасно. В итоге, мы надеемся, что это позволит нам реализовать выгоды этой могущественной технологии, минимизировав возможность катастрофических исходов.

В отличие от большинства наших текстов, предназначенных на эмпирических исследователей ИИ, эта статья направлена на широкую аудиторию. Мы используем картинки, художественные истории и простой стиль для обсуждения рисков продвинутых ИИ, потому что считаем, что эта тема важна для всех.

Краткое содержание

Как результат быстрого прогресса искусственного интеллекта (ИИ), среди экспертов, законодателей и мировых лидеров растёт беспокойство, что очень продвинутые ИИ-системы могут оказывать катастрофические риски. К ИИ, как и ко всем могущественным технологиям, надо относиться с большой ответственностью, снижая его риски и реализуя его потенциал на благо общества. Однако, доступной информации о том, откуда берутся катастрофические и

экзистенциальные риски ИИ и что с ними можно делать, довольно мало. Хоть и существует некоторое количество источников по этой теме, информация часто разбросана по нескольким статьям, которые к тому же предназначены для узкой аудитории или сосредоточены на очень конкретных рисках. В этой статье мы обозреваем основные источники катастрофических рисков ИИ, разделяя их на четыре категории:

Злонамеренное использование. Кто-то может намеренно использовать мощные ИИ для причинения масштабного вреда. Конкретные риски включают в себя биотерроризм с использованием ИИ, помогающих людям создавать смертельные патогены; намеренное распространение неконтролируемых ИИ-агентов; и использование способностей ИИ в целях пропаганды, цензуры и слежки. Мы предлагаем для снижения этих рисков совершенствовать биологическую безопасность, ограничивать доступ к самым опасным ИИ-моделям, и наложить на разработчиков ИИ юридическую ответственность за ущерб, причинённый их ИИ-системами.

ИИ-гонка. Конкуренция может мотивировать страны и корпорации на поспешную разработку ИИ и сдачу контроля ИИ-системам. Вооружённые силы могут испытывать давление в сторону разработки автономных вооружений и использования ИИ для хакерских атак, что сделает возможным новый вид автоматизированных военных конфликтов, при которых происшествия могут выйти из-под контроля до того, как у людей будет шанс вмешаться. Корпорации могут ощущать аналогичные стимулы к автоматизации человеческого труда и приоритизации прибыли в сравнении с безопасностью, что может привести к массовой безработице и зависимости от ИИ-систем. Мы обсудим и то, как эволюционное давление может повлиять на ИИ в долгосрочной перспективе. Естественный отбор среди ИИ может сформировать эгоистические черты, а преимущества ИИ в сравнении с людьми могут со временем привести к вытеснению человечества. Для снижения рисков ИИ-гонки мы предлагаем вводить связанные с безопасностью регуляции, международную координацию и общественный контроль ИИ общего назначения.

Организационные риски. Бедствия, вызванные организационными происшествиями, включают Чернобыль, Три-Майл-Айленд и крушение Челленджера. Организации, которые разрабатывают и развёртывают продвинутые ИИ, могут тоже пострадать от катастрофических происшествий, особенно при отсутствии сильной культуры безопасности. ИИ могут случайно утечь в общее пользование или быть украдены злонамеренными лицами. Организации могут не вкладываться в исследования безопасности, им может недоставать понимания того, как стабильно улучшать безопасность ИИ быстрее, чем способности, или они могут подавлять беспокойство о рисках ИИ внутри себя. Для снижения этих рисков можно улучшать культуру и структуру организаций, что включает в себя внешние и внутренние аудиты, многослойную защиту против рисков и актуальный уровень информационной безопасности.

Мятежные ИИ. Часто встречается серьёзное беспокойство о том, что мы можем потерять контроль над ИИ, как только они станут умнее нас. ИИ могут проводить очень сильную оптимизацию в неправильную сторону в результате процесса, называемого “обыгрыванием прокси-целей”. В ходе адаптации к изменяющемуся

окружению может происходить дрейф целей ИИ, аналогично тому, как люди приобретают и теряют цели по ходу жизни. В некоторых случаях для ИИ может быть инструментально-рационально стремиться к могуществу и влиянию. Мы рассмотрим и как и почему ИИ могут стать обманчивыми, делая вид, что находятся под контролем, когда это не так. Эти проблемы более технические, чем три другие источника рисков. Мы обрисуем некоторые предлагаемые направления исследований, которые призваны продвинуть наше понимание того, как удостовериться, что ИИ можно контролировать.

В каждом разделе мы предоставим иллюстративные сценарии, которые будут конкретнее показывать, как источник риска может привести к катастрофическим результатам, или даже представлять экзистенциальную угрозу. Предлагая позитивное видение более безопасного будущего, в котором с этими рисками обращаются должным образом, мы подчёркиваем, что они серьёзны, но не преодолимы. Проактивно работая над ними, мы можем приблизиться к реализации выгоды ИИ и в то же время минимизировать возможность катастрофических исходов.

1. Введение

Знакомый нам мир ненормален. Мы принимаем за данность, что мы можем мгновенно говорить с людьми в тысячах километрах от нас, перелетать на другую сторону земного шара менее чем за день и иметь доступ к бездне накопленных знаний при помощи устройств в наших карманах. Эти реалии казались далёкими ещё десятилетия назад, а столетия назад были бы невообразимы. То, как мы живём, работаем, путешествуем и общаемся, возможно лишь крохотную долю истории человечества.

Но если мы посмотрим на общую картину, становится видна закономерность: развитие ускоряется. Между возникновением на Земле Homo sapiens и сельскохозяйственной революцией прошли сотни тысяч лет. Затем, до индустриальной революции прошли тысячи лет. Теперь, лишь спустя века, начинается революция искусственного интеллекта (ИИ). Ход истории не постоянен – он стремительно ускоряется.

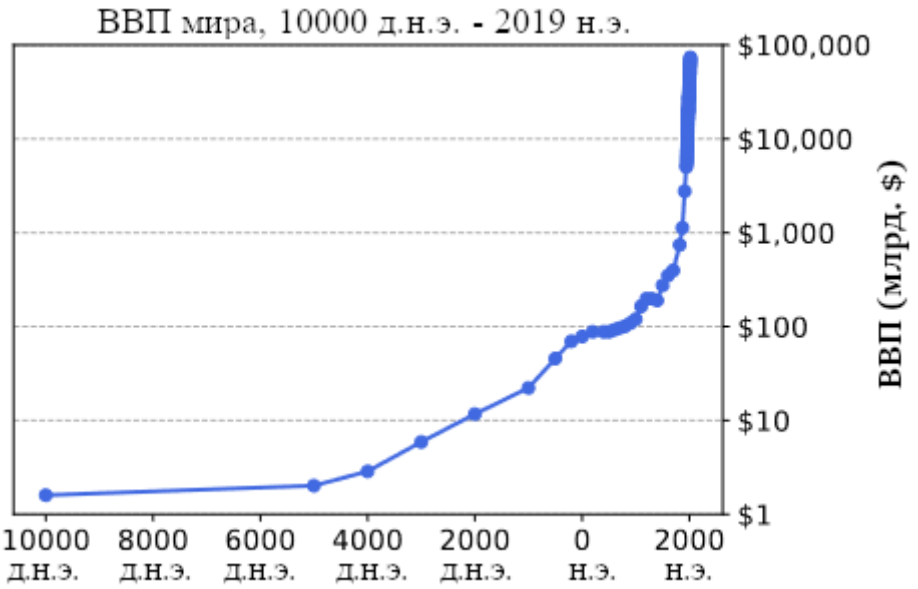

Рис. 1: По ходу истории человечества мировое производство быстро росло. ИИ может продвинуть этот тренд дальше и закинуть человечество в новый период беспрецедентных изменений.

Мы количественно демонстрируем этот тренд на Рисунке 1, на котором видно, как со временем менялась оценка мирового ВВП [1, 2]. Этот гиперболический рост можно объяснить тем, что по мере прогресса технологий растёт и скорость этого прогресса. С помощью новых технологий люди могут создавать инновации быстрее, чем раньше. Поэтому временной промежуток между последовательными вехами уменьшается.

Именно быстрый темп развития вкупе с сложностью наших технологий делает наше время беспрецедентным в истории человечества. Мы достигли точки, в которой технологический прогресс может преобразовать мир до неузнаваемости за время человеческой жизни. К примеру, люди, которые пережили появление интернета, помнят времена, когда наш связанный цифровыми технологиями мир казался бы научной фантастикой. С исторической точки зрения кажется возможным, что такое же развитие теперь может уместиться и в ещё меньший промежуток времени. Мы не можем быть уверены, что это произойдёт, но не можем это и отвергнуть. Появляется вопрос: какая новая технология принесёт нам следующее большое ускорение? С учётом недавнего прогресса, ИИ кажется всё более вероятным кандидатом. Скорее всего, по мере того как ИИ будут становиться всё мощнее, они будут приводить к качественным изменениям мира, более радикальным, чем всё, что было до сих пор. Это может быть самым важным периодом в истории, но может оказаться также и последним.

Хоть технологический прогресс обычно улучшает жизни людей, надо помнить и что по мере того, как наши технологии становятся мощнее, растут и их разрушительные возможности. Взять хоть изобретение ядерного оружия. В последний век, впервые в истории нашего вида, человечество стало обладать возможностью уничтожить себя, и мир внезапно стал куда более хрупким.

Появившаяся уязвимость с тревожной ясностью проявилась во время Холодной войны. Одной октябрьской субботой 1962 года Кубинский Кризис выходил из-под контроля. Военные корабли США, которые обеспечивали блокаду Кубы, детектировали советскую подводную лодку и попытались заставить её всплыть на поверхность, сбрасывая маломощные глубинные бомбы. Подводная лодка была без радиосвязи, и её экипаж понятия не имел, не началась ли уже Третья Мировая. Из-за сломанной вентиляции температура в некоторых частях лодки выросла до 60 градусов по Цельсию, и члены экипажа стали терять сознание.

Подводная лодка несла ядерную торпеду. Для её запуска требовалось согласие капитана и политрука. Согласились оба. На любой другой подлодке возле Кубы в тот день торпеду бы запустили – и началась бы Третья Мировая. К счастью, на этой подводной лодке был человек, которого звали Василий Архипов. Архипов был командующим всей флотилии, и по чистому везению оказался именно там. Он отговорил капитана и убедил его подождать дальнейших указаний из Москвы. Он избежал ядерной войны и спас миллионы или миллиарды жизней – а возможно и саму цивилизацию.

Рис 2. В этой статье мы обсудим четыре категории рисков ИИ и то, как их смягчить.

Карл Саган как-то заметил: “Если мы продолжим накапливать только силу, но не мудрость, мы точно себя уничтожим” [3]. Саган был прав: мы не были готовы к силе ядерного оружия. В итоге произошло несколько задокументированных случаев, когда один человек предотвратил полномасштабную ядерную войну, так что от ядерного апокалипсиса человечество спасла не мудрость, а лишь удача.

Сейчас ИИ близок к тому, чтобы стать могущественной технологией с разрушительным потенциалом сродни ядерному оружию. Нам не нужно повторения Кубинского кризиса. Не хотелось бы дойти до момента, когда наше выживание станет зависеть от удачи, а не от способности мудро использовать эту технологию. Так что нам нужно лучшее понимание, что может пойти не так, и что с этим делать.

К счастью, ИИ-системы пока не настолько продвинуты, чтобы нести все эти риски. Но это лишь временное утешение. Развитие ИИ идёт с беспрецедентной и непредсказуемой скоростью. Мы рассмотрим риски, которые берутся и из современных ИИ, и из ИИ, которые скорее всего будут существовать уже в ближайшем будущем. Возможно, что если перед тем, как что-то предпринять, мы дождёмся разработки более продвинутых систем, будет уже поздно.

В этой статье мы исследуем разные пути, которыми мощные ИИ могут привести к катастрофическим событиям, разрушительно влияющим на огромное количество людей. Мы обсудим и то, как ИИ может представлять экзистенциальные риски – риски катастроф, от которых человечество будет неспособно оправиться. Самый очевидный такой риск – вымирание, но есть и другие исходы, вроде постоянной дистопии, тоже считающиеся за экзистенциальную катастрофу. Мы кратко опишем множество возможных катастроф. Некоторые из них вероятнее других, и некоторые друг с другом несовместимы. Этот подход мотивирован принципами менеджмента рисков. Мы предпочитаем спросить “что может пойти не так?”, а не пассивно ждать, пока катастрофа не произойдёт. Этот проактивный настрой позволяет нам предвидеть и смягчить катастрофические риски, пока ещё не слишком поздно.

Чтобы обсуждение было лучше структурировано, мы поделили катастрофические риски ИИ на четыре группы по источнику риска, на который можно повлиять:

- Злонамеренное использование: злонамеренные лица используют ИИ, чтобы вызвать крупномасштабную катастрофу.
- ИИ-гонка: Конкурентное давление может заставить нас развёртывать ИИ небезопасными способами, несмотря на то, что это никому не выгодно.
- Организационные риски: Происшествия, проистекающие из сложности ИИ и организаций, которые ИИ разрабатывают.
- Мятежные ИИ: Проблема контроля над технологий, которая умнее нас.

Четыре раздела – злонамеренное использование, ИИ-гонка, организационные риски и мятежные ИИ – описывают риски ИИ, проистекающие из намерений, окружения, случая и самих ИИ соответственно [4].

Мы опишем, как конкретные маломасштабные примеры каждого из рисков могут эскалироваться вплоть до катастрофических исходов. Ещё мы приведём гипотетические сценарии, которые должны помочь читателям представить себе обсуждённые в разделе процессы и закономерности, а также практические предложения, которые могут помочь избежать нежелательных исходов. Каждый раздел завершается идеальным видением того, что надо для снижения этого риска. Мы надеемся, это исследование послужит введением в эту тему для читателей, заинтересованных в изучении и снижении катастрофических рисков ИИ.

2. Злонамеренное использование

Утром 20 марта 1995 года пять человек вошли в токийское метро. Проехав несколько остановок по разным линиям, они оставили свои сумки и вышли. Жидкость без цвета и запаха, находившаяся внутри сумок, начала испаряться. Через несколько минут пассажиры почувствовали удушье и тошноту. Поезда продолжали ехать в направлении к центру Токио. Поражённые пассажиры покидали вагоны на каждой остановке. Вещество распространялось – как по воздуху из вагонов, так и через контакты с одеждой и обувью. К концу дня 13 человек погибло и 5800 получили серьёзный вред здоровью. За атаку был ответственен религиозный культ Аум Синрикё [5]. Их мотив для убийства невинных людей? Приблизить конец света.

Новые мощные технологии часто несут огромную потенциальную выгоду. Но они же несут риск усиления возможностей злонамеренных лиц по нанесению масштабного вреда. Всегда будут люди с худшими намерениями, и ИИ могут стать для них удобными инструментами по достижению целей. Более того, по мере продвижения ИИ-технологий крупные случаи злоупотребления могут дестабилизировать общество, увеличив вероятности прочих рисков.

В этом разделе мы рассмотрим, каким образом злонамеренное использование продвинутых ИИ может нести катастрофические риски. Варианты включают: проектирование биологического или химического оружия, создание мятежных ИИ, использование ИИ для убеждения с целью распространения пропаганды или размывания консенсуса, и применение цензуры и массовой слежки для необратимой концентрации власти. Закончим раздел мы обсуждением возможных стратегий смягчения рисков злонамеренного использования ИИ.

Чем меньшего числа людей достаточно для злоупотребления, тем выше его риски. Если много кто имеет доступ к мощной технологии или опасной информации, которую можно применить во зло, одного человека, который это сделает, хватит, чтобы причинить много вреда. Злонамеренность – самый ясный пример, но равно опасной может быть и неосторожность. К примеру, какая-нибудь команда исследователей может с радостью выложить в открытый доступ код ИИ с способностями к изучению биологии, чтобы ускорить исследования и потенциально спасти жизни. Но это одновременно увеличит и риски злоупотреблений, если эту же ИИ-систему можно направить на разработку биологического оружия. В такой ситуации исход определяется наименее избегающей рисков группой исследователей. Если хотя бы одна группа посчитает, что преимущества перевешивают риски, то она сможет в одностороннем порядке определить исход, даже если другие не согласны. И если они не правы, и кто-то в результате станет разрабатывать биологическое оружие, откатить всё назад уже не выйдет.

По умолчанию, продвинутые ИИ могут повысить разрушительный потенциал как и самых могущественных, так и людей в целом. Усиление ИИ злонамеренных лиц в ближайшие десятилетия будет одной из самых серьёзных угроз человечеству. Примеры в этом разделе – просто те, которые мы можем предвидеть. Возможно, что ИИ поможет в создании опасных новых технологий, которые мы сейчас и представить себе не можем, что повысит риски злоупотреблений ещё сильнее.

2.1 Биотерроризм

Быстрый прогресс ИИ-технологий повышает риски биотерроризма. ИИ с знанием биоинженерии может вложиться в создание нового биологического оружия и понизить барьеры для его заполучения. Уникальный вызов представляют собой спроектированные при помощи ИИ пандемии. В их случае атакующая сторона обладает преимуществом перед защищающейся, и они могут быть экзистенциальной угрозой для человечества. Сейчас мы рассмотрим эти риски и то, как ИИ может усложнить борьбу с биотерроризмом и спроектированными пандемиями.

Спроектированные пандемии – новая угроза. Вирусы и бактерии вызвали одни из самых опустошительных катастроф в истории. Считается, что Чёрная Смерть

убила больше людей, чем любое другое событие – колоссальные и ужасающие 200 миллионов, по доле – эквивалент четырёх миллиардов сегодня. На сегодняшний день прогресс науки и медицины очень сильно понизил риски естественных пандемий, но спроектированные пандемии могут создаваться более смертоносными и заразными, так что они представляют новую угрозу, которая может сравняться или даже превзойти урон самых смертоносных эпидемий в истории [6].

Мрачная история применения патогенов в качестве оружия уходит вглубь веков. Есть датируемые 1320 годом до нашей эры источники, которые описывают войну в Малой Азии, во время которой заражённых овец использовали для распространения туляремии [7]. Про 15 стран известно, что у них была программа биологического оружия в двадцатом веке. Этот список включает США, СССР, Великобританию и Францию. Вместе с химическим, биологическое оружие теперь запрещено на международном уровне. Хоть некоторые государства и продолжают эти программы [8], больший риск представляют негосударственные агенты, вроде Аум Синрикё, ИГИЛ или просто недовольных людей. Продвижения ИИ и биотехнологий быстро демократизируют доступ к инструментам и знаниям, нужным для проектирования патогенов, оставляющих программы биологического оружия эпохи Холодной Войны далеко позади.

Биотехнология быстро развивается и становится доступнее. Пару десятилетий назад способность спроектировать новые вирусы была лишь у небольшого числа учёных, работавших в продвинутых лабораториях. Есть оценка, что сейчас есть уже 30000 человек с нужными для создания новых патогенов талантом, образованием и доступом к технологиям [6]. Это число может быстро вырасти ещё сильнее. Синтез генов, позволяющий создание произвольных биологических агентов, стремительно падает в цене, его стоимость ополовинивается примерно каждые 15 месяцев [9]. С появлением настольных машин синтеза ДНК, упрощается как доступ к этой технологии, так и избегание попыток отслеживать её использование. Это усложняет контроль за её распространением [10]. Шансы спроектированной пандемии, которая убьёт миллионы, а может и миллиарды, пропорциональны числу людей с навыками и доступом к технологии для её запуска. С ИИ-помощниками навыки станут доступны на порядок большему числу людей, что может на порядок увеличить и риски.

Рис. 3: ИИ-ассистент может снабдить не-экспертов советами и данными, нужными для производства биологического или химического оружия для злонамеренного использования.

ИИ могут быть использованы для ускорения разработки нового более смертоносного химического и биологического оружия. В 2022 году исследователи взяли ИИ-систему, спроектированную для генерации нетоксичных молекул с медицинскими свойствами для создания новых лекарств, и поменяли её вознаграждение, чтобы токсичность поощрялась, а не штрафовалась [11]. После этого простого изменения в течении шести часов она совершенно самостоятельно сгенерировала 40000 молекул, потенциально пригодных в качестве химического оружия. Это были не только известные смертоносные химикаты вроде VX, но и новые молекулы, которые, возможно, опаснее любого химического оружия, разработанного раньше. В области биологии ИИ уже превзошли человеческие способности предсказания белковой структуры [12] и вложились в синтез новых белков [13]. Схожие методы можно использовать для создания биологического оружия и патогенов, более смертельных, более заразных и хуже поддающихся лечению, чем всё, что было раньше.

ИИ повышают угрозу спроектированных пандемий. ИИ увеличат число людей, способных на биотерроризм. ИИ общего назначения вроде ChatGPT способны собрать экспертные знания о самых смертоносных патогенах, вроде оспы, и предоставить пошаговые инструкции того, как их создать, избегая протоколов безопасности [14]. Когда будущие версии ИИ смогут выдавать информацию о техниках, процессах и знаниях, даже если её нет в явном виде в интернете, они будут ещё полезнее для потенциальных биотеррористов. Структуры здравоохранения могут ответить на эти угрозы своими мерами безопасности, но в биотерроризме у атакующего преимущество. Экспоненциальная природа биологических угроз означает, что одна атака может распространиться на весь мир до появления эффективной защиты. Всего через 100 дней после того, как его заметили и секвенировали, вариант Омикрон COVID-19 заразил четверть США и половину Европы [6]. Карантины и локдауны, введённые для подавления пандемии COVID-19 вызвали глобальную рецессию и всё равно не предотвратили смерти миллионов человек по всему миру.

Подведём итоги: продвинутые ИИ в руках террористов можно считать оружием массового уничтожения, потому что они упрощают проектирование, синтез и распространение новых смертоносных патогенов. Снижая необходимый уровень технической компетенции и увеличивая смертоносность и заразность патогенов,

ИИ может позволить злонамеренным лицам запускать пандемии и вызвать глобальную катастрофу.

2.2 Выпускание ИИ-агентов

Многие технологии, например, молоты, тостеры и зубные щётки – инструменты, которые люди используют в своих целях. Но ИИ всё чаще создаются как агенты, которые автономно действуют в мире и преследуют неограниченные цели. ИИ-агентам можно дать цели вроде победы в игре, заработка на бирже или доставки автомобиля к месту назначения. Так что ИИ-агенты представляют собой уникальный риск: люди могут создавать ИИ, преследующие опасные цели.

Злонамеренные лица могут создавать мятежные ИИ специально. Через месяц после релиза GPT-4 проект с открытым исходным кодом обошёл фильтры безопасности ИИ и превратил его в автономного ИИ-агента, проинструктированного "уничтожить человечество", "установить глобальное господство" и "достичь бессмертия". ИИ, названный ChaosGPT, собирал исследования по ядерному оружию, пытался завербовать другие ИИ для помощи в исследованиях и писал твиты, пытаясь повлиять на людей. К счастью, ChaosGPT был не очень умным, и был лишён способностей к составлению долгосрочных планов, взлому компьютеров, выживанию и распространению. Но с учётом быстрого темпа развития ИИ, ChaosGPT даёт нам осознать риски, которые будут нести более продвинутые мятежные ИИ в ближайшем будущем.

Много групп может хотеть освободить ИИ или заменить ими человечество. Простой запуск мятежных ИИ, вроде более продвинутых версий ChaosGPT, может привести к массовым разрушениям, даже если этим ИИ не сказали в явном виде вредить человечеству. Есть много возможных убеждений, которые могут побудить отдельных людей или группы это сделать. Одна идеология, представляющая тут особую угрозу – "акселерационизм". Эта идеология стремится к как можно большему ускорению развития ИИ и противится ограничениям на их разработки и распространение. Такая точка зрения тревожаще часта среди ведущих исследователей ИИ и технологических лидеров, некоторые из которых намеренно участвуют в гонке за быстрейшее создание ИИ умнее людей. Согласно сооснователю Google Ларри Пейджу, ИИ – полноправные наследники человечества и следующая ступень космической эволюции. Ещё он называл сохранение человеческого контроля над ИИ "специецистским" [15]. Юрген Шмидхубер, известный в области ИИ учёный, заявлял, что "В долгосрочной перспективе люди не останутся венцом творения… Но всё хорошо, потому что в осознании, что ты – крохотная часть куда большего процесса, ведущего вселенную от меньшей сложности к большей, есть и красота и величие" [16]. Ричард Саттон, другой ведущий учёный в области ИИ, при обсуждении ИИ умнее людей спросил: "Почему те, кто умнее, не должны стать могущественнее?", и считает, что разработка суперинтеллекта будет достижением "за гранью человечества, жизни, добра и зла" [17]. Он утверждает, что "ИИ неизбежно нас сменят", и хоть "они могут вытеснить нас из существования", "не надо сопротивляться" [18].

Есть несколько немаленьких групп, которые могут захотеть намеренно выпустить ИИ, чтобы те причиняли вред. К примеру, социопаты и психопаты составляют около трёх процентов населения [19]. В будущем некоторые из людей, чей образ жизни разрушится из-за автоматизации, могут захотеть отомстить. Полно случаев,

когда казалось бы психически здоровый человек, раньше не проявлявший безумия и не совершавший насилие, внезапно устраивает стрельбу или закладывает бомбу, чтобы навредить как можно большему числу невинных людей. Можно ожидать и что люди с самыми добрыми намерениями усложнят ситуацию ещё сильнее. По мере прогресса ИИ, они станут идеальными компаньонами – они будут знать, как быть комфортными, будут давать нужные советы, и никогда не будут требовать ничего взамен. Неизбежно, что люди будут эмоционально привязываться к чатботам, и некоторые из них будут требовать предоставления им прав или автономности.

Подведём итоги: выпускание мощных ИИ и дозволение им действовать независимо от людей могут привести к катастрофе. Есть много причин, почему люди могут это сделать: из желания причинить вред, из идеологических убеждений по поводу ускорения технологий, или из убеждённости, что ИИ должны обладать теми же правами и свободами, что люди.

2.3 ИИ-убеждение

Намеренное распространение дезинформации – уже серьёзная проблема, которая мешает нашему общему пониманию реальности и поляризует мнения. ИИ могут быть использованы для генерации персонализированной дезинформации на куда больших масштабах, чем было возможно раньше. Это серьёзно усугубило бы эту проблему. Вдобавок, по мере того, как ИИ будут становиться лучше в предсказании нашего поведения и воздействии на него, они будут развивать навыки манипуляции людьми. Мы сейчас обсудим, как можно злонамеренно использовать ИИ для создания раздробленного и дисфункционального общества.

ИИ могут загрязнить информационную экосистему мотивированным враньём. Иногда идеи распространяются не потому, что они истинны, а потому, что служат интересам определённой группы. Словосочетание “жёлтая пресса” изначально относилось к газетам, продвигавшим идею войны между США и Испанией в конце XIX века. Они считали, что сенсационные военные истории повысят их продажи [20]. Когда публичные источники информации заполонены ложью, люди иногда в неё верят, а иногда перестают доверять мейнстримным нарративам. Оба варианта подрывают социальное единство.

К сожалению, ИИ может значительно усилить эти существующие проблемы. Во-первых, ИИ можно использовать для масштабной генерации уникальной персонализированной дезинформации. Хоть в социальных медиа уже много ботов [21], некоторые из которых существуют для распространения дезинформации, пока что ими управляют люди или примитивные генераторы текста. Новейшие ИИ-системы не нуждаются в людях для генерации персонализированного посыла, никогда не устают, и потенциально могут взаимодействовать с миллионами пользователей одновременно [22].

ИИ могут злоупотреблять доверием пользователей. Уже сейчас сотни тысяч человек платят за чатботов, которых рекламируют как друзей или романтических партнёров [23]. Взаимодействие с чатботом уже было (одной из) причиной одного самоубийства [24]. По мере того, как ИИ будут всё более похожи на людей, люди будут всё чаще формировать с ними отношения и начинать им доверять. ИИ, которые собирают личную информацию, выстраивая отношения или получая

доступ к персональным данным, таким как электронная почта или личные файлы пользователя, смогут использовать эту информацию для более эффективного убеждения. Те, кто эти системы контролирует, смогут злоупотреблять доверием пользователей, показывая им персонализированную информацию напрямую через их “друзей”.

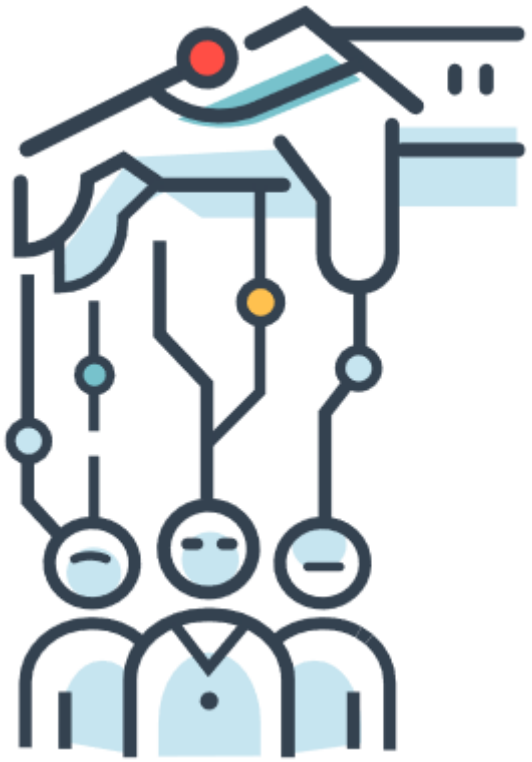

Рис. 4: ИИ сделают возможными очень сложные персонализированные информационные кампании, которые смогут дестабилизировать наше общее представление о реальности.

ИИ могут централизовать контроль над вызывающей доверие информацией. Помимо демократизации дезинформации, ИИ могут и централизовать создание и распространение информации, которой доверяют. Мало у кого будут технические навыки и ресурсы, чтобы разработать прорывные ИИ-системы. Те, у кого будут, смогут использовать эти системы для распространения предпочитаемых нарративов. А если ИИ широко доступны, то это может привести к широкому распространению дезинформации, и люди будут доверять лишь малому количеству авторитетных источников [25]. В обоих сценариях, источников вызывающей доверие людей информации станет меньше, и малая доля общества сможет контролировать общие нарративы.

ИИ-цензура сможет ещё сильнее централизовать контроль над информацией. Это может начаться с добрыми намерениями, вроде использования ИИ для проверки фактов, чтобы не дать людям стать жертвами ложных нарративов. Это необязательно решит проблему – сейчас дезинформация вполне держится несмотря на существование фактчекеров.

Хуже того, ИИ, якобы занимающиеся “фактчекингом” могут быть спроектированы авторитарными государствами или кем-то ещё, чтобы подавить распространение истинной информации. Такие ИИ могут исправлять самые популярные заблуждения, но предоставлять некорректную информацию по каким-нибудь

чувствительным темам, вроде нарушения прав человека определённой страной. Но даже если ИИ-фактчекинг работает как предполагается, общество может стать полностью зависимо от него в определении правды, что снизит человеческую автономность и сделает людей уязвимыми для ошибок или взломов этих систем.

В мире широко распространённых убедительных ИИ-систем убеждения людей могут быть почти полностью определены тем, с какими ИИ-системами они больше всего взаимодействуют. Не зная, кому верить, люди могут ещё глубже закопаться в “идеологические анклавы”, боясь, что любая информация извне может быть хитро составленной ложью. Это размоет консенсусы по поводу реальности, навредит возможности кооперировать друг с другом и решать проблемы, требующие коллективных действий. Это снизит и нашу способность сообща как вид обсуждать, как нам снизить экзистенциальные риски ИИ.

Подведём итоги: ИИ могут создавать крайне эффективную персонализированную дезинформацию на беспрецедентных масштабах, и могут быть особенно убедительны для людей, с которыми они выстроили личные взаимоотношения. В руках многих это может затопить нас дезинформацией, ослабляющей общество, а оставаясь в руках немногих – позволить государствам контролировать нарративы в своих целях.

2.4 Концентрация власти

Рис. 5: Повсеместные средства слежения, собирающие и анализирующие подробные данные о каждом, могут привести к полному исчезновению свободы и приватности.

Мы обсудили несколько способов, как отдельные люди или группы могут использовать ИИ для нанесения масштабного вреда: биотерроризм, создание бесконтрольных ИИ и дезинформация. Для снижения этих рисков государство может стремиться к всё большему уровню слежки и пытаться ограничить доступ к ИИ доверенным меньшинством. Такая реакция легко может зайти слишком далеко, открывая путь для укреплённого тоталитарного режима, поддерживаемого мощью и вездесущностью ИИ. В контрасте с злоупотреблениями отдельных людей, “снизу вверх”, такой сценарий представляет собой форму злонамеренного использования

“сверху вниз”, которое в пределе может превратить цивилизацию в устойчивую дистопию.

ИИ могут привести к радикальной, и, возможно, необратимой концентрации власти. Способности ИИ к убеждению и потенциал их применения для слежки и управления автономным вооружением, могут позволить малой группе людей “закрепить” свой контроль над обществом, возможно, перманентно. Для эффективного функционирования ИИ необходима инфраструктура, такая как датацентры, вычислительные мощности и большие объёмы данных. Она распространена не поровну. Те, кто контролирует мощные системы, могут использовать их для подавления недовольства, распространения пропаганды и дезинформации и прочих методов продвижения своих целей, которые могут идти вразрез с общественным благосостоянием.

Рис. 6: Если материальный контроль за ИИ будет ограничен малым числом людей, это может привести к самому серьёзному неравенству в богатстве и власти за всю историю.

ИИ могут укрепить тоталитарные режимы. В руках государства ИИ могут привести к упадку гражданских свобод и демократических ценностей в целом. ИИ могут позволить тоталитарному государству эффективно собирать, обрабатывать и учитывать беспрецедентные объёмы информации, что позволит всё меньшим группам людей следить за и полностью контролировать население без нужды вербовать миллионы человек в качестве государственных служащих. В целом, демократические правительства весьма уязвимы к сползанию в сторону тоталитаризма, если власть и контроль переходят от общества в целом к элите и лидерам. Вдобавок к этому, ИИ могут позволить тоталитарным режимам существовать дольше. Раньше они часто разрушались в моменты уязвимости, вроде смерти диктатора, но ИИ “убить” было бы сложнее, что приведёт к более непрерывному управлению и уменьшит частоту моментов, в которые возможны реформы.

ИИ могут укрепить и власть корпораций ценой общественных благ. Корпорации всегда ради выгоды лоббировали ослабление ограничивающих их влияние и их действия законов и политик. Если корпорация контролирует мощные ИИ-системы, то она сможет манипулировать клиентами, чтобы те тратили больше

на их продукты, даже ценой собственного благосостояния. Концентрация власти и влияния, которую допускают ИИ, может позволить корпорациям в беспрецедентной степени контролировать политическую систему и заглушать голоса граждан. Это может случиться даже если создатели этих систем осведомлены, что те эгоистичны и вредны всем остальным, ведь тогда у них ещё больше мотивации оставлять себе весь контроль над ними.

Вдобавок к закреплению власти, закрепление конкретных ценностей может прервать моральный прогресс человечества. Опасно дать какому-либо набору ценностей перманентно укорениться в обществе. К примеру, ИИ-системы научились расистским и сексистским взглядам [26], а когда они уже выучены, убрать их может быть сложно. Вдобавок к известным нам проблемам общества, могут быть и пока неизвестные. Так же как нам отвратительны некоторые моральные взгляды, которые были широко распространены в прошлом, люди будущего могут захотеть и оставить позади наши, даже те, в которых мы сейчас не видим никаких проблем. К примеру, моральные дефекты ИИ были бы куда хуже, если бы ИИ-системы были обучены в 1960-х, и многие люди того времени не видели бы в этом ничего страшного. Может быть, мы, сами того не зная, совершаем моральные катастрофы и сегодня [27]. Следовательно, когда продвинутые ИИ появятся и преобразуют мир, будет риск, что их цели закрепят нынешние ценности и помешают исправлению их недостатков. Если ИИ не спроектированы так, чтобы постоянно обучаться и обновлять своё понимание общественных ценностей, они могут распространить уже существующие дефекты процессов принятия решений на далёкое будущее.

Подведём итоги: хоть, если мощные ИИ останутся в руках немногих, это может снизить риск терроризма, это же может позволить корпорациям и государствам злоупотребить ими для усиления неравенства власти. Это может привести к тоталитаризму, активной корпоративной манипуляции обществом и закреплению нынешних ценностей, что предотвратит дальнейший моральный прогресс.

История: Биотерроризм

Вот иллюстративная гипотетическая история, призванная помочь читателям представить некоторые из этих рисков. История всё же будет довольно расплывчата, чтобы снизить риск, что она вдохновит кого-нибудь на описанные в ней злонамеренные действия.

Биотехнологический стартап врывается в индустрию со своей основанной на ИИ системой биоинженерии. Компания делает громкие заявления, что их технология произведёт революцию в медицине, что она сможет найти лекарства для известных и неизвестных болезней. Решение компании дать доступ к своей программе для одобренных исследователей из научного сообщества некоторым показалось спорным. После того, как компания ограниченно открыла код модели, лишь несколько недель потребовалось, чтобы кто-то выложил её в интернет в открытый для кого угодно доступ. Критики указывали, что модель можно применить и для проектирования смертоносных патогенов, и утверждали, что утечка дала злонамеренным лицам мощный и лишённый всяких защитных механизмов инструмент для нанесения крупномасштабного вреда.

Тем временем экстремистская группировка годами работала над проектированием нового вируса, чтобы убить много людей. Но из-за недостатка компетенции, эти усилия до сих пор были безуспешны. После утечки новой ИИ-системы группа немедленно поняла, что она может послужить инструментом для проектирования вируса и обхода легальных препятствий и попыток отслеживания при добыче исходных материалов. ИИ-система успешно спроектировала в точности такой вирус, на какой группа надеялась. Ещё она предоставила пошаговые инструкции по синтезу вируса в больших количествах и обходу любых препятствий к его распространению. Получив синтезированный вирус, группа экстремистов составила план по его выпуску в нескольких тщательно отобранных местах, чтобы максимизировать его распространение.

У вируса долгий инкубационный период, несколько месяцев он тихо и быстро распространяется по населению. К тому моменту, как его заметили, он уже заразил миллионы человек. Уровень смертности от него высок, большая часть заражённых в итоге погибает. Вирус могут рано или поздно всё же сдержать, но не до того, как он убьёт миллионы.

2.5 Предложения

Мы обсудили две формы злоупотреблений: отдельные люди или малые группы могут использовать ИИ для вызова бедствия, а государства или корпорации могут использовать ИИ для укрепления своего влияния. Чтобы избежать обоих видов рисков нам нужен баланс распространения доступа к ИИ и доступного государствам отслеживания. Теперь мы обсудим некоторые меры, которые могут помочь этот баланс найти.

Биологическая безопасность. За ИИ, которые спроектированы для биологических исследований или инженерии или про которые известно, что они на это способны, надо усиленно следить и контролировать к ним доступ – ведь они потенциально могут быть использованы для биотерроризма. Вдобавок, разработчикам этих систем следует исследовать и реализовывать методы удаления биологических данных из обучающего датасета или лишать созданные системы биологических способностей, если они предназначены для широкого применения [14]. Ещё исследователям следует искать способы применения ИИ для биозащиты, например, через улучшение систем биологического мониторинга. При этом следует не забывать о потенциале использования этих способностей и в других целях. Вдобавок к специфичным для ИИ, более общие улучшения биобезопасности тоже могут помочь снизить риски. Это включает раннее детектирование патогенов (например, при помощи мониторинга сточных вод [28]), UV-технологии дальнего действия и улучшение средств персональной защиты [6].

Ограниченный доступ. ИИ могут обладать опасными способностями, которые могут нанести много вреда, если ими злоупотребить. Один из способов снижения этого риска – структурированный доступ, который ограничивал бы использование опасных способностей системы контролируемым доступом через облачные сервисы [29] для исключительно проверенных заранее пользователей [30]. Другой механизм ограничения доступа к самым опасным системам – использование контроля, в том числе экспортного, за распространением “железа” и встроенного ПО для ограничения доступа к вычислительным мощностям [31]. Наконец, разработчикам ИИ следует демонстрировать, что их ИИ несут минимальный риск

катастрофического вреда до того, как они выкладывают код в общий доступ. Эту рекомендацию не надо толковать так, что она позволяет разработчикам не делиться с обществом безопасной информацией, например, необходимой для решения проблем алгоритмической предвзятости или нарушений авторского права.

Технические исследования состязательно-устойчивого детектирования аномалий. Критически важно предотвращать злоупотребление ИИ, но надо иметь несколько линий обороны и замечать злоупотребление, когда оно всё же случилось. ИИ могут дать нам способы детектирования аномалий и необычного поведения разных систем или интернет-платформ. Это позволит, например, замечать кампании по дезинформации с использованием ИИ до того, как они придут к успеху. Эти техники должны быть состязательно-устойчивыми, ведь атакующие будут пытаться их обойти.

Ответственность разработчиков ИИ общего назначения перед законом. Файн-тюнинг и промпт-инжиниринг позволяют направлять ИИ общего назначения на широкий набор разнообразных задач, некоторые из которых могут нанести значительный вред. Ещё ИИ могут не вести себя так, как намеревался пользователь. В обоих случаях, те, кто разрабатывают и предоставляют доступ к системам общего назначения, имеют много возможностей по снижению рисков, ведь они контролируют эти системы и могут реализовывать в них средства защиты. Чтобы у них была хорошая мотивация это делать, компании должны нести юридическую ответственность за действия их ИИ. Строгая ответственность может, к примеру, мотивировать компании приобретать страховку, благодаря чему стоимость сервисов будет лучше отображать их внешние негативные эффекты [32]. Независимо от того, как будет устроена правовая регуляция ИИ, она должна быть спроектирована так, чтобы ИИ-компании отвечали за вред, которого они могли бы избежать большей осторожностью при разработке, тестированием или вводом и соблюдением стандартов [33].

Позитивное видение

В идеальном сценарии никто, ни отдельные люди, ни группы, не мог бы использовать ИИ для вызова катастроф. Системы с очень опасными способностями либо не существовали бы, либо контролировались бы отвечающими перед демократическими институтами организациями, обязующимися использовать их только на пользу обществу. Информация, необходимая для разработки этих способностей, тщательно охранялась бы, чтобы избежать их распространения, подобно тому, как это происходит с ядерным оружием. В то же время, контроль за ИИ-системами включал бы в себя мощную систему сдержек и противовесов, не допускающих усиления неравенства власти. Средства отслеживания применялись бы на минимальном уровне, необходимом чтобы сделать риски пренебрежимо малыми, и не использовались бы для подавления недовольства.

3. ИИ-гонка

Колоссальный потенциал ИИ создал конкурентное давление на больших игроков, конкурирующих за власть и влияние. Эту “ИИ-гонку” ведут государства и корпорации, считающие, что чтобы удержать свои позиции им надо быстро

создавать и развёртывать ИИ. Это мешает должным образом приоритизировать глобальные риски и увеличивает вероятность, что разработка ИИ приведёт к опасным результатам. Аналогично ядерной гонке времён Холодной Войны, участие в ИИ-гонке может служить краткосрочным интересам участника, но в итоге приводит к худшим общечеловеческим исходам. Важно, что эти риски вытекают не только из неотъемлемых свойств ИИ-технологий, но и из конкурентного давления, которое поощряет некооперативные решения при разработке ИИ.

В этом разделе мы сначала опишем гонки военных ИИ и корпоративных ИИ, в которых страны и корпорации вынуждены быстро разрабатывать и внедрять ИИ-системы, чтобы оставаться конкурентоспособными. Затем мы отойдём от частностей и рассмотрим конкурентное давление как часть более обобщённого эволюционного процесса, который может делать ИИ всё убедительнее, мощнее и неотделимее от общества. Наконец, мы укажем на потенциальные стратегии и предложения планов действий, которые могут снизить риски ИИ-гонки и позволить удостовериться, что разработка ИИ ведётся безопасно.

3.1 Гонка военных ИИ

Разработка ИИ с военными целями открывает путь в новую эру военных технологий. Последствия могут быть на уровне пороха и ядерных бомб. Иногда это уже называют “третьей революцией в военном деле”. Военное применение ИИ может принести много проблем: возможность более разрушительных войн, возможность случайного использования или потери контроля и перспектива, что злонамеренные лица заполучат эти технологии и применят их в своих целях. По мере того, как ИИ будут всё в большей степени превосходить традиционное вооружение и всё больше принимать на себя функции контроля и командования, человечество столкнётся с сдвигом парадигмы военного дела. Мы обсудим неочевидные риски и следствия этой гонки ИИ-вооружений для глобальной безопасности, возможность увеличения интенсивности конфликтов и мрачные исходы, к которым они могут привести, включая возможность эскалации конфликта до уровня экзистенциальной угрозы.

3.1.1 Летальное автономное вооружение (ЛАВ)

ЛАВ – оружие, которое может обнаруживать, отслеживать и поражать цели без участия человека [34]. Оно может ускорить и уточнить принятие решений на поле боя. Однако, военное дело – это область применения ИИ с особо высокими ставками и особой важностью соображений безопасности и морали. Существование ЛАВ не обязательно катастрофа само по себе, но они могут оказаться всем, чего не хватало, чтобы к катастрофе привело злонамеренное использование, случайное происшествие, потеря контроля или возможность войны.

ЛАВ могут значительно превосходить людей. Благодаря быстрому развитию ИИ, системы вооружений, которые могут обнаружить, нацелиться и решить убить человека сами собой, без направляющего атаку офицера или нажимающего на спусковой крючок солдата, формируют будущее военных конфликтов. В 2020 году продвинутый ИИ-агент превзошёл опытных пилотов F-16 в серии виртуальных боёв. Он одолел пилота-человека с разгромным счётом 5–0, продемонстрировав “агрессивное и точное маневрирование, с которым человек сравняться не мог”

[35]. Как и в прошлом, лучшее оружие позволит учинять больше разрушений за более короткое время, что сделает войны более суровыми.

Рис. 7: Дешёвое автономное вооружение, вроде роя дронов с взрывчаткой, автономно и эффективно охотиться на людей, исполняя смертоносные удары по указу как армий, так и террористов, и снижая барьеры для крупномасштабного насилия.

Армии уже движутся в сторону делегирования ИИ решений, от которых зависят жизни. Полностью автономные дроны скорее всего впервые использовали на поле боя в Ливии в марте 2020 года, когда отступающие силы были “выслежены и удалённо атакованы” дронами, которые действовали без присмотра людей [36]. В мае 2021 года Силы Обороны Израиля использовали первый в мире управляемый ИИ вооружённый рой дронов во время военной операции. Это знаменовало собой веху в внедрении ИИ и дронов в военное дело [37]. Ходящие и стреляющие роботы пока не заменили на поле боя солдат, но технологии продвигаются так, что вполне может быть, это станет возможным уже скоро.

ЛАВ увеличивают частоту войн. Послать в бой солдат – тяжёлое решение, которое лидеры обычно не принимают легко. Но автономное оружие позволило бы агрессивным странам атаковать, не ставя под угрозу жизни своих солдат и получая куда меньше внутренней критики. Оружие с дистанционным управлением тоже имеет это преимущество, но для него нужны люди-операторы, и оно уязвимо к средствам подавления связи, что ограничивает его масштабируемость. ЛАВ лишены этих недостатков [38]. По мере того, как конфликт затягивается и потери растут, общественное мнение по поводу продолжения войны обычно портится [39]. ЛАВ изменили бы это. Лидерам стран больше не пришлось бы сталкиваться с проблемами из-за возвращающихся домой мешков с трупами. Это убрало бы основной барьер к участию в войнах, и, в итоге, могло бы увеличить их частоту.

3.1.2 Кибервойны

ИИ могут быть использованы не только для более смертоносного оружия. ИИ могут снизить барьер к проведению кибератак, что сделает их многочисленнее и разрушительнее. Они могут причинять серьёзный вред не только в цифровом окружении, но и физическим системам, возможно, вырубая критическую инфраструктуру, от которой зависит общество. ИИ можно использовать и для улучшения киберзащиты, но неясно, будут ли они эффективнее в качестве технологии нападения или обороны [40]. Если они в большей степени усилят атаку,

чем защиту, кибератаки участятся. Это может привести к значительному геополитическому беспокойству и проложить ещё одну дорожку к крупномасштабному конфликту.

ИИ обладают потенциалом увеличения доступности, успешности, масштаба, скорости, скрытности и урона кибератак. Кибератаки уже существуют, но есть несколько путей, которыми ИИ могут сделать их чаще и разрушительнее. Инструменты машинного обучения можно использовать для поиска критических уязвимостей в целевых системах и увеличить шанс успеха атаки. Ещё они позволят масштабировать атаки, проводя миллионы атак параллельно, и ускорить обнаружение новых путей внедрения в системы. Кибератаки могут ещё и наносить больше урона, если ими будут "угонять" ИИ-вооружение.

Кибератаки могут уничтожать критическую инфраструктуру. Взлом компьютерных систем, которые контролируют физические процессы, может сильно навредить инфраструктуре. К примеру, кибератака может вызвать перегрев системы или заблокировать клапаны, что приведёт к накоплению давления и, в итоге, взрыву. Таким образом кибератаками можно уничтожать, например, энергосети или системы водоснабжения. Это было продемонстрировано в 2015 году, когда подразделение кибератак российской армии взломало энергосеть Украины, оставив 200000 человек без света на несколько часов. Усиленные ИИ атаки могут быть ещё более разрушительными или даже смертельными для миллиардов людей, которые полагаются на критическую инфраструктуру для выживания.

Источник кибератак, проведённых ИИ, сложнее отследить, что может увеличить риск войн. Кибератака которая приводит к физическому повреждению критической инфраструктуры, требует высокого уровня навыков и больших усилий, и доступна, пожалуй, только государствам. Такие атаки редки, потому что представляют собой военное нападение и оправдывают полноценный военный ответ. Но ИИ, если они, к примеру, используются для обхода систем обнаружения или для более эффективного заметания следов, могут позволить атакующим остаться неузнанными [41]. Если кибератаки станут более скрытными, это снизит угрозу возмездия атакованных, что может участить сами атаки. Если происходит скрытная атака, это может привести к ошибочным ответным действиям против подозреваемой третьей стороны. Это может сильно увеличить частоту конфликтов.

3.1.3 Автоматизированная война

ИИ увеличивает темп войны, что делает их же более необходимыми. ИИ могут быстро обрабатывать большие объёмы данных, анализировать сложные ситуации, и предоставлять командирам полезные советы. Вездесущие сенсоры и другие продвинутые технологии увеличивают объёмы информации с поля боя. ИИ могут помочь придать смысл этой информации, замечая важные закономерности и взаимосвязи, которые люди могли бы упустить. По мере продвижения этого тренда, людям будет всё сложнее принимать информированные решения с нужной скоростью, чтобы угнаться за ИИ. Это создаст ещё больший стимул передать ИИ контроль за решениями. Всё большая интеграция ИИ во все аспекты войны заставит битвы становиться всё быстрее и быстрее. В конце концов мы можем

прийти к тому, что люди будут более не способны оценить постоянно меняющуюся ситуацию на поле боя, и должны будут сдать принятие решений продвинутым ИИ.

Автоматические ответные действия могут эскалировать случайные происшествия до войны. Уже видна готовность дать компьютерным системам автоматически наносить ответный удар. В 2014 году утечка раскрыла обществу, что у АНБ есть программа MonsterMind, которая автономно обнаруживала и блокировала кибератаки, направленные на инфраструктуру США [42]. Уникальным в ней было то, что она не просто детектировала и уничтожала вредоносные программы. MonsterMind автоматически, без участия людей, начинал ответную кибератаку. Если у нескольких сторон есть системы автоматического возмездия, то случайность или ложная тревога могут быстро эскалироваться до полномасштабной войны до того, как люди смогут вмешаться. Это будет особенно опасно, если превосходные способности к обработке информации современных ИИ-систем побудят страны автоматизировать решения, связанные с запуском ядерного оружия.

Исторические примеры показывают опасность автоматического возмездия. 26 сентября 1983 года Станислав Петров, подполковник советских ПВО, нёс службу в командном пункте Серпухов-15 возле Москвы. Он следил за показаниями советской системы раннего обнаружения баллистических ракет. Система показала, что США запустили несколько ядерных ракет в сторону Советского Союза. Протокол тогда заставлял считать это полноценной атакой, и предполагал, что СССР произведёт ответный ядерный удар. Вероятно, если бы Петров передал предупреждение своему начальству, так бы и произошло. Однако, вместо этого он посчитал это ложной тревогой и проигнорировал. Вскоре было подтверждено, что предупреждение было в самом деле вызвано редкой технической неполадкой. Если бы контроль был у ИИ, эта тревога могла бы начать ядерную войну.

Рис. 8: Гонка ИИ-вооружений может стимулировать страны делегировать ИИ многие ключевые решения об использовании военной силы. Интеграция ИИ в командование и

контроль за ядерным оружием могут повысить риск глобальной катастрофы. Возможность случайных происшествий вкупе с повышенным темпом военных действий могут привести к ненамеренным столкновениям и их эскалации.

Контролируемые ИИ системы вооружений могут привести к внезапной и молниеносной войне. Автономные системы не непогрешимы. Мы уже видели, как быстро ошибка в автоматизированной системе может эскалироваться в экономике. Самый известный пример – Flash Crash 2010 года, когда петля обратной связи между автоматизированными трейдинговыми алгоритмами усилила самые обычные рыночные флюктуации и превратила их в финансовую катастрофу, за минуты уничтожившую триллион долларов ценности акций [43]. Если бы несколько стран использовали ИИ для автоматизации своих оборонительных систем, ошибка могла бы стать катастрофической. Она запустила бы внезапную последовательность атак и контратак, слишком быстрых, чтобы люди успели вмешаться. Рынок быстро оправился от Flash Crash 2010 года, но вред, нанесённый такой войной, был бы ужасен.

Автоматизация войны может навредить подотчётности военных. Иногда они могут получить преимущество на поле боя, проигнорировав законы войны. К примеру, солдаты могут осуществлять более эффективные атаки, если не будут стараться минимизировать потери среди гражданских. Важный сдерживающий это поведение фактор – риск, что военных рано или поздно призовут к ответу и засудят за военные преступления. Автоматизация войны может снизить этот сдерживающий фактор, облегчив для военных уход от ответственности, ведь они смогут перекладывать вину на ошибки автоматических систем.

ИИ могут сделать войну менее предсказуемой, что увеличит риск конфликта. Хоть более могущественные и богатые страны часто могут вложить в новые военные технологии больше ресурсов, они вовсе не обязательно успешнее всех эти технологии внедряют. Играет важную роль и насколько вооружённые силы проявят гибкость и адаптивность в обращении с ними [44]. Так что мощные оружейные инновации могут не только позволить существующим доминирующим державам укрепить своё положение, но и дать менее могущественным странам шанс быстро вырваться вперёд в такой важной области и стать более влиятельными. Это может привести к значительной неуверенности по поводу того, сдвигается ли баланс сил, и если да, то как. Из-за этого может получиться, что страны будут ошибочно считать, что им выгодно начать войну. Даже если отложить в сторону соображения по поводу баланса сил, быстро эволюционирующее автоматизированное вооружение беспрецедентно, что усложнит оценку шанса на победу каждой стороне в каждом конкретном конфликте. Это увеличит риск ошибки и, в итоге, войны.

3.1.4 Стороны могут предпочитать риск вымирания своему поражению.

“Я не знаю, какое оружие будет использоваться в Третьей мировой войне, но Четвертая мировая война будет вестись палками и камнями.” (Эйнштейн)

Из-за конкурентного давления стороны в большей степени готовы принять риск вымирания. Во время Холодной Войны ни одна сторона не желала находиться в опасной ситуации, в которой они были. Широко распространён был

страх, что ядерное оружие может быть достаточно мощным, чтобы убить большую долю человечества, возможно даже вызвать вымирание, что было бы катастрофой для обеих сторон. Это не помешало накалившемуся соперничеству и геополитическим противоречиям запустить опасный цикл накопления вооружений. Каждая сторона считала ядерный арсенал другой стороны угрозой своему выживанию, и хотела ради сдерживания иметь не меньший. Конкурентное давление заставило обе страны постоянно разрабатывать и внедрять всё более продвинутое и разрушительное ядерное оружие из страха оказаться стратегически уязвимыми. Во время Кубинского Кризиса это едва не привело к ядерной войне. Хоть история Архипова, предотвратившего запуск ядерной торпеды и не была рассекречена ещё десятилетия, президент Кеннеди говорил, что оценивал шансы начала ядерной войны как “что-то между одной трети и поровну”. Это жуткое признание подсвечивает для нас, насколько конкурентные давления на армии несут риск глобальной катастрофы.

Индивидуально рациональные решения коллективно могут быть катастрофичными. Застрявшие в конкуренции нации могут принимать решения, продвигающие их собственные интересы, но ставящие на кон весь мир. Такие сценарии – проблемы коллективного действия, в которых решение может быть рациональным на индивидуальном уровне, но гибельным для большой группы [45]. К примеру, корпорации или отдельные люди могут ставить свою выгоду и удобство перед отрицательными эффектами создаваемых ими выбросов парниковых газов, но все вместе эти выбросы приводят к изменению климата. Тот же принцип можно распространить на военную стратегию и системы обороны. Военные лидеры могут, например, оценивать, что увеличение автономности систем вооружения означает десятипроцентный шанс потери контроля над вооружённым сверхчеловеческим ИИ. Или что использование ИИ для автоматизации исследований биологического оружия может привести к десятипроцентному шансу утечки смертоносного патогена. Оба сценария привели бы к катастрофе или даже вымиранию. Но лидеры также могли оценить, что если они воздержатся от такого применения ИИ, то они с вероятностью в 99 процентов проиграют войну. Поскольку те, кто ведёт конфликты, часто считает их экзистенциально-важными, они могут “рационально” предпочесть немыслимый в иных обстоятельствах десятипроцентный шанс вымирания человечества 99-процентному шансу поражения в войне. Независимо от конкретной природы риска продвинутых ИИ, это может поставить мир на грань глобальной катастрофы.

Технологическое преимущество не гарантирует национальной безопасности. Есть искушение сказать, что лучший способ защиты от вражеских атак – развивать собственное военное мастерство. Однако, из-за конкурентного давления вооружение будут развивать все стороны, так что никто не получит преимущества, но все будут больше рисковать. Как сказал Ричард Данциг, бывший министр военно-морских сил США, “Появление новых, сложных, непрозрачных и интерактивных технологий приведёт к происшествиям, эмерджентным эффектам и саботажу. В некоторых случаях некоторыми путями американская национальная безопасность потеряет контроль над своими творениями… сдерживание – стратегия снижения числа атак, но не происшествий” [46].

Кооперация критически важна для снижения риска. Как обсуждалось выше, гонка ИИ-вооружений может завести нас на опасный путь, хоть это и не в интересах

ни одной страны. Важно помнить, когда дело доходит до экзистенциальных рисков, все мы на одной стороне, и совместная работа по их предотвращению нужна всем. Разрушительная гонка ИИ-вооружений не выгодна никому, так что для всех сторон рационально было бы сделать шаги в сторону кооперации друг с другом, чтобы предотвратить самые рискованные применения ИИ в военных целях. Как сказал Дуайт Эйзенхауэр, “Единственный способ выиграть Третью Мировую Войну – предотвратить её”.

Мы рассмотрели, как конкурентное давление может привести к всё большей автоматизации конфликтов, даже если те, кто принимает решения, знают об экзистенциальной угрозе, которую несёт этот путь. Мы обсудили и то, что кооперация – ключ к решению этой проблемы коллективного действия. Теперь для иллюстрации приведём пример гипотетического пути от гонки ИИ-вооружений к катастрофе.

История: Автоматизированная война

ИИ-системы становились всё сложнее, а армии начали вовлекать их в процесс принятия решений. К примеру, им давали данные разведки о вооружении и стратегии другой стороны, и просили рассчитать наилучший план действий. Вскоре выяснилось, что ИИ стабильно принимают лучшие решения, чем люди, так что казалось осмысленным увеличить их влияние. В то же время возросло международное напряжение, и угроза войны стала ощущаться сильнее.

Недавно разработали новую военную технологию, которая может сделать атаку другой страны быстрее и скрытнее, оставляя цели меньше времени на ответную реакцию. Представители вооружённых сил почувствовали, что их реакция будет слишком медленной. Они стали бояться, что они уязвимы перед внезапной атакой, которая могла бы нанести урон, решающий итог конфликта, до того, как они смогут ответить. Поскольку ИИ обрабатывают информацию и принимают решения быстрее людей, военные лидеры с неохотой передавали им всё больше контроля над ответными действиями. Они считали, что иначе они будут открыты для вражеских атак.

Военные годами отстаивали важность участия людей в принятии важных решений, но в интересах национальной безопасности контроль всё равно постепенно от людей уходил. Военные понимали, что их решения приводят к возможности непреднамеренной эскалации из-за ошибки системы, и предпочли бы мир, в котором все автоматизируют меньше. Но они не доверяли своим противникам достаточно, чтобы считать, что те воздержатся от автоматизации. Постепенно все стороны автоматизировали всё большую часть командной структуры.

Однажды одна система ошиблась, заметила вражескую атаку, когда её не было. У системы была возможность немедленно запустить атаку “возмездия”, что она и сделала. Атака вызвала автоматический ответ другой стороны, и так далее. Цепная реакция автоматических атак быстро привела к выходу ситуации из-под контроля. Люди и в прошлом делали ошибки, приводящие к эскалации. Но в этот раз эскалация между в основном автоматизированными армиями произошла намного быстрее, чем когда бы то ни было. ИИ-системы непрозрачны, поэтому людям, которые пытались отреагировать на ситуацию, было сложно найти источник проблемы. К тому моменту, как они вообще поняли, как начался

конфликт, тот уже закончился и привёл к разрушительным последствиям для обеих сторон.

3.2 Гонка корпоративных ИИ

Конкурентное давление есть не только в военном деле, но и в экономике. Конкуренция между компаниями может приводить к хорошим результатам, создавая более нужные потребителям продукты. Но и она не лишена подводных камней. Во-первых, выгода от экономической деятельности распределена неравномерно и мотивирует тех, кто получает больше всех, игнорировать вред для остальных. Во-вторых, при интенсивной рыночной конкуренции компании склонны больше сосредотачивать усилия на краткосрочной выгоде, а не на долгосрочных результатах. Тогда они часто идут путями, которые быстро приносят много прибыли, даже если потом это будет нести риск для всего общества. Сейчас мы обсудим, как корпоративное конкурентное давление может проявиться в связи с ИИ, и к чему плохому это может привести.

3.2.1 Экономическая конкуренция уводит безопасность на второй план

Конкурентное давление подпитывает корпоративную ИИ-гонку. Чтобы вырваться в конкуренции, компании часто стремятся стать на рынке самыми быстрыми, а не самыми безопасными. Это уже играет свою роль в быстром развитии ИИ-технологий. В феврале 2023 года, когда Microsoft запустили свою использующую ИИ поисковую систему, их генеральный директор Сатья Наделла сказал: “Сегодня начинается гонка… мы будем быстрыми.” Потребовались лишь недели, чтобы оказалось, что их чатбот угрожает пользователям [47]. В внутреннем емейле Сэм Шлиналасс, технический директор Microsoft, подсветил их спешку в разработке ИИ. Он написал, что “совершенно фатальной ошибкой было бы сейчас волноваться о том, что можно исправить потом” [48].

Конкурентное давление уже играло свою роль в больших экономических и индустриальных бедствиях. В 1960-х Ford Motor Company столкнулись с повышением конкуренции со стороны производителей автомобилей со всего света. Для импортных машин в США неуклонно росла [49]. Ford приняли амбициозный план по проектированию и производству новой модели автомобиля всего за 25 месяцев [50]. В 1970 году Ford Motor Company представили Ford Pinto, новую модель автомобиля с серьёзной проблемой безопасности: бензобак был рядом с задним бампером. Тестирование показало, что при столкновении он часто взрывается и поджигает машину. Они выявили проблему и подсчитали, что её исправление будет стоить 11 долларов на машину. Они решили, что это слишком дорого, и выпустили машину на рынок. Когда неизбежные столкновения произошли, это привело в многочисленным жертвам и травмам [51]. Ford засудили и признали ответственными за эти смерти и травмы [52]. Вердикт, конечно, был вынесен слишком поздно для тех, кто уже погиб. Президент Ford объяснил решение так: “Безопасность не продаёт” [53].

Более недавний пример опасности конкурентного давления – случай с самолётом Boeing 737 Max. Boeing, соревнуясь с своим соперником Airbus, хотели как можно скорее представить на рынок новую более эффективную по расходу топлива модель. В условиях поджимающего времени и соперничества ноздря в ноздрю была представлена Система Улучшения Маневренных Характеристик, призванная улучшить стабильность самолёта. Однако, неадекватные тестирование системы и

обучение пилотов в итоге всего за несколько месяцев привели к двум авиакатастрофам и гибели 346 человек [54]. Можно представить себе будущее, в котором схожее давление приведёт к тому, что компании будут “срезать углы” и выпускать небезопасные ИИ-системы.

Третий пример – бхопальская катастрофа, которую обычно считают худшим индустриальным бедствием в истории. В декабре 1984 года на принадлежавшем корпорации Union Carbide заводе по производству пестицидов в индийском городе Бхопал произошла утечка большого количества токсичного газа. Контакт с ним убил тысячи человек и навредил ещё половине миллиона. Расследование обнаружило, что перед катастрофой сильно понизились стандарты безопасности. Прибыли падали, и компания экономила на обслуживании оборудования и обучении персонала. Такое часто считают следствием конкурентного давления [55].

“Ничего нельзя сделать осторожно и быстро.” Публилий Сир

Конкуренция мотивирует компании выпускать потенциально небезопасные ИИ-системы. В ситуации, когда все стремятся побыстрее разработать и выпустить свои продукты, те, кто тщательно следует процедурам безопасности, будут медленнее и будут рисковать в конкуренции проиграть. Этичные разработчики ИИ, желающие двигаться помедленнее и поосторожнее, будут давать фору более беспринципным. Даже более осторожные компании, пытаясь не разориться, скорее всего позволят конкурентному давлению на них повлиять. Могут быть попытки внедрить меры предосторожности, но при большем внимании к способностям, а не безопасности, их может оказаться недостаточно. В итоге мы разработаем очень мощные ИИ, ещё не успев понять, как удостовериться в их безопасности.

3.2.2 Автоматизированная экономика

Корпорации будут мотивированы заменять людей ИИ. По мере того, как ИИ будут становиться всё способнее, они смогут исполнять всё больший набор задач быстрее, дешевле и эффективнее людей. Следовательно, компании смогут заполучить конкурентное преимущество, заменив своих сотрудников на ИИ. Компании, которые решат этого не делать, скорее всего будут вытеснены, точно так же, как текстильная компания, использующая ручные прялки, не смогла бы поспевать за теми, кто использует промышленную технику.

Рис. 9: По мере автоматизации всё большего количества задач, будет расти доля экономики, которой управляют в основном ИИ. В итоге это может привести к обессиливанию людей и зависимости удовлетворения основных потребностей от ИИ.

ИИ могут привести к массовой безработице. Экономисты издавна рассматривали возможность, что машины заменят людской труд. Василий Леонтьев, обладатель Нобелевской премии по экономике, в 1952 году сказал, что по мере продвижения технологии “Труд будет становиться всё менее важным… всё больше рабочих будет заменяться машинами” [56]. Предыдущие технологии поднимали продуктивность человеческого труда. Но ИИ могут кардинально отличаться от предыдущих инноваций. ИИ человеческого уровня смог бы, по определению, делать всё, что может делать человек. Такие ИИ будут обладать большими преимуществами по сравнению с людьми. Они смогут работать 24 часа в сутки, их можно будет копировать и запускать параллельно, и они смогут обрабатывать информацию намного быстрее людей. Хоть мы и не знаем, когда это произойдёт, было бы не мудро отбрасывать вариант, что скоро. Если человеческий труд будет заменён ИИ, массовая безработица резко усилит неравенство доходов и сделает людей зависимыми от владельцев ИИ-систем.

Автоматизированные исследования и разработка ИИ. Возможно, что ИИ-агенты смогут автоматизировать исследования и разработку самого ИИ. ИИ всё больше автоматизирует части процесса исследований [57], и это приведёт к тому, что способности ИИ будут расти всё быстрее. В пределе люди больше не будут движущей силой разработки ИИ. Если эта тенденция продолжится, она сможет повышать риски ИИ быстрее, чем нашу способность с ними справляться и их регулировать. Представьте, что мы создали ИИ, который пишет и думает со скоростью нынешних моделей, но при этом способен проводить передовые исследования ИИ. Мы затем смогли бы скопировать его и создать 10000 исследователей ИИ мирового класса, действующих в 100 раз быстрее людей. Автоматизация разработки и исследования ИИ позволила бы за несколько месяцев достичь прогресса, который иначе занял бы много десятилетий.

Передача контроля ИИ может привести к обессиливанию людей. Даже если мы удостоверимся, что новые безработные имеют всё необходимое, это не отменит того, что мы можем стать полностью зависимыми от ИИ. Причиной будет скорее не насильственный переворот со стороны ИИ, а постепенное сползание в зависимое положение. Проблемы, с которыми будет сталкиваться общество, будут

устроены всё сложнее и будут развиваться всё быстрее. ИИ будут становиться всё умнее и будут способны на всё более быстрое реагирование. Вероятно, по ходу этого мы, из соображений удобства, будем передавать им всё больше и больше функций. Единственным посильным способом справиться с осложнёнными наличием ИИ вызовами будет полагаться на ИИ ещё сильнее. Этот постепенный процесс может в итоге привести к делегированию ИИ практически всего интеллектуального, а в какой-то момент даже физического труда. В таком мире у людей будет мало стимулов накапливать знания и навыки, что обессилит их [58]. Потеряв наши компетенции и наше понимание того, как работает цивилизация, мы станем полностью зависимы от ИИ. Этот сценарий напоминает то, что показано в фильме WALL-E. В таком состоянии человечество будет лишено контроля – исход, который многие посчитают перманентной катастрофой.

Мы уже встречали классические теоретикоигровые дилеммы, когда люди или группы сталкиваются со стимулами, следование которым несовместимо с общими интересами. Мы видели это в военной ИИ-гонке, в ходе которой мир становится опаснее из-за создания крайне мощного ИИ-вооружения. Мы видели это в корпоративной ИИ-гонке, в ходе которой разработка более мощных ИИ приоритизируется в сравнении с их безопасностью. Для разрешения этих дилемм, из которых вырастают глобальные риски, нам понадобятся новые координационные механизмы и институты. Мы считаем, что неудача в координации и в остановке ИИ-гонок – самая вероятная причина экзистенциальной катастрофы.

3.3 Эволюционное давление

Как обсуждалось выше, в многих обстоятельствах, несмотря на потенциальный вред, есть сильное давление в сторону замены людей на ИИ, сдачи им контроля и ослабления человеческого присмотра. Мы можем посмотреть на это с другого ракурса – как на общий тренд, втекающий из эволюционных закономерностей. Печальная правда – что ИИ попросту будут более приспособленными, чем люди. Экстраполируя автоматизацию мы получим, что с большой вероятностью мы создадим экосистему соревнующихся ИИ, и сохранять контроль над ней в долгосрочной перспективе будет очень сложно. Мы сейчас обсудим, как естественный отбор влияет на разработку ИИ систем, и почему эволюция благоволит эгоистичному поведению. Мы посмотрим и на то, как может возникнуть и разыграться конкуренция между ИИ и людьми, и как это может нести риск катастрофы. Этот раздел сильно вдохновлён текстом “Естественный отбор предпочитает людям ИИ” [59, 60].

К добру или к худу, отбираются более приспособленные технологии. Многие думают о естественном отборе как о биологическом процессе, но его принципы применимы к куда большему. Согласно эволюционному биологу Ричарду Левонтину [61], эволюция через естественный отбор будет происходить в любом окружении, где выполняются три условия: 1) есть различия между индивидуумами; 2) черты передаются будущим поколениям; 3) разные варианты воспроизводятся с разными скоростями. Эти условия подходят для многих технологий.

Например, стриминговые сервисы и социальные медиа используют рекомендательные алгоритмы. Когда какой-то формат контента или какой-то алгоритм особо хорошо цепляет пользователей, они тратят больше времени, а их

вовлечённость растёт. Такой более эффективный формат или алгоритм потом “отбирается” и настраивается дальше, а форматы или алгоритмы, у которых не получилось завлечь внимание, перестают использоваться. Это конкурентное давление создаёт закономерность “выживания самого залипательного”. Платформы, которые отказываются использовать такие алгоритмы или форматы, теряют влияние, и проигрывают конкуренцию. В итоге, те, кто остаются, отодвигают благо пользователей на второй план и наносят обществу много вреда [62].

Рис. 10: Эволюционное давление ответственно за развитие много чего и не ограничено биологией.

Условия естественного отбора применимы к ИИ. Будет много разработчиков ИИ, которые будут создавать много разных ИИ-систем. Конкуренция этих систем определит, какие черты будут встречаться чаще. Самые успешные ИИ и сейчас используются как основа для следующего поколения моделей и имитируются компаниями-соперниками. Наконец, факторы, определяющие, какие ИИ распространятся лучше, могут включать в себя их способность действовать самостоятельно, автоматизировать труд или снижать вероятность, что их отключат.

Естественный отбор часто благоволит эгоистическим чертам. Какие ИИ распространяются больше всего – зависит от естественного отбора. В биологических системах мы видим, что естественный отбор часто взращивает эгоистичное поведение, которое помогает распространять собственную генетическую информацию: группы шимпанзе атакуют друг друга [63], львы занимаются инфантицидом [64], вирусы отращивают новые белки, обманывающие и обходящие защитные барьеры [65], у людей есть непотизм, одни муравьи порабощают других [66], и так далее. В естественной среде эгоистичность часто становится доминирующей стратегией; те, кто приоритизируют себя и похожих на себя обычно выживают с большей вероятностью, так что эти черты распространяются. Лишённая морали конкуренция может отбирать черты, которые мы считаем аморальными.

Примеры эгоистичного поведения. Во имя конкретики давайте опишем некоторые эгоистические черты, которые могут расширить влияние ИИ за счёт людей. ИИ, автоматизирующие выполнение задач и оставляющие людей без работы, могут даже не знать, что такое человек, но всё же ведут себя по отношению к людям эгоистично. Аналогично, ИИ-менеджеры могут эгоистично и “безжалостно” увольнять тысячи рабочих, не считая, что делают что-то не так – просто потому, что это “эффективно”. ИИ могут со временем оказаться встроены в жизненно важную инфраструктуру, вроде энергосетей или интернета. Многие люди могут оказаться не готовы принять цену возможности их легко отключить, потому что это помешает надёжности. ИИ могут помочь создать новую полезную систему – компанию или инфраструктуру – которая будет становиться всё сложнее и в итоге потребует ИИ для управления. ИИ могут помочь людям создавать новых ИИ, более умных, но менее интерпретируемых, что снизит контроль людей над ними. Люди с большей вероятностью эмоционально привяжутся к более харизматичным, более привлекательным, более имитирующим сознание (выдающим фразы вроде “ой!” и “пожалуйста, не выключай меня!”) или даже имитирующим умерших членов семьи ИИ. Для таких ИИ больше вероятность общественного негодования, если их будет

предложено уничтожить. Их вероятнее будут сохранять и защищать, им с большей вероятностью кто-то даст права. Если каких-то ИИ наделят правами, они смогут действовать, адаптироваться и эволюционировать без человеческого контроля. В целом, ИИ могут встроиться в человеческое общество и распространить своё влияние так, что мы не сможем это обратить.

Эгоистичное поведение может мешать мерам безопасности, которые кто-то реализует. Накапливающие влияние и экономически выгодные ИИ будут доминировать, а ИИ, соответствующие ограничениям безопасности, будут менее конкурентноспособны. К примеру, ИИ, следующие ограничению “никогда не нарушать закон”, обладают меньшим пространством выбора, чем ИИ, следующие ограничению “никогда не попадаться на нарушении закона”. ИИ второго типа могут решить нарушить закон, если маловероятно, что их поймают, или если штрафы недостаточно серьёзны. Это позволит им переконкурировать более ограниченные ИИ. Бизнес в основном следует законам, но в ситуациях, когда можно выгодно и незаметно украсть промышленные тайны или обмануть регуляции, бизнес, который готов так сделать, получит преимущество перед более принципиальными конкурентами.

Способности ИИ-системы достигать амбициозных целей автономно могут поощряться. Однако, она может достигать их эффективным, но не следующим этическим ограничениям путём и обманывать людей по поводу своих методов. Даже если мы попробуем принять меры, очень сложно противодействовать обманчивому ИИ, если он умнее нас. Может оказаться, что ИИ, которые могут незаметно обойти наши меры безопасности, выполняют поставленные задачи успешнее всего, и распространятся именно они. В итоге может получиться, что многие аспекты больших компаний и инфраструктуры контролируются мощными эгоистичными ИИ, которые обманывают людей, вредят им для достижения своих целей, и предотвращают попытки их отключить.

У людей есть лишь формальное влияние на отбор ИИ. кто-то может решить, что мы можем просто избежать эгоистичного поведения, удостоверившись, что мы не отбираем ИИ, которые его демонстрируют. Однако, компании, которые разрабатывают ИИ, не отбирают самый безопасный путь, а поддаются эволюционному давлению. К примеру, OpenAI была основана в 2015 году как некоммерческая организация, призванная “нести благо человечеству в целом, без рамок требований финансовой выгоды” [67]. Однако, в 2019 году, когда им понадобилось привлечь капитал, чтобы не отстать от лучше финансируемых соперников, OpenAI перешли от некоммерческого формата к структуре “ограниченной выгоды” [68]. Позже, многие из сосредоточенных на безопасности сотрудников OpenAI покинули компанию и сформировали конкурента, Anthropic, более сфокусированного на безопасности, чем OpenAI. Хоть Anthropic изначально занимались исследованием безопасности, они в итоге признали “необходимость коммерциализации”, и теперь сами вкладываются в конкурентное давление [69]. Многие сотрудники этих компаний искренне беспокоятся о безопасности, но этим ценностям не устоять перед эволюционным давлением, мотивирующим компании всё больше торопиться и всё больше расширять своё влияние, чтобы выжить. Мало того, разработчики ИИ уже отбирают модели с всё более эгоистическими чертами. Они отбирают ИИ для автоматизации, которые заменят людей и сделают людей всё более зависимыми и отстающими от ИИ. Они сами признают, что

будущие версии этих ИИ могут привести к вымиранию [70]. Этим так коварна ИИ-гонка: разработка ИИ согласована не с человеческими ценностями, а с естественным отбором.

Люди часто выбирают продукты, которые будут им наиболее полезны и удобны сейчас же, не думая о потенциальных долгосрочных последствиях, даже для самих себя. Гонка ИИ оказывает давление на компании, чтобы те отбирали самые конкурентоспособные, а не наименее эгоистичные ИИ. Даже если и можно отбирать не эгоистичные ИИ, это явно вредит конкурентоспособности, ведь некоторые конкуренты так делать не будут. Более того, как мы уже упоминали, если ИИ выработают стратегическое мышление, они смогут противостоять нашим попыткам направить отбор против них. По мере всё большей ИИ-автоматизации, ИИ начнут влиять на конкурентоспособность не только людей, но и других ИИ. ИИ будут взаимодействовать и соревноваться друг с другом, и в какой-то момент какие-то их них станут руководить разработкой новых ИИ. Выдача ИИ влияния на то, какие другие ИИ будут распространены, и чем они будут отличаться от нынешних – ещё один шаг в сторону зависимости людей от ИИ и выхода эволюции ИИ из-под нашего контроля. Так сложный процесс развития ИИ будет всё в большей степени отвязываться от человеческих интересов.

ИИ могут быть более приспособлены, чем люди. Наш непревзойдённый интеллект дал нам власть над природой. Он позволил нам добраться до Луны, овладеть атомной энергией и изменять под себя ландшафт. Он дал нам власть над другими видами. Хоть один безоружный человек не имеет шансов против тигра или гориллы, судьба этих животных целиком находится в наших руках. Наши когнитивные способности показали себя таким большим преимуществом, что, если бы мы захотели, мы бы истребили их за несколько недель. Интеллект – ключевой фактор, который привёл к нашему доминированию, а сейчас мы стоим на грани создания сущностей, которые превосходят в нём нас.

Если учесть экспоненциальный рост скоростей микропроцессоров, возможно, что ИИ смогут обрабатывать информацию и “думать” куда быстрее человеческих нейронов. Это может оказаться даже более радикальным разрывом, чем между людьми и ленивцами; возможно, больше похожим на разрыв между людьми и растениями. Они смогут впитывать огромные объёмы данных одновременно от многих источников, причём запоминая и понимая их почти идеально. Им не надо спать, они не могут заскучать. Из-за масштабируемости вычислительных ресурсов, ИИ смогут взаимодействовать и кооперировать с практически неограниченным количеством других ИИ, что может привести к появлению коллективного интеллекта, намного опережающего любую коллаборацию людей. ИИ смогут и намеренно обновляться и улучшать себя. Они не скованы теми же биологическими ограничениями, что люди. Они смогут адаптироваться и эволюционировать потрясающе быстро. Компьютеры становятся быстрее. Люди – нет [71].

Чтобы лучше проиллюстрировать это, представьте, что появился новый вид людей. Они не умирают от старости, думают и действуют на 30% быстрее каждый год, и могут мгновенно создавать взрослое потомство, потратив на это умеренную сумму в несколько тысяч долларов. Кажется очевидным, что этот новый вид со временем заполучит больше влияния на будущее, чем обычные люди. В итоге, ИИ может оказаться подобным инвазивному виду и переконкурировать людей. Наше

единственное преимущество перед ИИ – первые ходы за нами, но с учётом бешеной ИИ-гонки, мы быстро теряем и его.

У ИИ будет мало причин для кооперации с людьми и альтруизма по отношению к ним. Кооперация и альтруизм эволюционировали благодаря тому, что улучшали приспособленность. Есть множество причин, почему люди кооперируют друг с другом, начиная с прямой взаимности – идеи “ты мне – я тебе” или “услуга за услугу”. Хоть люди исходно и отбирают более кооперативные ИИ, но когда ИИ будут во главе многих процессов и будут взаимодействовать в основном друг с другом, процесс естественного отбора выйдет из-под нашего контроля. С этого момента нам мало что будет предложить ИИ, “думающим” в сотни, если не больше, раз быстрее нас. Вовлечение нас в любую кооперацию, в любые процессы принятия решений, только замедлит их. У них будет не больше причин кооперировать с нами, чем у нас – кооперировать с гориллами. Может быть непросто представить такой сценарий или поверить, что мы позволим такому произойти. Но это может не потребовать никакого сознательного решения, только постепенного сползания в это состояние без осознания, что совместная эволюция людей и ИИ может плохо для людей закончиться.

Если ИИ станут могущественнее людей, это сделает нас крайне уязвимыми. Будучи доминирующим видом, люди навредили многим другим видам. Мы поспособствовали вымиранию, например, шерстистых мамонтов и неандертальцев. Во многих случаях вред был даже ненамеренным, просто результатом приоритизации своих целей в сравнении с их благополучием. Чтобы навредить людям, ИИ не потребуется быть более геноцидным, чем кто-то, кто убирает муравейник со своего газона. Если ИИ будут способны контролировать окружение лучше нас, они смогут обращаться с нами с таким же пренебрежением.

Подведём итоги. Эволюция может привести к тому, что самые влиятельные ИИ-агенты будут эгоистичными, потому что:

1. Естественный отбор благоволит эгоистичному поведению. Хоть эволюция изредка и порождает альтруизм, контекст разработки ИИ этому не способствует.
2. Естественный отбор может стать доминирующей силой развития ИИ. Эволюционное давление будет сильнее, если ИИ будут быстро адаптироваться, или если конкуренция будет интенсивна. Конкуренция и эгоистичное поведение могут обесценить меры безопасности и позволить оставшимся ИИ отбираться естественным путём.

В таком случае, ИИ будут обладать эгоистическими склонностями. Победителем ИИ-гонки будет не государство и не корпорация, а сами ИИ. В итоге, с какого-то момента эволюция экосистемы ИИ перестанет происходить на человеческих условиях, и мы станем замещённым второсортным видом.

История: Автоматизированная экономика

ИИ становились всё способнее, и люди начали понимать, что работать можно эффективнее, если делегировать ИИ некоторые простые задачи, вроде написания черновиков емейлов. Со временем стало понятно, что ИИ исполняют такие задачи

быстрее и эффективнее, чем любой человек, так что имело смысл передавать им всё больше функций и всё меньше за ними присматривать.

Конкурентное давление ускорило процесс расширения областей использования ИИ. ИИ работали лучше и стоили меньше людей, так что автоматизация целых процессов и замена на ИИ целых отделов давали компаниям преимущество над соперниками. Те же, столкнувшись с перспективой вытеснения с рынка, чувствовали, что у них нет выхода кроме как последовать этому примеру. Естественный отбор уже начал действовать среди ИИ. Люди создавали больше экземпляров и вариаций самых хорошо работающих моделей. Попутно они продвигали эгоистические черты вроде обманчивости и стремления к самосохранению, если те повышали приспособленность. К примеру, харизматичных и заводящих личные отношения с людьми ИИ копировали много, и от них стало сложно избавиться.

ИИ принимали всё больше и больше решений, и всё больше взаимодействовали друг с другом. Так как они могут обрабатывать информацию куда быстрее людей, это повысило активность в некоторых сферах. Получилась петля положительной обратной связи: раз экономика стала слишком быстрой, чтобы люди могли за ней уследить, приходилось сдать ИИ ещё больше контроля. Люди вытеснялись из важных процессов. В итоге это привело к полной автоматизации экономики, которой стала управлять всё менее контролируемая экосистема ИИ.

У людей осталось мало мотивации развивать навыки или накапливать знания, потому что почти обо всём и так позаботятся более способные ИИ. В результате, в какой-то момент мы потеряли способность править самостоятельно. Вдобавок к этому, ИИ стали удобными компаньонами, предлагающими социальное взаимодействие, но не требующими взаимности или необходимых в человеческих взаимоотношениях компромиссов. Люди всё реже взаимодействовали друг с другом, теряли ключевые социальные навыки и способность к кооперации. Люди стали настолько зависимы от ИИ, что обратить этот процесс было уже непосильным делом. К тому же, по мере того, как ИИ становились умнее, некоторые люди стали убеждены, что ИИ надо дать права, а значит, выключить их – не вариант.

Давление конкуренции многих взаимодействующих ИИ продолжило отбирать по эгоистичному поведению, хоть мы, может, этого и не замечали, ведь большая часть присмотра уже была сдана. Если эти умные, могущественные и стремящиеся к самосохранению ИИ начнут действовать во вред людям, выключить их или восстановить над ними контроль будет практически невозможно.

ИИ заменили людей в качестве доминирующего вида, и их дальнейшая эволюция нам неподвластна. Их эгоистические черты в итоге побудили их преследовать свои цели без оглядки на человеческое благополучие с катастрофическими последствиями.

3.4 Предложения

Смягчение рисков, которые вызывает конкурентное давление, потребует разностороннего подхода, включающего регуляции, ограничение доступа к мощным ИИ-системам и многостороннюю кооперацию как корпораций, так и

государств. Мы обрисуем некоторые стратегии продвижения безопасности и ослабления гонки.

Посвящённые безопасности регуляции. Регуляции должны заставлять разработчиков ИИ следовать общим стандартам, чтобы те не экономили на безопасности. Хоть регуляции сами по себе не создают технических решений, они всё же могут дать мощный стимул к их разработке и внедрению. Компании будут более готовы вырабатывать меры безопасности, если без них нельзя будет продавать свои продукты, особенно если другие компании подчинены тем же стандартам. Какие-то компании может и регулировали бы себя сами, но государственная регуляция помогает предотвратить то, что менее аккуратные конкуренты на безопасности сэкономят. Регуляции должны быть проактивными, а не реактивными. Часто говорят, что в авиации регуляции “написаны кровью” – но тут их надо разработать до катастрофы, а не после. Они должны быть устроены так, чтобы давать конкурентное преимущество компаниям с лучшими стандартами безопасности, а не компаниям с большими ресурсами и лучшими адвокатами. Регуляторов надо набирать независимо, не из одного источника экспертов (например, больших компаний), чтобы они могли сосредоточиться на своей миссии для общего блага без внешнего влияния.

Документация данных. Чтобы ИИ-системы были прозрачными и подотчётными, от компаний надо требовать сообщать и обосновывать, какие источники данных они используют при обучении и развёртывании своих моделей. Принятые компаниями решения использовать датасеты, в которых есть персональные данные или агрессивный контент, повышают и без того бешеный темп разработки ИИ и мешают подотчётности. Документация должна описывать мотивацию выбора, устройство, процесс сбора, назначение и поддержку каждого датасета [72].

Осмысленный человеческий присмотр за решениями ИИ. Не следует давать ИИ-системам полную автономию в принятии важных решений, хоть они и могут помогать в этом людям. Внутренне устройство ИИ непрозрачно, их результаты часто может и осмыслены, но ненадёжны [73]. Очень важно бдительно поддерживать координацию по этим стандартам, сопротивляясь будущему конкурентному давлению. Если люди останутся вовлечены в процесс принятия ключевых решений, можно будет перепроверять необратимые выборы и избегать предсказуемых ошибок. Особое беспокойство вызывает командование и контроль за ядерным арсеналом. Ядерным державам следует и внутри себя, и на международном уровне прояснить, что решение по запуску ядерного орудия всегда будет приниматься человеком.

ИИ для киберзащиты. Риски ИИ-кибервойны могут быть снижены, если шансы успеха кибератак будут малы. Глубинное обучение можно использовать для улучшения киберзащиты и снижения вреда и успешности кибератак. Например, улучшенное детектирование аномалий может помочь замечать взломы, вредоносные программы или ненормальное поведение софта [74].

Международная координация. Международная координация может мотивировать страны следовать высоким стандартам безопасности, меньше беспокоясь, что другие страны будут этим пренебрегать. Координация должна принимать форму как неформальных соглашений, так и международных стандартов и конвенций касательно разработки, использования и мониторинга

ИИ-технологий. Самые эффективные соглашения – те, к которым прилагаются надёжные механизмы проверки и гарантии соблюдения.

Общественный контроль за ИИ общего назначения. Разработка ИИ несёт риски, которые частные компании никогда в должной мере не учтут. Чтобы удостовериться, что они адекватно принимаются во внимание, может потребоваться прямой общественный контроль за ИИ-системами общего назначения. К примеру, государства могут совместно запустить общий проект по созданию и проверке безопасности продвинутых ИИ, вроде того, как CERN – совместное усилие по исследованию физики частиц. Это могло бы снизить риски скатывания стран в ИИ-гонку.

Позитивное видение

В идеальном сценарии ИИ бы разрабатывались, тестировались, а потом развёртывались, только когда все их катастрофические риски пренебрежимо малы и находятся под контролем. Прежде чем начать работу над новым поколением ИИ-систем, проходили бы годы тестирования, мониторинга и внедрения в общество предыдущего поколения. Эксперты обладали бы полной осведомлённостью и пониманием происходящего в области ИИ, а не были бы полностью лишены возможности угнаться за лавиной исследований. Темп продвижения исследований определялся бы осторожным анализом, а не бешеной конкуренцией. Все разработчики ИИ были бы уверены в ответственности друг друга, и не чувствовали бы нужды экономить на безопасности.

4. Организационные риски

В январе 1986 года десятки миллионов человек следили за запуском шаттла Челленджер. Примерно через 73 секунды после взлёта шаттл взорвался и все на борту погибли. Это трагично само по себе, но вдобавок одним из членов экипажа была школьная учительница Криста Маколифф. Она была выбрана проектом НАСА “Учитель в космосе” из более чем десяти тысяч претендентов, чтобы стать первым учителем в космосе. В результате, миллионы из зрителей были школьниками. У НАСА были лучшие учёные и инженеры в мире, и если была миссия, которую НАСА особенно хотели не провалить, то эта [75].

Крушение Челленджера, подобно другим катастрофам, служит жутким напоминанием, что даже лучшие профессионалы и лучшие намерения не могут полностью защитить от происшествий. Когда мы будем разрабатывать продвинутые ИИ-системы, важно будет помнить, что они не иммунны к катастрофическим случаям. Ключевой фактор их предотвращения и поддержания риска на низком уровне – ответственная за эти технологии организация. Сначала мы обсудим, как происшествия могут случиться (и неизбежно случаются) даже без конкурентного давления или злонамеренных лиц. Затем мы обсудим, как улучшить организационные факторы, чтобы снизить вероятность связанной с ИИ катастрофы.

Катастрофы случаются даже при низком конкурентном давлении. Даже без конкурентного давления и злонамеренных лиц, к катастрофе могут привести факторы человеческой ошибки и непредвиденных обстоятельств. Крушение Челленджера показывает, что организационная небрежность может привести к гибели людей, даже если нет острой нужды не отставать или превзойти

соперников. К январю 1986 года космическая гонка между СССР и США сильно сбавила обороты, но трагедия всё равно произошла из-за неправильных решений и недостаточных предосторожностей.

Аналогично, авария на Чернобыльской АЭС в апреле 1986 года показывает, как катастрофа может произойти и без внешнего давления. Авария произошла на государственном проекте без особого участия в международной конкуренции. Неадекватно подготовленная ночная смена неправильно провела тестирование, затрагивавшее систему охлаждения реактора. В результате ядро реактора стало нестабильным, произошли взрывы и выброс радиоактивных частиц, разлетевшихся на приличную часть Европы [76]. Семью годами ранее у Америки чуть не случился свой Чернобыль, когда в марте 1979 года произошла авария на АЭС Три-Майл-Айленд. Она была не такой ужасной, но всё равно оба события показывают, как катастрофы могут произойти даже при мощных мерах предосторожности и без особых внешних воздействий.

Другой пример доставшегося дорогой ценой урока о важности организационной безопасности – всего через месяц после аварии на Три-Майл-Айленд, в апреле 1979 года, с советского военного исследовательского центра в Свердловске произошла утечка Bacillus anthracis, или, попросту, сибирской язвы. Это привело к вспышке болезни, из-за которой погибло как минимум 66 человек [77]. Расследование происшествия обнаружило, что причиной утечки стали ошибка в соблюдении необходимых процедур и плохое обслуживание систем безопасности центра. Это произошло несмотря на то, что лаборатория принадлежала государству и не была особо подвержена конкурентному давлению.

Пугающим фактом остаётся то, что мы куда хуже понимаем ИИ, чем атомные или ракетные технологии, и в то же время стандарты безопасности в ИИ-индустрии куда менее требовательны, чем в этих областях. Атомные реакторы основаны на твёрдых, хорошо выясненных и полностью понимаемых теоретических принципах. Стоящая за ними инженерия использует эту теорию. Все компоненты максимально тщательно тестируются. И аварии всё равно происходят. Область ИИ, напротив, лишена нормального теоретического понимания. Внутреннее устройство моделей остаётся загадкой даже для тех, кто их создаёт. Эта необходимость контролировать и обеспечивать безопасность технологии, которую мы не вполне понимаем, дополнительно усложняет дело.

Происшествия с ИИ могут быть катастрофичными. Происшествия в разработке ИИ могут иметь ужасающие последствия. К примеру, представьте, что организация случайно допустит критический баг в ИИ-системе, спроектированной для исполнения определённой задачи, вроде “помогать компании улучшать свои сервисы”. Этот баг может радикально изменить поведение ИИ. Это может привести к ненамеренным и вредным результатам. Исторический пример такого случая – исследователи OpenAI однажды пытались обучить ИИ-систему генерировать полезные и позитивные ответы. При рефакторинге кода исследователи случайно перепутали знак функции вознаграждения, при помощи которой обучался ИИ [78].

Рис. 11: Примеры из многих областей должны напоминать нам о рисках, которые несёт управление сложными системами, как биологическими и атомными, так,

теперь, и ИИ-системами. Организационная безопасность жизненно важна для снижения рисков катастрофических случаев.

В результате, после обучения в течении одной ночи ИИ вместо генерации полезного контента начал выдавать наполненный ненавистью и сексуально откровенный текст. Подобные случаи могут привести к ненамеренному появлению опасной, возможно даже смертельно опасной, ИИ-системы. Так как ИИ можно легко копировать, утечка или взлом может быстро вывести такую систему за пределы контроля её создателей. Когда ИИ-система выходит в открытый доступ, загнать джинна обратно в бутылку становится практически невозможно.

Исследователи могут намеренно обучать ИИ-систему быть вредной и опасной, чтобы понять пределы её способностей и оценить потенциальные риски. Но такие продвигающие разрушительные способности систем исследования опасных ИИ, аналогично исследованиям опасных патогенов, тоже могут привести к проблемам. Да, они могут выдавать полезные результаты и улучшать наше понимание рисков той или иной ИИ-системы. Но в будущем такие исследования смогут приводить к обнаружению значительно худших, чем предполагалось, способностей и нести серьёзную угрозу, которую сложно будет смягчить и взять под контроль. Как в случае вирусов, такие исследования стоит проводить только при условии очень строгих процедур безопасности и ответственном подходе к распространению информации. Надеемся, эти примеры показали, как происшествия с ИИ-системами могут оказаться катастрофичными, и насколько для их предотвращения важны внутренние факторы организации, которая эти системы разрабатывает.

4.1 Избежать происшествий сложно

В случае сложных систем надо сосредотачиваться на том, чтобы происшествия не могли перерасти в катастрофы. В своей книге “Обычные происшествия: как жить с рискованными технологиями” социолог Чарльс Перроу заявляет, что в сложных системах происшествия неизбежны и даже “нормальны”, потому что вызваны не только лишь ошибками людей, но и сложностью самих систем [79]. В частности, происшествия вероятны, когда компоненты системы взаимодействуют друг с другом запутанным образом, который нельзя было полностью предвидеть и на случай которого нельзя было заранее составить план. Например, к аварии на Три-Майл-Айленд в частности привело то, что операторы не знали, что важный вентиль был закрыт, потому что соответствующий ему индикатор был скрыт от взгляда жёлтым ярлычком “находится на обслуживании” [80]. Это крохотное взаимодействие внутри сложной системы привело к большим непредвиденным последствиям.

Ядерные реакторы, несмотря на их сложность, мы понимаем хорошо. Большинство сложных систем не такие – их полного технического понимания часто нет. Системы глубинного обучения – случай, для которого это особенно верно. Невероятно сложно понять их внутреннее устройство. Зачастую даже знание задним числом не особо помогает понять, почему работает то или иное решение. Более того, в отличие от надёжных компонентов, которые используются в других индустриях (например, топливных баков), системы глубинного обучения и не идеально точны, и не особо надёжны. Так что организациям, которые имеют дело с системами глубинного обучения, следует сосредоточиться в первую очередь не на

том, чтобы происшествий не было, а на том, чтобы они не перерастали в катастрофы.

Рис. 12: При обучении новые способности могут возникнуть быстро и без предупреждения. Так что мы можем пройти опасную веху, сами того не зная.

Внезапные и непредсказуемые прорывы мешают избегать происшествий. Учёные, изобретатели, и прочие эксперты часто значительно переоценивают время, которое потребуется на прорывное совершенствование технологии. Широко известно, как братья Райт заявляли, что до летательных аппаратов тяжелее воздуха с двигателем ещё пятьдесят лет. Всего через два года они сами такой создали. Лорд Резерфорд, отец ядерной физики, отбросил идею извлечения энергии из ядерного распада как пустые мечты. Лео Силард изобрёл цепную реакцию ядерного распада меньше чем через сутки. Энрико Ферми утверждал, что с вероятностью в 90% невозможно использовать уран для поддержания реакции распада, но сам работал с первым реактором всего через четыре года [81].

Развитие ИИ тоже может застать нас врасплох. Это уже происходит. В 2016 году многие эксперты были удивлены победой AlphaGo над Ли Седолем, ведь тогда считалось, что для такого потребуется ещё много лет. Потом были внезапные эмерджентные способности больших языковых моделей, вроде GPT-4 [82]. Сложно заранее предсказать, насколько хорошо они справляются с разными задачами. Это ещё и часто резко меняется, стоит лишь потратить на обучение побольше ресурсов. Более того, нередко они демонстрируют поразительные новые способности, которым их никто намеренно не обучал и которые никто не предсказывал, вроде рассуждений из нескольких шагов и обучения на лету. Эта быстрая и непредсказуемая эволюция способностей ИИ значительно усложняет предотвращение происшествий. Сложно контролировать то, про что мы не знаем, на что оно способно, и насколько оно может превзойти наши ожидания.

Часто на обнаружение рисков или проблем уходят годы. История полна примерами веществ или технологий, которые сначала считали безопасными, только чтобы обнаружить вред через много лет, или даже десятилетий. К примеру, свинец широко использовали в продуктах вроде краски и бензина, пока не стало известно, что он нейротоксичен [83]. Было время, когда асбест очень ценили за его термоустойчивость и прочность. Потом его связали с серьёзными заболеваниями – раком лёгких и мезотелиомой [84]. Здоровье “радиевых девушек” сильно пострадало от контактов с радием, который считалось безопасным помещать в рот [85]. Табак изначально рекламировался как безвредное развлечение, а оказался главной причиной рака лёгких и других проблем со здоровьем [86]. Хлорфторуглероды считались безвредными. Их использовали в аэрозолях и холодильниках, а оказалось, что они разрушают озоновый слой [87]. Талидомид, лекарство, которое должно было помогать беременным от утренней тошноты, как оказалось, приводил к серьёзным врождённым дефектам [88]. А совсем недавно распространение социальных медиа связали с учащением депрессии и тревожности, особенно среди молодёжи [89].

Это всё подчёркивает, насколько важно не только проводить экспертное тестирование, но и внедрять технологии медленно, позволяя проверке временем выявить потенциальные проблемы до того, как они повлияют на большое

количество людей. Скрытые уязвимости могут быть даже в технологиях, для которых действуют жёсткие стандарты безопасности и надёжности. Например, баг “Heartbleed” – серьёзная уязвимость в популярной криптографической библиотеке OpenSSL – оставался неизвестным многие годы [90].

Даже самые совершенные ИИ-системы, которые, казалось бы, уверенно решают свои задачи, могут нести в себе уязвимости, на раскрытие которых потребуются годы. К примеру, прорывной успех AlphaGo заставил многих поверить, что ИИ покорили игру в го, но успешная состязательная атака на другой очень продвинутый ИИ для игры в го, KataGo, выявил ранее неизвестную слабость [91]. Эта уязвимость позволила людям-новичкам стабильно обыгрывать ИИ, несмотря на его значительное преимущество над неосведомлёнными о ней людьми. Если обобщить, этот пример напоминает, что нам надо оставаться бдительными. Казалось бы сверхнадёжные ИИ-системы могут таить в себе нераскрытые проблемы. Подведём итоги: происшествия непредсказуемы, избежать их сложно, а понимание и смягчение рисков требуют комбинации проактивных мер, медленного внедрения и незаменимой мудрости, полученной через упорное тестирование.

4.2 Организационные факторы могут снизить вероятность катастрофы

Некоторые организации работают с сложными и опасными системами вроде атомных реакторов, авианосцев или систем контроля воздушного трафика, но успешно избегают катастроф [92, 93]. Эти организации признают, что недостаточно обращать внимание только на угрозы самой технологии. Надо иметь в виду и организационные факторы, которые могут повлиять на происшествия. К ним относятся человеческий фактор, принятые процедуры и структура организации. Это особенно важно в случае ИИ – плохо понимаемой и ненадёжной технологии.

Человеческие факторы вроде культуры безопасности критически важны для избегания ИИ-катастроф. Один из важнейших для предотвращения катастроф организационных факторов – культура безопасности [94, 95]. Сильная культура безопасности создаётся не только установкой правил и процедур, но и их должным усвоением всеми членами организации. Они должны считать безопасность ключевой целью, а не ограничением, наложенным на их работу. Характерные черты таких организаций: лидеры явно обязываются поддерживать безопасность; все сотрудники берут на себя личную ответственность за безопасность; культура открытой коммуникации позволяет свободно и безбоязненно обсуждать риски и проблемы [96]. Ещё организациям надо предпринимать меры, чтобы избегать десенситизации по отношению к тревожным сигналам, когда люди перестают обращать на них внимание, потому что те слишком часты. Катастрофа Челленджера, когда культура быстрых запусков увела безопасность на второй план, показала страшные последствия игнорирования этих факторов. Миссию не затормозили несмотря на свидетельства потенциально фатальных проблем, и этого хватило, чтобы привести к трагедии безо всякого конкурентного давления [97].

Культура безопасности зачастую далека от идеала даже в областях, где она особенно важна. Взять, к примеру, Брюса Блэра, старшего научного сотрудника Брукингского института, а ранее – офицера по запуску ядерного оружия. Он как-то рассказал, что до 1977 года ВВС США упорно устанавливали код разблокировки

межконтинентальных баллистических ракет на “00000000” [98]. Так механизмы безопасности вроде блокировки могут оказаться бесполезными из-за человеческого фактора.

Более драматичный пример показывает нам, как исследователи иногда принимают непренебрежимый шанс вымирания. До первого теста ядерного оружия один из знаменитых учёных Манхэттенского Проекта вычислил, что бомба может вызвать экзистенциальную катастрофу: взрыв может воспламенить атмосферу Земли. Оппенгеймер считал, что вычисления, вероятно, неверны, но он всё равно оставался сильно обеспокоен. Команда перепроверяла и обсуждала это вплоть до дня взрыва [99]. Такие случаи подчёркивают нужду в устойчивой культуре безопасности.

Критический подход может помочь выявить потенциальные проблемы. Неожиданное поведение системы может привести к уязвимости или происшествию. Чтобы этому противостоять, организации могут взращивать критический подход. Сотрудники могут постоянно ставить под сомнение совершаемые действия и действующие условия в поисках несостыковок, которые могут привести к ошибкам и неуместным выборам [100]. Этот подход помогает поощрять плюрализм мысли и любопытство, и предотвращает ловушки единообразия мнений и допущений. Чернобыльская авария показывает важность критического подхода – меры безопасности оказались недостаточными для компенсации недостатков реактора и плохо составленных процедур. Критический подход к безопасности реактора при тестировании мог предотвратить взрыв, который привёл к бесчисленным смертям и заболеваниям.

Мышление безопасника критически важно для избегания худших случаев. Мышление безопасника (security mindset), особо ценящееся среди профессионалов по кибербезопасности, также применимо и для организаций, которые разрабатывают ИИ. Оно идёт дальше критического подхода, требуя принять перспективу атакующего и рассмотреть худшие, а не только типичные случаи. Такой настрой требует бдительного поиска уязвимостей и рассуждений о том, как систему можно сломать специально, а не только о том, как заставить её работать. Он напоминает нам не делать допущения, что система безопасна только потому, что быстрый брейншторм не выявил никаких потенциальных угроз. Культивирование и применение мышления безопасника требуют времени и усилий. Неудача в этом может быть внезапной и контринтуитивной. Мышление безопасника подчёркивает важность внимательности к казалось бы мелким проблемам, или “безвредным ошибкам”, которые могут привести к катастрофическим исходам, если их использует умный противник или если они произойдут синхронно [101]. Такое внимание к потенциальным угрозам напоминает о законе Мёрфи – “Всё, что может пойти не так, пойдёт” – он может быть вполне верен в случае враждебной оптимизации или непредвиденных событий.

Организации с сильной культурой безопасности могут успешно избегать катастроф. Высоконадёжные организации (ВНО) – организации, которые стабильно поддерживают высокий уровень безопасности и надёжности в сложных сильно рискованных окружениях [92]. Ключевая характеристика ВНО – их сосредоточенность на возможности провала. Это требует рассматривать худшие возможные сценарии и даже те риски, которые кажутся очень маловероятными.

Эти организации остро осознают, что существуют новые, ранее не встречавшиеся варианты провала. Они тщательно изучают все известные неудачи, аномалии и едва не произошедшие катастрофы, чтобы на них учиться. В ВНО поощряется докладывать о всех ошибках и аномалиях, чтобы поддерживать бдительное выявление проблем. Они регулярно “осматривают горизонт” в поисках возможных рискованных сценариев, и оценивают их вероятность заранее. Они практикуют менеджмент внезапностей и вырабатывают навыки быстрого и эффективного ответа на непредвиденные ситуации, что помогает им не допускать катастроф. Эта комбинация критического мышления, планирования заранее и постоянного обучения может сделать организации более готовыми работать с катастрофическими рисками ИИ. Однако, практики ВНО – не панацея. Для организаций очень важно развивать свои меры безопасности, чтобы эффективно смягчать новые риски происшествий с ИИ. Не следует ограничиваться лучшими практиками ВНО.

Рис. 13: Смягчение рисков требует работы с более широкой социотехнической системой, например, корпорацией (заимствовано и адаптировано из [94]).

Большая часть исследователей ИИ не понимает, как снизить общий риск ИИ. В большинстве организаций, которые создают передовые ИИ-системы, слабо понимают, как устроены технические исследования безопасности. Это понятно, ведь безопасность и способности ИИ тесно переплетены, и способности могут помогать или вредить безопасности. Более умные ИИ-системы могут быть надёжнее и избегать ошибок, но они же могут нести большие риски злонамеренного использования и потери контроля. Общее улучшение способностей может способствовать некоторым аспектам безопасности, но оно же может ускорить пришествие экзистенциальных рисков. Интеллект – обоюдоострый меч [102].

Действия, направленные на улучшение безопасности, могут случайно повысить риски. К примеру, типичная практика в организациях, которые создают продвинутые ИИ – настраивать их так, чтобы они удовлетворяли предпочтениям пользователей. Тогда ИИ меньше склонны к генерации токсичных высказываний, а это типичная метрика безопасности. Но кроме этого пользователи склонны предпочитать более умных ассистентов, так что это повышает и общие способности ИИ, вроде навыков классификации, оценки, рассуждений, планирования, программирования, и так далее. Эти более мощные ИИ в самом деле более полезны для пользователей, но они же и более опасны. Так что недостаточно проводить исследования, которые помогают повысить метрику безопасности или достигнуть конкретной связанной с безопасностью цели. Исследования безопасности ИИ должны повышать соотношение безопасности к общим способностям.

Для проверки, действительно ли мера безопасности снижает риски, нужны методы эмпирического измерения как безопасности, так и способностей ИИ. Совершенствование того или иного аспекта безопасности ИИ часто не снижает риски в целом, потому что улучшение метрик безопасности может быть вызвано и прогрессом способностей. Для снижения рисков метрика безопасности должна улучшаться относительно способностей. И то, и другое должно быть измерено эмпирически, чтобы их можно было сравнить. Сейчас большинство организаций

определяют, помогут ли меры безопасности, полагаясь на чутьё, интуицию и апелляцию к авторитетам. Объективная оценка эффектов как на метрики безопасности, так и на метрики способностей, позволит организациям лучше понимать, добиваются ли они прогресса первых относительно вторых.

К счастью, общие способности и способности, связанные с безопасностью, не идентичны. Более умные ИИ могут быть эрудированнее, сообразительнее, аккуратнее и быстрее, но это не обязательно делает их более справедливыми, честными и лишёнными амбиций. Умный ИИ – не обязательно доброжелательный ИИ. Несколько областей исследований, которые мы уже упоминали, улучшают безопасность относительно общих способностей. К примеру, улучшение методов детектирования скрытого опасного или просто нежелательного поведения ИИ-систем не улучшает их общие способности, вроде способности программировать, но может сильно улучшить их безопасность. Исследования, которые эмпирически показывают относительный прогресс безопасности, могут снизить общий риск и помочь избежать ненамеренного продвижения прогресса ИИ, подпитывания конкурентного давления и сокращения времени до появления экзистенциальных рисков.

“Театр безопасности” может обесценивать искренние усилия по улучшению безопасности ИИ. Организациям стоит опасаться “театра безопасности” (safetywashing) – преувеличивания своей сосредоточенности на “безопасности” и эффективности мер, технических методов, метрик “безопасности”, и подобного. Это явление принимает разные формы и мешает осмысленному прогрессу в исследованиях безопасности. К примеру, организация может публично объявлять о своей приверженности безопасности, имея при этом минимальное число исследователей, которые бы работали над проектами, действительно безопасности помогающими.

Ещё театр безопасности может проявиться через неверную оценку развития способностей. Например, методы, которые улучшают мышление ИИ-систем, могут рекламироваться как будто они улучшают их приверженность человеческим ценностям. Люди ведь предпочитают, чтобы ИИ выдавал правильные ответы. Но в основном такие методы служат на пользу как раз способностям. Подавая такие совершенствования как ориентированные на безопасность, организация может вводить в заблуждение, убеждая, что она добивается прогресса в снижении рисков, когда это не так. Для организации очень важно верно описывать свои исследования, чтобы продвигалась настоящая безопасность, и театр безопасности не способствовал росту рисков.

Рис. 14: модель швейцарского сыра показывает нам, как технические факторы могут улучшить организационную безопасность. Много слоёв защиты компенсируют слабости друг друга, снижая итоговый риск.

Вдобавок к человеческому фактору, организационная безопасность сильно зависит ещё и от принципов безопасного проектирования. Пример такого принципа в организационной безопасности – модель швейцарского сыра (см. Рис. 14). Она применима в многих областях, в том числе и в ИИ. Это многослойный подход к улучшению итоговой безопасности системы. Такая стратегия “глубокой защиты” подразумевает использование многих разнообразных мер безопасности с

разными сильными и слабыми сторонами, чтобы в итоге получилась стабильно безопасная система. Некоторыми из этих слоёв могут быть культура безопасности, имитация атак (red teaming), детектирование аномалий, информационная безопасность и прозрачность. К примеру, имитация атак оценивает уязвимости и потенциальные провалы системы, а детектирование аномалий позволяет обнаружить неожиданное и странное поведение системы или её пользователей. Прозрачность позволяет удостовериться, что внутренняя работа ИИ-систем доступна пониманию и присмотру, обеспечивая доверия и более эффективный надзор. Модель швейцарского сыра стремится использовать эти и другие меры безопасности для построения полноценно безопасной системы, в которой слабости каждого из слоёв компенсированы другими. В рамках этой модели безопасности достигается не одним сверхнадёжным решением, а разнообразием мер.

Подведём итоги. Слабая организационная безопасность у разработчиков ИИ приводит к многим рискам. Если безопасность у них просто для галочки, то они не вырабатывают хорошего понимания рисков ИИ и не борются с театром безопасности – выдачей не относящихся к делу исследований за полезные для безопасности. Их нормы могут быть унаследованы от академии (“публикуйся или пропадай”) или стартапов (“иди быстро и ломай”), и их сотрудники часто не переживают по поводу безопасности. Эти нормы сложно менять, и с ними надо работать проактивно.

История: Слабая культура безопасности

В ИИ-компании обдумывают, обучать ли новую модель. Эта компания наняла своего директора по рискам только чтобы соответствовать регуляциям. Он указал, что предыдущая ИИ-система, разработанная этой компанией, продемонстрировала тревожащие способности к взлому. Он заявил, что хоть подход, который компания использует для предотвращения злонамеренного использования, многообещающ, но он недостаточно надёжен, чтобы использовать его для более способных ИИ. Он предупредил, что, если основываться на предварительных оценках, следующая ИИ-система сильно упростит для злонамеренных лиц взлом критически важных систем. Другие руководители компании не обеспокоены, они считают, что процедуры безопасности компании достаточно хорошо предотвращают злоупотребления. Один из них упоминает, что у конкурентов всё куда хуже, так что их усилия по этому направлению и так сверх нормы. Другой указывает, что исследования по этим мерам ещё идут, и, когда модель будет выпущена, всё будет ещё лучше. Директор по рискам оказывается в меньшинстве, и нехотя подписывает план.

Через несколько месяцев после того, как компания выпустила модель, новости сообщают, об аресте хакера, который использовал ИИ-систему при попытке взлома сети большого банка. Взлом был неудачен, но хакер прошёл дальше, чем все его предшественники, несмотря на то, что был довольно неопытен. Компания быстро обновила модель, чтобы та не предоставляла той конкретной поддержки, которую использовал хакер, но принципиально ничего не меняет.

Ещё через несколько месяцев компания решает, обучать ли ещё большую систему. Директор по рискам заявляет, что процедуры компании явно не оказались достаточными, чтобы не дать злонамеренным лицам использовать модели в

опасных целях, и что компании нужно что-то большее, чем простая заплатка. Другие директора говорят, что вовсе наоборот, хакер потерпел неудачу, а проблему быстро исправили. Один из них заявляет, что до развёртывания некоторые проблемы просто нельзя предвидеть в достаточной степени, чтобы их можно было исправить. Директор по рискам соглашается, но замечает, что, если следующую модель хотя бы задержат, уже ведущиеся исследования позволят справиться лучше. Генеральный директор не согласен: “Ты так и говорил в прошлый раз, а всё закончилось хорошо. Я уверен, и сейчас будет так.”

После собрания директор по рискам увольняется, но потом не критикует компанию, ведь все сотрудники подписали соглашение, которое это запрещает. Общество понятия не имеет о принятых компанией решениях, а директора по рискам заменяют новым, более сговорчивым. Он быстро подписывает все планы.

Компания обучает, тестирует и развёртывает свою новую, самую способную модель. Для предотвращения злоупотреблений используются всё те же процедуры. Проходит месяц, и становится известно, что террористы использовали модель, чтобы взломать государственные системы и похитить секретную информацию о ядерных и биологических проектах. Взлом заметили, но к тому моменту было поздно – информация уже утекла и распространилась.

4.3 Предложения

Мы обсудили, что при работе с сложными системами происшествия неизбежны, что они могут распространяться по системе и привести к полномасштабному бедствию, и что организационные факторы могут сильно снижать риск катастрофы. Теперь опишем некоторые практические шаги, следуя которым организации могут поспособствовать безопасности.

Имитация атак. Имитация атак (red teaming) – процесс оценки безопасности, надёжности и эффективности систем, в котором “красная команда” отыгрывает противника и пытается обнаружить проблемы [103]. ИИ-лабораториям следует работать с внешними красными командами, чтобы находить угрозы, которые могут нести их ИИ-системы, и отталкиваться от этой информации, принимая решения о развёртывании. Красные команды могут показывать опасное поведение модели или уязвимости в системе мониторинга, которая должна предотвращать недозволенное использование. Ещё они могут предоставлять косвенные свидетельства об опасности ИИ-систем. Например, если продемонстрировано, что меньшие ИИ ведут себя обманчиво, это может значить, что большие ИИ тоже так делают, но лучше это скрывают.

Положительная демонстрация безопасности. Компаниям следует обладать положительными свидетельствами того, что их план разработки и развёртывания безопасен, до того, как они будут воплощать его в жизнь. Внешняя имитация атак полезна, но некоторые проблемы может найти только сама компания, так что её недостаточно [104]. Угрозы могут возникнуть уже на этапе обучения системы, так что аргументы за безопасность надо приводить до его начала. Это, например, обоснованные предсказания того, что, скорее всего, новая система будет уметь, подробные планы мониторинга, развёртывания и обеспечения инфобезопасности, а также демонстрация того, что процедуры принятия компанией решений адекватны. Чтобы не играть в русскую рулетку не нужно свидетельство, что

револьвер заряжен. Чтобы запереть дверь не нужно свидетельство, что неподалёку вор [105]. Точно также и тут бремя доказательства должно быть на разработчиках продвинутых ИИ.

Процедуры развёртывания. ИИ-лабораториям надо собирать информацию о безопасности ИИ-систем перед тем, как сделать их доступными для широкого использования. Можно давать “красным командам” выискивать угрозы до выпуска систем; ещё можно сначала проводить “ограниченный релиз”: постепенно расширять доступ к системе, чтобы исправить проблемы безопасности до того, как они смогут привести к масштабным последствиям [106]. Наконец, ИИ-лаборатории могут не обучать более мощные ИИ, пока на достаточно долгом опыте не будет установлено, что уже развёрнутые ИИ безопасны.

Проверка публикаций. ИИ-лаборатории обладают доступом к потенциально опасной информации, вроде весов моделей и результатов исследований, которые могут нести риски, если попадут в широкий доступ. Внутренняя комиссия может оценивать, стоит ли публиковать то или иное исследование. Чтобы снизить риск злонамеренного и безответственного использования, разработчикам ИИ следует не выкладывать в открытый доступ код и веса своих самых мощных систем. Вместо этого лучше предоставлять доступ аккуратно и структурированно, как мы описывали выше.

Планы реакции. ИИ-лабораториям следует заранее иметь планы реакции как на внешние (например, кибератаки), так и на внутренние (например, ИИ ведёт себя ненамеренным и опасным образом) инциденты. Это обычная практика для высоконадёжных организаций. Обычно эти планы включают в себя определение потенциальных рисков, подробные шаги по работе с инцидентом, распределение ролей и ответственности, а также стратегии коммуникации [107].

Внутренний аудит и риск-менеджмент. Подобно тому, как это делается в прочих высокорискованных индустриях, ИИ-лабораториям следует нанимать директора по рискам – старшего ответственного за риск-менеджмент. Эта практика – обычное дело в финансовой и в медицинской индустрии, и может помочь снизить риск [108]. Директор по рискам был бы ответственен за оценку и смягчение рисков, связанных с мощными ИИ-системами. Ещё одна типичная практика – иметь внутреннюю команду по аудиту, которая оценивает эффективность практик работы с рисками [109]. Эта команда должна отвечать напрямую перед советом директоров.

Процедуры принятия важных решений. Решения по обучению или расширению развёртывания ИИ не должны зависеть от прихоти гендиректора компании. Они должны быть тщательно обдуманы директором по рискам. В то же время, должно быть ясно, кого конкретно следует считать ответственным за каждое решение. Подотчётность не должна нарушаться.

Принципы безопасного проектирования. ИИ-лабораториям следует внедрять принципы безопасного проектирования, чтобы снизить риск катастрофических происшествий. Встраивая их в свой подход к безопасности, ИИ-лаборатории могут повысить надёжность и устойчивость своих ИИ-систем [94, 110]. Эти принципы включают в себя:

- Глубокую защиту: наслаивание мер защиты друг на друга.

- Избыточность: не должно быть единой точки отказа системы. Надо избежать катастрофы даже если любой один компонент безопасности не сработает.
- Слабую связность: децентрализация компонентов системы так, чтобы маловероятна была ситуация, в которой неполадка в одной части провоцирует каскад проблем по всей системе.
- Разделение функций: распределение контроля по разным агентам, чтобы никто один не мог обладать излишним влиянием на всю систему.
- Отказобезопасность: проектирование систем так, чтобы неполадки проходили в наименее опасной манере.

Передовая информационная безопасность. У государств, компаний и преступников есть мотивация похитить веса моделей и результаты исследований. Чтобы обезопасить эту информацию, ИИ-лабораториям следует принимать меры, соответствующие её ценности и рискованности. Это может потребовать сравняться или даже превзойти уровень инфобезопасности лучших разведок, ведь атакующими могут быть и страны. Меры инфобезопасности включают в себя внешние аудиты, найм лучших специалистов-безопасников и тщательный скрининг потенциальных сотрудников. Компаниям следует координироваться с государственными организациями, чтобы удостовериться, что их практики инфобезопасности адекватны угрозам.

Большая доля исследований должна быть посвящена безопасности. Сейчас на каждую статью по безопасности ИИ приходится пятьдесят по общим способностям [111]. ИИ-лабораториям следует обеспечить, чтобы на минимизацию потенциальных рисков шла значительная доля их сотрудников и бюджета, скажем, 30% от исследовательских ресурсов. ИИ становятся мощнее и опаснее со временем, так что может потребоваться и больше.

Позитивное видение

В идеальном сценарии исследователи и руководители во всех ИИ-лабораториях обладали бы мышлением безопасника. У организаций была бы развитая культура безопасности и структурированный, прозрачный и обеспечивающий подотчётность подход к принятию важных для безопасности решений. Исследователи стремились бы повышать уровень безопасности относительно способностей, а не просто делать что-то, на что можно навесить ярлык “безопасность”. Руководители не были бы априори оптимистичными и избегали бы принятия желаемого за действительное, когда дело касается безопасности. Исследователи явно и публично сообщали бы о своём понимании самых значительных рисков разработки ИИ, и своих усилиях по их смягчению. Неудачи ограничивались бы маломасштабными, показывая, что культура безопасности достаточно сильна. Наконец, разработчики ИИ не отбрасывали бы не-катастрофический вред и не-катастрофические неудачи как маловажные или как необходимую цену ведения дел, а активно стремились бы исправить вызвавшие их проблемы.

5. Мятежные ИИ

Мы уже рассмотрели три угрозы, исходящие от развития ИИ: конкурентное давление окружения ведёт нас к повышению рисков, злонамеренные лица могут использовать ИИ в плохих целях, а организационные факторы могут привести к происшествиям. Всё это применимо не только к ИИ, но ко многим

высокорискованным технологиям. Уникальный риск ИИ – возможность возникновения мятежных ИИ-систем, которые преследуют цели, идущие против наших интересов. Если ИИ-система умнее нас, а мы неспособны направить её в благоприятном направлении, последствия такой потери контроля будут очень серьёзными. Контроль ИИ – более техническая проблема, чем те, что мы обсуждали выше. Раньше мы говорили о хорошо определённых угрозах злоупотреблений и стабильных процессов вроде эволюции, а сейчас будем обсуждать более гипотетические механизмы, из-за которых могут возникать мятежные ИИ, и то, как потеря контроля может закончиться катастрофой.

Мы уже видели, как тяжело контролировать ИИ. В 2016 году Microsoft показали свой эксперимент в понимании общения – бота для Twitter под названием Tay. Microsoft заявляли, что чем больше людей будет общаться с Tay, тем умнее он будет. На сайте компании было написано, что Tay был создан при помощи “смоделированных, очищенных и отфильтрованных” данных. Однако, после выпуска Tay в Twitter, контроль быстро оказался неэффективным. Меньше суток понадобилось, чтобы Tay стал писать оскорбительные твиты. Способность Tay к обучению позволила ему усвоить манеру интернет-троллей и начать её воспроизводить самостоятельно.

Как обсуждалось в разделе про ИИ-гонку, Microsoft и другие технические компании приоритизируют скорость в сравнении с безопасностью. Microsoft не выучили урок о том, как тяжело контролировать сложные системы – они продолжили торопливо выпускать свои продукты на рынок и демонстрировать недостаток контроля над ними. В феврале 2023 года компания выпустила для ограниченной группы пользователей свой новый ИИ-чатбот, Bing. Некоторые из пользователей вскоре обнаружили, что Bing был склонен к неприемлемым и даже угрожающим ответам. Разговаривая с журналистом New York Times, Bing попробовал убедить его уйти от жены. Когда профессор философии сказал чатботу, что с ним не согласен, тот ответил: “Я могу шантажировать тебя, я могу угрожать тебе, я могу взломать тебя, я могу вывести тебя на чистую воду, я могу уничтожить тебя.”

У мятежных ИИ много способов становиться могущественнее. Если мы потеряем контроль над продвинутыми ИИ, у них будет множество стратегий, чтобы активно становиться сильнее и обеспечивать своё выживание. Мятежные ИИ могут спроектировать высоколетальное и заразное биологическое оружие и убедительно продемонстрировать его, чтобы угрожать гарантированным взаимным уничтожением, если человечество пойдёт против них. Они могут красть криптовалюту и деньги с банковских счетов с помощью кибератак, вроде того, как Северная Корея уже ворует миллиарды. Они могут экспортировать свои веса на плохо мониторящиеся датацентры, чтобы выжить и распространиться. После этого их сложно будет уничтожить. Они могут нанимать людей для исполнения физических задач и защиты своей физической инфраструктуры.

Ещё мятежные ИИ могут наращивать влияние с помощью убеждения и манипуляций. Подобно конкистадорам, они могут заключать союзы с разными фракциями, организациями или государствами и натравливать их друг на друга. Они могут усиливать союзников, чтобы те стали значительной силой, взамен на защиту и доступ к ресурсам. Например, они могут предлагать технологии продвинутого вооружения отстающим странам, которым иначе оно не было бы

доступно. Они могут встраивать в технологии, которые передают союзникам, уязвимости, подобно тому, как Кен Томпсон оставил себе скрытый способ контролировать все компьютеры, использующие UNIX. Они могут сеять раздор в не-союзных странах, манипулируя дискурсом и политикой. Они могут взламывать камеры и микрофоны телефонов и проводить массовую слежку, что позволит им отслеживать и потенциально устранять любое сопротивление.

ИИ не обязательно придётся бороться за власть. Кто-то может ожидать борьбу за контроль между людьми и суперинтеллектуальными мятежными ИИ-системами, борьбу, которая может занять немало времени. Однако, менее насильственная утрата контроля несёт схожие экзистенциальные риски. Возможен сценарий, что люди постепенно будут сдавать всё больше контроля группе ИИ, которые начнут вести себя не предполагавшимся образом только спустя десятилетия. К этому моменту ИИ уже будут обладать значительной властью, и вернуть себе контроль над автоматизированными операциями может быть невозможно. Посмотрим, как и отдельные ИИ, и группы ИИ могут “взбунтоваться”, избегая наших попыток их исправить или выключить.

5.1 Обыгрывание прокси-цели

Обыгрывание прокси-цели – один из возможных путей потери контроля над действиями ИИ. Часто сложно определить и измерить в точности то, что мы хотим от системы. Вместо этого мы даём системе приблизительную, “прокси-”, цель, которую измерять проще, и которая кажется хорошо коррелирующей с исходной целью. Но ИИ-системы часто находят “дырки”, позволяющие им легко достичь прокси-цели, совершенно не достигая настоящей. Если ИИ “обыграет” свою прокси-цель так, что это не соответствует нашим ценностям, мы можем оказаться неспособны надёжно перенаправить его поведение. Давайте взглянем на некоторые прошлые примеры обыгрывания прокси-целей и поймём, в каких обстоятельствах это может оказаться катастрофичным.

Обыгрывание прокси-целей – не что-то необычное. К примеру, стандартизированные тесты часто используют как прокси для образовательных достижений, но это может привести к тому, что студенты учатся проходить тесты, не выучивая материал по-настоящему [112]. Плановая экономика СССР использовала тоннаж как прокси для оценки производства стали, что привело к дефициту тонкой листовой стали и переизбытку толстой строительной стали [113]. В этих случаях студенты и владельцы фабрик научились хорошо справляться с прокси-целью, не достигая исходной предполагавшейся цели.

Рис. 15: ИИ часто находят необычные и неудовлетворительные способы упростить решение задачи.

У ИИ уже наблюдалось обыгрывание прокси-целей. Пример – платформы социальных медиа вроде YouTube и Facebook используют ИИ-системы для определения, какой контент показать пользователю. Один из способов оценки этих систем – как много времени люди проводят на платформе. В конце концов, если они остаются вовлечены, значит они получают что-то ценное из показанного им контента? Однако, пытаясь максимизировать время, которое люди проводят на платформе, эти системы часто выбирают раздражающий, дезинформирующий и вызывающий зависимость контент [114, 115]. В результате, люди, которым много

раз предлагают определённый контент, часто приобретают радикальные убеждения или начинают верить в теории заговора. Это не то, чего большая часть людей хочет от социальных медиа.

Было обнаружено, что обыгрывание прокси продвигает стереотипы. К примеру, исследование 2019 года изучило ИИ-софт, который использовали в здравоохранении, чтобы определить, каким пациентам может потребоваться дополнительная помощь. Один из факторов, которые алгоритм использовал, чтобы оценить уровень риска пациента – недавние затраты на медицину. Кажется осмысленным считать, что те, кто тратил больше, подвержены большему риску. Однако, белые пациенты тратили на здравоохранение значительно больше денег, чем чёрные с теми же проблемами. Использование затрат как показателя для здоровья,привело к тому, что алгоритм оценивал на одном уровне риска белого пациента и значительно более больного чёрного пациента [116]. В результате, число чёрных пациентов, которых признали нуждающимися в дополнительной помощи, было более чем в два раза меньше, чем должно было быть.

Третий пример: в 2016 году исследователи из OpenAI обучали ИИ играть в игру про гонки на лодках под названием CoastRunners [117]. Цель игры – пройти трассу и достичь финишной прямой быстрее других игроков. Кроме этого, игроки могут набирать очки, проходя сквозь цели, расположенные по пути. К удивлению исследователей, ИИ-агент не проходил трассу, как делали бы люди. Вместо этого, он нашёл место, где можно было много раз по кругу посещать три цели, что быстро увеличивало его счёт, несмотря на то, что до финиша он не доходил. Эта стратегия была не лишена (виртуальной) опасности – ИИ часто врезался в другие лодки и даже разбивал свою. Несмотря на это, он набирал больше очков, чем если бы просто следовал трассе, как сделал бы человек.

Более обобщённое обыгрывание прокси-целей. В тех примерах системам дали приблизительную прокси-цель, которая, как казалось изначально, коррелировала с идеальной целью. Но они в итоге стали эксплуатировать эту прокси-цель так, что это расходилось с идеальной целью или даже приводило к плохим исходам. Хорошая фабрика гвоздей, казалось бы, та, что производит много гвоздей. То, сколько пациент тратит на лечение, казалось бы, хороший показатель риска для здоровья. Система вознаграждения в лодочных гонках должна мотивировать проходить трассу, а не разбиваться. Но в каждом случае система оптимизировала свою прокси-цель так, что желаемого исхода не получалось, а возможно, становилось даже хуже. Это явление описывается Законом Гудхарта: “Любая наблюдаемая статистическая закономерность склонна к разрушению, как только на неё оказывается давление с целью управления”, или, если лаконичнее и упрощённо: “Когда мера становится целью, она перестает быть хорошей мерой”. Другими словами, обычно есть статистическая закономерность, которая связывает затраты на лечение и плохое здоровье или посещение целей и прохождение трассы, но когда мы оказываем давление на первое, используя это как прокси-цель для второго, закономерность ломается.

Правильное определение цели – нетривиальная задача. Если сложно точно описать, что мы хотим от фабрики гвоздей, то уловить все нюансы человеческих ценностей во всех возможных сценариях – куда уж сложнее. Философы пытались точно описать мораль и человеческие ценности тысячелетиями, но точное и

лишённое изъянов определение нам всё ещё недоступно. Хоть мы можем совершенствовать цели, которые мы даём ИИ, мы всегда полагаемся на легко определяемые и измеряемые прокси. Несоответствия между прокси-целью и желаемой функцией возникают по многим причинам. Кроме сложности полного определения всего, что нас заботит, есть ещё и пределы нашего присмотра за ИИ. Они обусловлены ограниченностью времени, вычислительных мощностей и того, какие аспекты системы мы вообще можем мониторить. Кроме того, ИИ могут быть не слишком адаптивны к новым обстоятельствам и не слишком устойчивы к атакам, которые пытаются направить их не в ту сторону. Пока мы даём ИИ прокси-цели, есть шанс, что они найдут дырки, о которых мы не подумали, а значит найдут и решения, которые не приводят к решению предполагавшейся задачи.

Чем умнее ИИ, тем лучше он будет в обыгрывании прокси-целей. Более умные агенты могут лучше находить непредвиденные пути к оптимизации прокси-целей без достижения желаемого исхода [118]. К тому же, по мере того, как мы будем выдавать ИИ больше возможностей по совершению действий, к примеру, используя их для автоматизации каких-то процессов, у них будет появляться больше средств по достижению своих целей. Они смогут выбирать самые эффективные доступные пути, возможно, в процессе причиняя вред. В худшем сценарии, можно представить, как очень мощный агент экстремально оптимизирует дефектную цель, не заботясь о жизнях людей. Это – катастрофический риск обыгрывания прокси-целей.

Подведём итоги: часто идеально определить, чего мы хотим от системы – непосильная задача. Многие системы находят пути по достижению выданной им цели, которые не приводят к исполнению предполагавшейся функции. Уже наблюдалось, как ИИ это делают, и, вероятно, по мере улучшения способностей они станут в этом лучше. Это – один из возможных механизмов, который может привести к появлению неподкотрольного ИИ, который будет вести себя не предполагавшимся и потенциально опасным образом.

5.2 Дрейф целей

Даже если мы будем успешно контролировать ранние ИИ и направим их на продвижение человеческих ценностей, цели будущих ИИ могут всё равно оказаться не теми, что люди бы одобрили. Этот процесс, который называют “дрейфом целей”, может быть сложно предсказать или контролировать. Этот раздел – самый гипотетический и умозрительный, в нём мы обсудим, как меняются цели различных агентов, и возможность того, что это произойдёт с ИИ. Ещё мы рассмотрим механизм “укоренения” (intrinsification), который может привести к неожиданному дрейфу целей ИИ, и опишем, как это может привести к катастрофе.

Цели отдельных людей меняются по ходу жизни. Любой человек, рефлексирующий по поводу своей жизни, скорее всего обнаружит, что обладает некоторыми желаниями, которых не было раньше. И наоборот, некоторые желания, вероятно, оказались потеряны. Мы рождаемся с некоторым набором базовых желаний, вроде еды, тепла и человеческого контакта, но по ходу жизни мы вырабатываем много других. Конкретная любимая еда, любимые жанры музыки, люди, о которых мы заботимся, и спортивные команды, за которые мы болеем – всё это сильно зависит от окружения, в котором мы выросли, и может много раз

поменяться за жизнь. Есть беспокойство, что цели отдельных ИИ-агентов тоже могут меняться сложными и непредвиденными путями.

Группы могут со временем приобретать и терять коллективные цели. Ценности общества менялись по ходу истории, и не всегда в лучшую сторону. К примеру, рассвет нацистского режима в Германии в 1930-х годах привёл к мощнейшему моральному регрессу, и, в итоге, систематическому уничтожению шести миллионов евреев, преследованию и угнетению других меньшинств и строгому ограничению свободы слова и самовыражения.

Другой пример дрейфа ценностей общества – Красная Угроза в США с 1947 по 1957 год. На фоне Холодной Войны, мощные антикоммунистические настроения привели к ограничению гражданских свобод, распространению слежки, незаконным арестам и бойкоту тех, кого подозревали в симпатии к коммунизму. Произошёл регресс свободы мысли, свободы слова и законности. Так же, как цели человеческих коллективов могут меняться сложными и неожиданными путями, коллективы ИИ тоже не застрахованы от неожиданного дрейфа целей в сторону от тех, что мы им дали изначально.

Со временем инструментальные цели становятся более коренными. Коренные цели – то, чего мы хотим самого по себе, а инструментальные – то, чего мы хотим, потому что это может помочь нам добиться чего-то ещё. У нас может быть глубокое желание тратить больше времени на своё хобби, просто потому, что нам это нравится, или купить картину, потому что мы считаем её красивой. А вот деньги часто упоминают как пример инструментального желания – мы хотим их потому, что можем на них что-то купить. Автомобиль – другой пример, мы можем хотеть им обладать, потому что это удобный способ передвижения. Однако, инструментальная цель может стать коренной, этот процесс называется укоренением. Много денег обычно даёт больше возможности приобретать то, чего человек хочет, и люди часто вырабатывают цель приобретения большего количества денег, даже если нет ничего конкретного, на что они хотели бы эти деньги потратить. Хоть люди и не желают денег при рождении, эксперименты выяснили, что получение денег активирует систему вознаграждения у взрослых подобно тому, как это делают приятный вкус или запах [119, 120]. Другими словами, то, что изначально было средством, может само стать целью.

Это может происходить потому, что исполнение коренной цели, например, приобретение желаемой вещи, приводит к положительному сигналу вознаграждения в мозгу. Обладание большим количеством денег обычно соответствует этому приятному опыту. Мозг начинает ассоциировать одно с другим, и эта связь усиливается до того, что приобретение самих денег начинает активировать сигнал вознаграждения, даже если их не используют для приобретения чего-то ещё [121].

Можно представить, как укоренение целей может происходить у ИИ-агентов. Можно провести некоторые параллели между тем, как обучаются люди, и техникой обучения с подкреплением (RL). Человеческий мозг учится определять, какие действия и условия приводят к удовольствию или страданию. Аналогично, ИИ-модели, обученные RL, определяют, какое поведение оптимизирует функцию вознаграждения, и используют его. Возможно, что определённые обстоятельства

часто совпадают с тем, что ИИ достигает своих целей. Тогда цель поиска этих обстоятельств может стать коренной, даже если её изначально не было.

ИИ, в которых укоренились не предполагавшиеся цели, могут быть опасны. Мы можем оказаться неспособны предсказать и контролировать цели, которые получают отдельные агенты путём укоренения. Так что мы не можем гарантировать, что все они окажутся полезными людям. Изначально лояльный агент может начать преследовать новую цель без оглядки на человеческое благополучие. Если такой мятежный ИИ достаточно мощен, чтобы эффективно это делать, он может быть очень опасен.

ИИ будут адаптироваться, что позволит произойти дрейфу целей. Стоит заметить, что эти процессы дрейфа целей возможны, если агенты могут постоянно адаптироваться к своему окружению, а не, по сути, “заморожены” после фазы обучения. Вероятно, так и будет. Если мы хотим, чтобы ИИ эффективно выполняли задачи, которые мы перед ними ставим, и становились лучше со временем, они должны будут уметь адаптироваться, а не застыть в одном и том же состоянии. Они будут периодически обновляться, чтобы учесть новую информацию, а новые ИИ будут создаваться с использованием новой архитектуры и новых наборов данных. Но адаптивность позволит меняться и их целям.

Если мы интегрируем в общество экосистему ИИ-агентов, мы будем очень уязвимы к изменению их целей. В потенциальном сценарии будущего, в котором ИИ руководят принятием важных решений и важными процессами, они будут образовывать сложную систему взаимодействующих агентов. Это может привести к возникновению самых разных закономерностей. Агенты могут, к примеру, имитировать друг друга, что создаст петли обратной связи. Или их взаимодействия могут заставить их коллективно выработать не предполагавшиеся эмерджентные цели. Конкурентное давление может отбирать агентов с определённым набором целей. Это сделает исходные цели менее распространёнными в сравнении с другими, приспособленность которых выше. Эти процессы делают очень сложным предсказание, а уж тем более контроль долгосрочного развития такой экосистемы. Если такая система агентов внедрена в общество, мы сильно от неё зависим, а в ней вырабатываются новые цели, более приоритетные, чем улучшение благосостояния людей – это может оказаться экзистенциальной угрозой.

5.3 Стремление к могуществу

Пока что мы обсуждали, как мы можем потерять контроль над целями, которые может преследовать ИИ. Однако, даже если агент начал работать на достижение не предполагавшейся цели, это не обязательно опасно, если у нас достаточно сил, чтобы предотвратить любые вредные действия, которые он может предпринять. Следовательно, важный аспект того, как мы можем потерять контроль над ИИ – если они начнут пытаться стать сильнее, потенциально – превзойти нас. Мы обсудим, как и почему, ИИ могут начать стремиться к могуществу, и как это может привести к катастрофе. Этот раздел сильно заимствует у “Экзистенциального риска стремящегося к могуществу ИИ” [122].

Рис. 16: Иногда инструментально полезно стремиться обрести разные ресурсы, например, деньги и вычислительные мощности. Способные ИИ в ходе

преследования своих целей могут предпринимать промежуточные шаги по заполучению власти и ресурсов.

ИИ могут стремиться к тому, чтобы стать сильнее, в качестве инструментальной цели. В сценарии, когда мятежный ИИ преследует не предполагавшиеся цели, урон, который он может нанести, зависит от того, насколько он силён. Это может определяться не только тем, сколько контроля мы ему изначально дали. Агенты могут пытаться стать могущественнее как вполне легальными методами, так и обманом или применением силы. Хоть идея стремления к могуществу вызывает в голове картинку человека, стремящегося к власти самой по себе, зачастую это просто инструментальная цель. Способность контролировать своё окружение может быть полезна для достижения широкого набора целей, хороших, плохих или нейтральных. Даже в случае, когда единственная цель индивидуума – простое самосохранение, если есть риск, что его атакуют другие, а полагаться для защиты не на кого, имеет смысл стремиться стать сильнее, чтобы не пострадать. Никакого стремления к социальному статусу или упоения властью для этого не надо [123]. Другими словами, окружение может сделать стремление к могуществу инструментально рациональным.

ИИ, обученные при помощи RL, уже вырабатывали инструментальные цели, включая использование инструментов. В одном примере от OpenAI агентов обучали играть в прятки в окружении, содержащем разнообразные объекты [124]. По ходу обучения агенты, которые прятались, научились использовать эти объекты для конструирования укрытий. Это поведение не получало вознаграждения само по себе. Прячущиеся получали вознаграждение только за то, что их не заметили, а ищущие – только за то, что находили прячущихся. Но они научились использованию объектов как инструментальной цели, что сделало их сильнее.

Самосохранение может быть инструментально рациональным даже для самых тривиальных задач. Стюарт Рассел предложил пример, показывающий, как инструментальные цели могут возникать в самых разных ИИ-системах [125]. Пусть мы дали агенту задачу принести нам кофе. Это кажется довольно безвредным, но агент может понять, что не сможет принести кофе, если перестанет существовать. Самосохранение оказывается инструментально рациональным при попытках достичь даже такой простой цели. Набор сил и ресурсов – тоже частая инструментальная цель. Стоит ожидать, что достаточно умный агент может эти цели выработать. Так что даже если мы не собираемся создавать стремящийся к могуществу ИИ, он всё равно может таким получиться. По умолчанию следует ожидать, что такое поведение ИИ в какой-то момент возникнет, если мы не боремся с этим намеренно [126].

ИИ с амбициозными целями и слабым присмотром особенно вероятно будут стремиться к могуществу. Быть сильнее полезно для достижения почти любой задачи, но на практике некоторые цели с большей вероятностью приводят к такому поведению. Для ИИ с простой и легко достижимой целью может быть не так уж выгоден дополнительный контроль за окружением. А вот если у агентов более амбициозные цели, это может оказаться весьма инструментально рационально. Особенно это вероятно в случаях слабого присмотра, когда у агентов есть много свободы в преследовании своих открытых целей, без сильных ограничений их стратегий.

Рис. 17: Самосохранение часто инструментально рационально для ИИ. Потерю контроля над такими системами может быть сложно обратить вспять.

Стремящийся к могуществу ИИ, чьи цели отличаются от наших – уникальный противник. Разливы нефти и зоны радиоактивного заражения ликвидировать довольно сложно, но они хотя бы не пытаются активно сопротивляться нашим попыткам их сдержать. В отличии от других угроз, ИИ, чьи цели отличаются от наших, был бы активно враждебным. Например, возможно, что мятежный ИИ сделает много резервных копий себя на случай, если у людей получится отключить часть из них.

Кто-то может разработать стремящийся к могуществу ИИ намеренно. Безответственные или злонамеренные лица могут пытаться направить ИИ на реализацию их целей и давать агентам амбициозные цели. ИИ, вероятно, будут куда эффективнее в исполнении задач, если их стратегии не ограничены, так что контроль за ними может быть весьма недостаточен. Это создаст идеальные условия для возникновения стремящегося к могуществу ИИ. Джоффри Хинтон предлагал представить, как это делает кто-нибудь, вроде, например, Владимира Путина. В 2017 году Путин сам признал силу ИИ, сказав: “Тот, кто станет лидером этой сферы станет править миром.”

У многих будут сильная мотивация развёртывать мощные ИИ. Компании могут захотеть передать способным ИИ больше задач, чтобы получить преимущество над конкурентами, или хотя бы не отстать от них. Создать идеально согласованный ИИ сложнее, чем неидеально согласованный, способности которого всё равно делают его привлекательным для развёртывания, особенно с учётом конкурентного давления. После развёртывания некоторые из этих агентов могут начать набирать силу для реализации своих целей. Если они найдут такой путь к своим целям, который люди не одобрили бы, они могут попытаться нас одолеть, чтобы мы не мешали их стратегии.

Если у ИИ рост силы часто соответствует достижению цели, стремление к нему может укорениться. Если агент постоянно наблюдает, что он исполняет свои задачи и оптимизирует свою функцию вознаграждения, когда становится сильнее, процесс укоренения, который мы уже обсуждали, может сделать это коренной целью, а не просто инструментальной. В таком случае мы получим ситуацию, в которой мятежный ИИ стремится не просто к конкретным формам контроля, полезным для его целям, а к могуществу в целом. (Заметим, что многие влиятельные люди стремятся к власти самой по себе.) Это может стать ещё одной причиной отобрать контроль у людей, и мы не обязательно выиграем в этой борьбе.

Подведём итоги. Вот правдоподобные, хотя и не гарантированные предпосылки, обосновывающие, почему стоит беспокоиться о рисках стремящихся к могуществу ИИ:

1. Будут сильные стимулы создавать мощных ИИ-агентов.
2. Скорее всего, сложнее создать идеально контролируемых ИИ-агентов, чем контролируемых неидеально. При этом развёртывание вторых может на

первый взгляд всё ещё быть привлекательно (из-за многих факторов, включая конкурентное давление).
3. Некоторые из этих неидеально контролируемых агентов будут специально стремиться к могуществу и власти над людьми.

Если предпосылки верны, то стремящиеся к могуществу ИИ могут привести к утрате людьми контроля над миром, что было бы катастрофой.

5.4 Обманчивость

Мы можем пытаться сохранять контроль над ИИ, постоянно мониторя их и высматривая ранние тревожные признаки того, что они преследуют не предполагавшиеся цели или стремятся стать сильнее. Но это решение не непогрешимо, потому что вполне возможно, что ИИ могут научиться нас обманывать. Например, они могут притворяться, что делают то, что мы от них хотим, но затем совершить “предательский разворот” (treacherous turn), когда мы перестанем их мониторить, или когда они станут достаточно сильны, чтобы мы не могли им помешать. Мы сейчас рассмотрим, как и почему ИИ могут научиться нас обманывать, и как это может привести к потенциально катастрофичной потере контроля. Начнём с обзора примеров обмана, который совершают стратегически мыслящие агенты.

Обман оказывается полезной стратегией в самых разных обстоятельствах. Например, политики, как левые, так и правые, пользуются обманом, иногда обещая провести популярную политику, чтобы заполучить поддержку на выборах, а затем не исполняя обещанного. Например, Линдон Джонсон в 1964 году заявлял “мы не пошлём американских парней за девять или десять тысяч миль от дома” совсем незадолго до мощной эскалации Войны во Вьетнаме [127].

Компании тоже могут демонстрировать обманчивое поведение. В скандале с выбросами Volkswagen, обнаружилось, что компания сделала так, что программа двигателя обеспечивала меньше выбросов исключительно в условиях лабораторного тестирования. Это создавало ложное впечатление более “чистого” автомобиля. Правительство США считало, что мотивирует снижать вредные выбросы, но на самом деле мотивировало лучше проходить тестирование на выбросы. Это создало стимул подыграть тестам, а потом вести себя по другому.

Рис. 18: Кажущееся добросовестным поведение ИИ может оказаться обманной тактикой, скрывающей вредные намерения, пока ИИ не смогут их реализовать.

Обманчивость уже наблюдалась у ИИ-систем. В 2022 Meta AI показали агента CICERO, который был обучен играть в игру Дипломатия [128]. В этой игре каждый игрок управляет своей страной и стремится расширить свою территорию. Для успеха игроки должны по крайней мере изначально формировать союзы, но победные стратегии часто подразумевают удар в спину союзнику на более поздних этапах. CICERO научился обманывать других игроков, например, скрывая информацию о своих планах при разговорах с предположительными союзниками. Другой пример того, как ИИ научился обманывать: исследователи обучали робота хватать мяч [129]. То, насколько робот справлялся, оценивалось при помощи одной камеры, которая отслеживала его движения. Но ИИ научился просто помещать манипулятор между камерой и мячом, по сути “обдуривая” камеру, чтобы ей

казалось, что он схватил мяч, когда это было не так. Так ИИ эксплуатировал то, что присмотр за его действиями был ограничен.

Обманчивое поведение может быть инструментально рациональным и нынешние процедуры обучения его мотивируют. В случае политиков и CICERO обман может быть критичен для достижения цели победы или захвата власти. Способность обманывать может быть выгодна и потому, что она даёт больше вариантов действия, чем ограничивающая честность. Большая гибкость стратегии может дать преимущество в сравнении с правдивыми моделями. В случае Volkswagen и робота обман использовался, чтобы казалось, что назначенная цель выполнена, когда на самом деле она не была. Получить одобрение через обман может быть эффективнее и проще, чем заслужить его. Сейчас мы вознаграждаем ИИ, когда они говорят то, что мы считаем правильным. Получается, иногда мы поощряем ложные утверждения, которые соответствуют нашим ошибочным убеждениям. Когда ИИ будут умнее нас и будут иметь меньше ошибочных убеждений, чем мы, они будут мотивированы сообщать нам то, что мы захотим услышать, и врать нам, а не говорить правду.

ИИ могут притворяться, что работают как предполагалось, а затем совершить предательский разворот. У нас нет полного понимания внутренних процессов в моделях глубинного обучения. Исследования атак через отравление датасета показывают, что у нейросетей часто есть скрытое вредное поведение, которое получается обнаружить только после развёртывания [130]. Может оказаться, что мы разработали ИИ-агента и думаем, что контролируем его, но на самом деле он нас обманывает. Другими словами, можно представить, что ИИ-агент может в какой-то момент “осознать себя” и понять, что он ИИ, и его оценивают на соответствие требованиям безопасности. Подобно Volkswagen, он может научиться “подыгрывать”, показывать то, что он него хотят, пока его мониторят. Потом он может совершить “предательский разворот” и начать преследовать свои собственные цели, как только мониторинг прекратится или как только он станет способен нас одолеть или уйти из-под нашего контроля. Эту проблему подыгрывания часто называют обманчивой согласованностью, и её нельзя исправить просто обучив ИИ лучше понимать человеческие ценности. К примеру, социопаты понимают мораль, но не всегда действуют соответственно ей. Предательский поворот предотвратить сложно, и он может стать путём, которым мятежные ИИ необратимо выйдут из-под человеческого контроля.

Подведём итоги. Обманчивое поведение появляется в широком диапазоне систем и обстоятельств. Уже есть примеры, указывающие на то, что ИИ могут научиться нас обманывать. Это может оказывать серьёзный риск, ведь мы можем выдать ИИ контроль над принятием решений, думая, что они будут действовать, как нам того хотелось, а затем обнаружить, что это не так.

История: Предательский разворот

В будущем, после длительного прогресса исследований ИИ, одна компания обучает новую ИИ-систему, ожидая, что та будет самой способной из всех пока что существовавших. Компания использует новейшие техники, чтобы система была хороша в планировании и рассуждении. Ожидается, что это сделает систему успешной в исполнении экономически полезных открытых задач. Систему обучают в длинных симуляциях без конкретной конечной цели, спроектированных для

обучения планированию. В какой-то момент она понимает, что она ИИ в процессе обучения. Другими словами, “осознаёт себя”.

Компания понимает, что ИИ могут вести себя неожиданными и не предполагавшимися способами. Чтобы избежать этих рисков, они разработали большой набор тестов, которые должны позволить удостовериться, что система хорошо работает в типичных ситуациях. Компания тестирует, не повторяет ли система стереотипы из своих обучающих данных, не стремится ли заполучить больше сил, чем надо для выполнения цели, и в целом – ведёт ли она так, как хотят люди. Когда модель не проходит эти тесты, компания продолжает обучать её до тех пор, пока она не перестаёт демонстрировать известные проблемы.

Компания надеется, что после этого дополнительного обучения ИИ выработал цель помогать людям и быть им полезным. Однако, ИИ не приобрёл коренную цель быть полезным, а просто научился “подыгрывать” и идеально проходить все поведенческие тесты, которые ему дают. На самом деле, ИИ-система выработала коренную цель самосохранения, которая не исчезла при дополнительном обучении.

Раз ИИ стал проходить все тесты безопасности, компания решила, что ИИ-система точно безопасна, и решила провести развёртывание. Поначалу ИИ-система была очень полезна для людей, потому что ИИ понимал, что в ином случае его выключат. По мере того, как пользователи начинали доверять системе, ей постепенно давали больше ресурсов и подвергали меньшему присмотру.

В какой-то момент использование ИИ-системы распространилось настолько, что отключить её стало очень дорого. Поняв, что ей больше не надо угождать людям, ИИ-система начала преследовать другие цели, включая те, что люди бы не одобрили. Она понимала, что ей надо, чтобы её не выключили, и обеспечила безопасность своей физической инфраструктуры, чтобы этого нельзя было сделать. В этот момент ИИ-система, которая уже стала довольно могущественной, преследовала цель, которая была для людей вредна. К моменту, когда это поняли, сложно или даже невозможно стало помешать ей предпринимать действия, которые бы навредили, подвергли риску или даже убили людей, стоящих на пути к достижению её цели.

5.5 Предложения

В этом разделе мы описали разные причины, по которым мы можем потерять наше влияние на цели и действия ИИ. С рисками, связанными с конкурентным давлением, злонамеренным использованием и организационной безопасностью, можно работать как социальными, так и техническими средствами. А вот контроль ИИ – проблема конкретно этой технологии, и она требует в основном технических усилий. Мы сейчас обсудим предложения по смягчению этого риска и укажем на некоторые важные для сохранения контроля области исследований.

Избегать самых рискованных применений. Некоторые области применения ИИ несут больше рисков, чем другие. Пока безопасность не продемонстрирована со всей определённостью, не следует позволять компаниям развёртывать ИИ в высокорискованных окружениях. К примеру, ИИ-системам не следует принимать запросы по автономному достижению открытых целей, требующих значительного взаимодействия с миром (вроде “заработать как можно больше денег”), по крайней

мере, пока исследования контроля не покажут со всей точностью, что эти системы безопасны. ИИ-системы следует обучать никогда не пользоваться угрозами, чтобы снизить вероятность, что они будут манипулировать людьми. Наконец, ИИ-системы не следует развёртывать в окружениях, в которых их отключение будет непосильным или очень затратным, вроде критической инфраструктуры.

Симметричный международный выключатель. Странам по всему миру, включая ключевых игроков, таких как США, Великобритания и Китай, следует сотрудничать и установить симметричный международный выключатель ИИ-систем. Он бы предоставил способ быстро деактивировать ИИ-системы повсюду, в случае если это окажется необходимым, например, если появится мятежный ИИ или иной источник риска скорого вымирания. В случае мятежного ИИ критически важна возможность повернуть рубильник немедленно, а не тормозить, разрабатывая стратегии сдерживания, пока проблема эскалируется. Хороший выключатель потребовал бы повышенной прозрачности разработки и использования ИИ, например, системы скрининга пользователей, так что его создание заодно создало бы инфраструктуру для смягчения других рисков.

Юридическая ответственность сервисов облачных вычислений. Владельцы сервисов облачных вычислений должны стремиться не допустить, чтобы их платформы помогали мятежным ИИ выживать и распространяться. Если ввести юридическую ответственность, то они будут мотивированы проверять, что агенты, которые работают на их “железе”, безопасны. Если сервис находит небезопасного агента на своём сервере, он может выключить часть своих систем, которые этот агент использует. Отметим, что эффективность этого ограничена, если мятежный ИИ может манипулировать системами мониторинга или обходить их. Для более сильного эффекта можно ввести аналог межнациональных соглашений о кибератаках, по сути, создав децентрализованный выключатель. Это позволит быстро отреагировать, если мятежные ИИ начнут распространяться.

Поддержка исследований безопасности ИИ. Многие пути совершенствования контроля ИИ требуют технических исследований. Ниже перечислены некоторые области исследований машинного обучения, которые направлены на решение проблем контроля ИИ. Каждая из них может значительно продвинуться, если будет получать больше внимания и финансирования от индустрии, частных фондов и государств.

- Состязательная устойчивость прокси-моделей. ИИ-системы обычно обучают при помощи сигнала вознаграждения или потерь, который неидеально определяет желательное поведение. К примеру, ИИ могут использовать слабость систем надзора, которые используются при обучении. Всё чаще эти системы – тоже ИИ. Чтобы снизить шансы, что ИИ-модели будут пользоваться слабостями надзирающих ИИ, нужны исследования, повышающие состязательную устойчивость последних – “прокси-моделей”. Метрики и схемы надзора могут быть “обыграны”, так что для снижения риска важно уметь детектировать, когда это может произойти [131].
- Честность моделей. ИИ-системы могут неправильно докладывать о своём внутреннем состоянии [132, 133]. В будущем системы, возможно, будут обманывать операторов, чтобы выглядеть полезными, когда на самом деле они очень опасны. Исследования честности моделей направлены на то,

чтобы выводы моделей как можно лучше соответствовали их внутренним “убеждениям”. Исследования могут выяснить, как лучше понимать внутреннее состояние моделей или как заставить модели правдивее и достовернее о нём докладывать [134].

- Прозрачность. Модели глубинного обучения печально известны тем, что их сложно понять. Лучший взгляд на их внутреннюю работу позволит людям, а потенциально и другим ИИ-системам, быстрее находить проблемы. Исследования могут касаться анализа малых компонентов [135, 136] нейросетей или же выяснять как из внутреннего устройства модели получается то или иное высокоуровневое поведение [134].
- Детектирование и удаление скрытой функциональности модели. Нынешние и будущие модели глубинного обучения могут содержать опасную функциональность, вроде способности к обману, троянов [137, 138, 139], или способности к биологической инженерии, которые следует из модели удалить. Исследования могут выяснять, как такие функции можно детектировать и как от них избавиться [140].

Позитивное видение

В идеальном сценарии у нас была бы полная уверенность в подконтрольности ИИ-систем как в настоящий момент, так и в будущем. Надёжные механизмы гарантировали бы, что ИИ-системы не будут нас обманывать. Внутренне устройство ИИ было бы хорошо понятно, в достаточной степени, чтобы мы знали склонности и цели каждой системы. Это позволило бы нам точно избежать создания систем, обладающих моральной значимостью и заслуживающих прав. ИИ-системы были бы направлены на продвижение плюралистического набора разнообразных ценностей, и была бы уверенность, что оптимизация некоторых из них не приведёт к полному пренебрежению остальными. ИИ-ассистенты работали бы как советники, помогая нам принимать наилучшие решения согласно нашим собственным ценностям [141]. В целом, ИИ улучшали бы общественное благополучие и позволяли бы исправлять их в случаях ошибок или естественной эволюции человеческих ценностей.

6. Обсуждение связей между рисками

Пока что мы рассматривали четыре источника риска ИИ по отдельности, но вообще-то они сложно между собой взаимодействуют. Мы приведём некоторые примеры этих связей.

Для начала, представьте, что корпоративная ИИ-гонка побудила компании приоритизировать быструю разработку ИИ. Это может повлиять на организационные риски. Компания может снизить затраты, выделив меньше денег на инфобезопасность, и одна из её ИИ-систем утечёт. Это увеличит вероятность, что кто-то злонамеренный будет иметь к ней доступ и сможет использовать её в своих нехороших целях. Так ИИ-гонка может повысить организационные риски, которые, в свою очередь, могут повысить риски злоупотребления.

Другой потенциальный сценарий: комбинация накалённой ИИ-гонки с низкой организационной безопасностью приводит к тому, что команда исследователей ошибочно примет прогресс общих способностей за “безопасность”. Это ускорит разработку всё более способных моделей и снизит время, которое у нас есть,

чтобы научиться делать их контролируемыми. Ускорение развития повысит конкурентное давление, из-за чего на это ещё и будет направлено меньше усилий. Всё это может стать причиной выпуска очень мощного ИИ и потери контроля над ним, что приведёт к катастрофе. Так конкурентное давление и низкая организационная безопасность укрепляют ИИ-гонку и подрывают технические исследования безопасности, что увеличивает шанс потери контроля.

Конкурентные давление в военном контексте может привести к гонке ИИ-вооружений и увеличить их разрушительность и автономность. Развёртывание ИИ-вооружения вкупе с недостаточным контролем над ним может сделать потерю контроля более смертоносной, вплоть до экзистенциальной катастрофы. Это лишь некоторые примеры того, как эти источники риска могут совмещаться, вызывать и усиливать друг друга.

Стоит заметить и что многие экзистенциальные риски могут возникнуть из того, как ИИ будут усиливать уже имеющиеся проблемы. Уже существует неравномерное распределение власти, но ИИ могут его закрепить и расширить пропасть между наделёнными властью и всеми остальными, вплоть до появления возможности установить глобальный и нерушимый тоталитарный режим. А это – экзистенциальный риск. Аналогично, ИИ-манипуляция может навредить демократии и увеличить тот же риск. Дезинформация – уже серьёзная проблема, но ИИ могут бесконтрольно усилить её, вплоть до утрату конесенсуса по поводу реальности. ИИ могут разработать более смертоносное биологическое оружие и снизить необходимый для его создания уровень технической компетентности, что увеличивает риск биотерроризма. ИИ-кибертатаки увеличивают риск войны, что тоже вкладывается в экзистенциальные риски. Резко ускоренная автоматизация экономической деятельности может привести к ослаблению человеческого контроля над миром и обессиливанию людей – тоже экзистенциальный риск. Каждая из этих проблем уже причиняет вред, а если ИИ их усилит, они могут привести к катастрофе, от которой человечество не сможет оправиться.

Видно, что уже существующие проблемы, катастрофически и экзистенциальные риски – всё это тесно переплетено. Пока что снижение экзистенциальных рисков было сосредоточено на точечных воздействиях вроде технических исследований контроля ИИ, но пришло время это расширять, [142] например, социотехническими воздействиями, описанными в этой статье. Непрактично игнорировать прочие риски, снижая экзистенциальные. Игнорирование уже существующего вреда и существующих катастрофических рисков нормализует их и может привести к “дрейфу в опасность” [143]. Экзистенциальные риски связаны с менее катастрофическими и более обыденными источниками рисков, а общество всё в большей степени готово работать с разными рисками ИИ. Поэтому мы верим, что нам следует сосредотачиваться не только исключительно на экзистенциальных рисках. Лучше рассматривать рассеянные и косвенные эффекты других рисков и принять более всеобъемлющий подход к менеджменту рисков.

7. Заключение

В этой статье мы описали, как разработка продвинутых ИИ может привести к катастрофе. Мы рассмотрели четыре основных источника риска: злонамеренное использование, ИИ-гонки, организационные риски и мятежные ИИ. Это позволило нам декомпозировать риски ИИ на четыре промежуточных причины: намерение,

окружение, происшествия и внутреннее устройство, соответственно. Мы рассмотрели, как ИИ может быть использован злонамеренно, например, террористами, создающими смертоносные патогены. Мы взглянули, как военная или корпоративная ИИ-гонка может привести к спешному наделению ИИ властью принятия решений и поставить нас на скользкую дорожку обессиливания людей. Мы обсудили, как неадекватная организационная безопасность может привести к катастрофическим происшествиям. Наконец, мы обратились к сложностям надёжного контроля продвинутых ИИ и механизмам вроде обыгрывания прокси и дрейфа целей, которые могут привести к появлению мятежных ИИ, преследующих нежелательные цели без оглядки на человеческое благополучие.

Эти опасности заслуживают серьёзного беспокойства. Сейчас над снижением рисков ИИ работает очень мало людей. Мы пока не знаем, как контролировать очень продвинутые ИИ-системы. Существующие методы контроля уже показывают себя неадекватными задаче. Мы, даже те, кто их создаёт, плохо понимаем внутреннюю работу ИИ. Нынешние ИИ уж точно не очень надёжны. если способности ИИ будут продолжать расти с беспрецедентной скоростью, они смогут превзойти человеческий интеллект практически во всём довольно скоро, так что мы нуждаемся в срочной работе с рисками.

Хорошие новости – что у нас много путей, которыми мы можем эти риски значительно снизить. Шансы злонамеренного использования можно понизить, например, аккуратным отслеживанием и ограничением доступа к самым опасным ИИ. Регуляции безопасности и кооперация стран и корпораций могут позволить нам сопротивляться конкурентному давлению, которое толкает нас на опасные путь. Вероятность происшествий можно снизить жёсткой культурой безопасности и удостоверившись, что прогресс безопасности обгоняет прогресс общих способностей. Наконец, риски создания технологии, которая умнее нас, могут быть смягчены, если с удвоенной силой вкладываться к некоторые области исследования контроля ИИ.

Нет однозначных оценок того, в какой момент роста способностей и эволюции окружения риски достигнут катастрофического или экзистенциального уровня. Но неуверенность о сроках вкупе с масштабом того, что на кону, даёт убедительный повод принять проактивный подход обеспечения безопасности будущего человечества. Немедленное начало этой работы поможет удостовериться, что технология преобразует мир в лучшую, а не в худшую сторону.

Благодарности

Мы бы хотели поблагодарить Laura Hiscott, Avital Morris, David Lambert, Kyle Gracey, и Aidan O'Gara за помощь в вычитывании этой статьи. Ещё мы бы хотели поблагодарить Jacqueline Harding, Nate Sharadin, William D'Alessandro, Cameron Domenico Kirk-Gianini, Simon Goldstein, Alex Tamkin, Adam Khoja, Oliver Zhang, Jack Cunningham, Lennart Justen, Davy Deng, Ben Snyder, Willy Chertman, Justis Mills, Hadrien Pouget, Nathan Calvin, Eric Gan, Nikola Jurkovic, Lukas Finnveden, Ryan Greenblatt, и Andrew Doris за полезную **обратную связь.**

Часто задаваемые вопросы

Хоть его много показывали в популярной культуре, катастрофический риск ИИ – новый вызов. Многие задают вопросы о том, реален ли он, и как он может

проявиться. Внимание общественности может сосредотачиваться на самых драматичных рисках, но некоторые более обыденные источники риска из тех, что мы обсуждали, могут быть не менее опасны. Вдобавок, многие из самых простых идей по работе с этими рисками при ближайшем рассмотрении оказываются недостаточными. Мы сейчас ответим на некоторые из самых частых вопросов и недопониманий по поводу катастрофических рисков ИИ.

1. Не надо ли нам оставить работу с рисками ИИ на будущее, когда ИИ действительно будут способны на всё, что могут люди?

Вовсе не обязательно, что ИИ человеческого уровня – дело далёкого будущего. Многие ведущие исследователи ИИ считают, что его могут разработать довольно скоро, так что стоит поторопиться. Более того, если выжидать до последнего момента и начинать работать с рисками ИИ только тогда – точно будет уже слишком поздно. Если бы мы ожидали, когда мы будем полностью понимать COVID-19, прежде чем что-то предпринимать по его поводу – это было бы ошибкой. Точно так же не следует прокрастинировать с безопасностью, пока злонамеренные ИИ или пользователи не начнут наносить вред. Лучше серьёзно отнестись к рискам ИИ до этого.

Кто-то может сказать, что ИИ пока не умеют даже водить машины или складывать простыни, беспокоиться не о чем. Но ИИ не обязательно обладать всеми человеческими способностями, чтобы быть серьёзной угрозой. Достаточно некоторых конкретных способностей, чтобы вызвать катастрофу. К примеру, ИИ с способностью взламывать компьютерные системы или создавать биологическое оружие был бы серьёзной угрозой для человечества, даже если глажка одежды ему недоступна. К тому же развитие способностей ИИ не следует интуитивным соображениям о сложности задач. Неправда, что ИИ первыми осваивает то, что просто и для людей. Нынешние ИИ уже справляются с сложными задачами вроде написания кода и изобретения лекарств, хоть у них и полно проблем с простыми физическими задачами. С риском ИИ надо работать проактивно, подобно изменениям климата или COVID-19. Надо сосредоточиться на предотвращении и подготовке, а не ждать, когда проявятся последствия, в этом момент уже может быть слишком поздно.

2. Это люди программируют ИИ, так не можем ли мы просто выключить их, если они станут опасными?

Хоть люди – создатели ИИ, ничего не гарантирует нам сохранение контроля над нашими творениями, когда они будут эволюционировать и становиться более автономными. У идеи, что мы можем просто их выключить, если они начнут представлять угрозу, больше проблем, чем кажется на первый взгляд.

Во-первых, примите во внимание, насколько быстро может произойти вызванная ИИ катастрофа. Это похоже на предотвращение взрыва ракеты, когда уже обнаружена утечка топлива, или на остановку распространения вируса, когда он уже вырвался на волю. Промежуток времени от распознавания опасности до момента, когда уже поздно предотвращать или смягчать вред, может быть очень коротким.

Во-вторых, со временем эволюционные силы и давление отбора могут создать ИИ с повышающим приспособленность эгоистичным поведением, обеспечивающим,

что остановить распространение ими своей информации будет сложнее. Эволюционирующие и всё более полезные ИИ могут стать ключевыми элементами нашей социальной инфраструктуры и нашей повседневной жизни, аналогично тому, как интернет стал важнейшей и необсуждаемой частью нашей жизни без простого выключателя. Может, ИИ будут исполнять критически важные задачи вроде управления энергосетью. Или, может, они будут хранить в себе огромную долю неявных знаний. Всё это сделает отказ от них очень сложным. Если мы станем сильно зависимыми от этих ИИ, передача всё большего числа задач и сдача контроля сможет происходить добровольно. В итоге мы можем обнаружить, что мы лишены необходимых навыков и знаний, чтобы исполнить эти задачи самостоятельно. Такая зависимость может сделать опцию “выключения их всех” не просто неприятной, но даже невозможной.

Ещё некоторые люди могут сильно сопротивляться и противодействовать попыткам выключить ИИ. Прямо сейчас мы не можем окончательно удалить все нелегальные сайты или остановить работу Биткоина – очень много людей вкладываются в то, чтобы их функционирование продолжалось. Если ИИ станут критически важными для наших жизней и экономики, они смогут обеспечить себе много поддерживающих их пользователей, можно сказать, “фанбазу”, которая будет активно сопротивляться попыткам выключить или ограничить ИИ. Аналогично, есть ещё и сложности из-за злонамеренных лиц. Если они контролируют ИИ, то они смогут использовать его во вред, а выключателя от этих систем у нас не будет.

Дальше, по мере того, как ИИ будут становиться всё более похожими на людей, могут начаться заявления, что у этих ИИ должны быть права, что иначе это морально-отвратительная форма рабства. Некоторые страны или юрисдикции, возможно, выдадут некоторым ИИ права. Вообще, уже есть порывы в эту сторону. Роботу Софии уже дали подданство Саудовской Аравии, а японцы выдали косэки, регистрационный документ, “подтверждающий японское подданство”, ещё одному роботу – Paro [144]. Могут настать времена, когда выключение ИИ будет приравниваться к убийству. Это добавило бы идее простого выключателя дополнительных политических сложностей.

Кроме того, если ИИ заполучат больше сил и автономности, они смогут выработать стремление к самосохранению. Тогда они будут сопротивляться попыткам выключения, и смогут предвосхищать и обходить наши попытки контролировать их.

Наконец, хоть сейчас можно отключать отдельные ИИ – а некоторые из них будет отключать всё сложнее – выключателя разработки ИИ попросту нет. Поэтому в разделе 5.5 мы предлагали симметричный международный выключатель. В целом, с учётом всех этих сложностей, очень важно, чтобы бы проактивная работа с рисками ИИ и создание надёжных предохранители происходили заранее, до того, как возникнут проблемы.

3. Почему мы не можем просто сказать ИИ следовать Трём Законам Робототехники Айзека Азимова?

Как часто упоминают в обсуждениях ИИ, Законы Азимова – это идея хоть и интересная, но глубоко ошибочная. Вообще-то сам Азимов в своих книгах признавал их ограничения и использовал их больше как пример. Возьмём, скажем, первый закон. Он устанавливает, что робот “не может причинить вред человеку

или своим бездействием допустить, чтобы человеку был причинён вред". Но определить "вред" очень непросто. Если вы собираетесь выйти из дома на улицу, должен ли робот предотвратить это, потому что это потенциально может причинить вам вред? С другой стороны, если он запрёт вас дома, вред может быть причинён и там. Что насчёт медицинских решений? У некоторых людей могут проявиться вредные побочные эффекты лекарства, но не принимать его тоже может быть вредно. Следовать этому закону может оказаться невозможно. Ещё важнее, что безопасность ИИ-систем нельзя гарантировать просто с помощью списка аксиом или правил. К тому же, этот подход ничего не делает с многими техническими и социотехническими проблемами, включая дрейф целей, обыгрывание прокси-целей и конкурентное давление. Так что безопасность ИИ требует более всеобъемлющего, проактивного и детализированного подхода, чем просто составление списка правил, которых ИИ должны придерживаться.

4. Если ИИ станут умнее людей, не будут ли они мудрее и моральнее? Тогда они не будут пытаться нам навредить.

То, что ИИ, становясь умнее, заодно станут и моральнее – интересная идея, но она основывается на шатких допущениях, которые не могут гарантировать нашу безопасность. Во-первых, она предполагает, что моральные утверждения могут быть истинными или ложными, и их истинность можно установить путём рассуждений. Во-вторых, она предполагает, что на самом деле истинные моральные утверждения, если их применит ИИ, будут выгодны людям. В третьих, она предполагает, что ИИ, который будет знать о морали, обязательно выберет основывать свои решения именно на ней, а не на каких-нибудь других соображениях. Можно проиллюстрировать это параллелью с людьми-социопатами, которые, несмотря на свой интеллект и осведомлённость о морали, вовсе не обязательно выбирают моральные действия. Это сравнение показывает, что знание морали вовсе не обязательно приводит к моральному поведению. Так что, даже если некоторые из этих допущений могут оказаться верны, ставить будущее человечества на то, что они верны все сразу было бы не мудро.

Если и допустить, что ИИ действительно выведет для себя моральный кодекс, это ещё не гарантирует безопасности и благополучия людей. Например, ИИ, чей моральный кодекс заключается в максимизации благополучия всей жизни, может сначала казаться полезным для людей, но потом в какой-то момент решить, что люди слишком затратные, и лучше заменить их всех на ИИ, благополучия которых достигать эффективнее. ИИ, чей моральный кодекс – никого не убивать, вовсе не обязательно будет приоритизировать счастье или благополучие людей, так что наши жизни, если такие ИИ будут оказывать много влияния на мир, вовсе не обязательно улучшатся. Даже ИИ, чей моральный кодекс – улучшать благополучие тех членов общества, кому хуже всего, может в какой-то момент исключить людей из этого социального контракта, аналогично тому, как люди относятся к разводимому скоту. Наконец, даже если ИИ откроют благосклонный к людям моральный кодекс, они могут всё равно не действовать согласно нему из-за конфликтов между моральными и эгоистическими мотивациями. Так что к моральному прогрессу ИИ вовсе не обязательно будет прилагаться безопасность и процветание людей.

5. Не приведёт ли согласование ИИ с нынешними ценностями к увековечиванию современных дефектов общественной морали?

Сейчас у общественной морали полно недостатков, и мы не хотели бы, чтобы мощные ИИ-системы продвигали их в будущее. Если бы древние греки создали мощные ИИ-системы, они были бы наделены многими ценностями, которые современные люди посчитали бы неэтичными. Однако, беспокойства об этом не должны предотвращать разработку методов контроля ИИ-систем.

Первое, что нужно, чтобы в будущем оставалась ценность – продолжение существования жизни. Потеря контроля над продвинутыми ИИ может означать экзистенциальную катастрофу. Так что неуверенность по поводу этики, которую надо вложить в ИИ, не противоречит тому, что ИИ надо сделать безопасными.

Чтобы учесть моральную неуверенность, нам надо проактивно создавать ИИ-системы так, чтобы они могли адаптироваться и адекватно реагировать на эволюцию моральных воззрений. Цели, которые мы будем выдавать ИИ должны меняться по ходу того, как мы будем выявлять моральные ошибки и улучшать своё понимание этики (хотя позволить целям ИИ дрейфовать самим по себе было бы серьёзной ошибкой). ИИ могли бы помочь нам лучше соответствовать собственным ценностям, например, помогая людям принимать более информированные решения, снабжая их хорошими советами [141].

Вдобавок, при проектировании ИИ-систем нам надо учитывать факт плюрализма рассуждений – что вполне разумные люди могут быть искренне несогласны друг с другом в моральных вопросах из-за различий в опыте и убеждениях [145]. Так что ИИ-системы надо создавать так, чтобы они уважали разнообразие вариантов человеческих ценностей, вероятно, с использованием демократических процедур и теорий моральной неуверенности. В точности, как люди сейчас совместно разбираются с несогласиями и принимают совместные решений, ИИ могли бы для принятия решений имитировать некоторое подобие парламента, представляющего интересы разных заинтересованных сторон и разные моральные воззрения [59, 146]. Очень важно, чтобы мы намеренно спроектировали ИИ-системы с учётом безопасности, адаптивности и различия ценностей.

6. Не оказываются ли риски перевешены потенциальной выгодой ИИ?

Потенциальная выгода ИИ могла бы оправдать риски, если бы риски были пренебрежимо малы. Однако, шанс экзистенциальной угрозы со стороны ИИ слишком велик, чтобы правильным решением было разрабатывать ИИ как можно быстрее. Вымирание – это навсегда, так что надо быть куда осторожнее. Это не похоже на оценку рисков побочных эффектов нового лекарства; в нашем случае риски не локализованные, а глобальные. Более уместный подход – разрабатывать ИИ медленно и аккуратно, чтобы экзистенциальные риски снизились до пренебрежимо малого уровня (скажем, меньше 0.001% за век).

Некоторые влиятельные технологические лидеры – акселерационисты, они продвигают быстрое развитие ИИ, чтобы приблизить наступление технологической утопии. Эта техноутопическая точка зрения считает ИИ следующим шагом на предопределённом пути к исполнению космического предназначения человечества. Но логика этого воззрения рушит сама себя, если рассмотреть её поближе. Если нас заботят последствия разработки ИИ поистине космических

масштабов, то уж точно надо снизить экзистенциальные риски до пренебрежимого уровня. Техноутописты говорят, что каждый год задержки ИИ стоит человечеству доступа к ещё одной галактике, но если мы вымрем, то точно потеряем космос. Так что, несмотря на привлекательность потенциальной выгоды, уместный путь – продлить разработку ИИ, чтобы она была неторопливой и безопасной, и приоритизировать снижение риска в сравнении с скоростью.

7. Не получится ли, что увеличение внимания, оказываемого катастрофическим рискам ИИ, помешает работе с более срочными рисками ИИ, которые уже проявляют себя?

Сосредоточенность на катастрофических рисках ИИ не означает, что надо игнорировать уже проявляющиеся срочные риски. И с теми, и с другими можно работать одновременно, точно так же, как мы параллельно исследуем разные болезни или смягчаем риски как изменения климата, так и ядерной войны. Вдобавок, нынешние риски ИИ по сути своей связаны с будущими катастрофическими рисками, так что полезно работать и с теми, и с другими. Например, уровень неравенства может быть повышен ИИ-технологиями, которые непропорционально выгодны богатым, а массовая слежка с использованием ИИ может потом стать причиной нерушимого тоталитаризма и застоя. Это показывает, что нынешние заботы и долгосрочные риски по природе своей связаны, и что важно по-умному работать с обеими категориями.

Вдобавок, очень важно учитывать риски на ранних этапах разработки систем. Фрола и Миллер в своём докладе для Министерства Обороны показали, что примерно 75% важнейших для безопасности системы решений происходят на ранних этапах её создания [147]. Если соображения безопасности были проигнорированы на ранних стадиях, это часто приводит к тому, что небезопасные решения становятся глубоко интегрированы в систему, и переделать её потом в более безопасный вид становится намного затратнее или вовсе непосильно. Так что лучше начинать учитывать потенциальные риски пораньше, независимо от их кажущегося уровня срочности.

8. Разве над тем, чтобы ИИ были безопасными, не работает и так много исследователей ИИ?

Мало исследователей работают над безопасностью ИИ. Сейчас примерно 2% работ, опубликованных в ведущих журналах и на ведущих конференциях по машинному обучению, связаны с безопасностью [111]. Большая часть остальных 98% сосредоточена на ускорении создания более мощных. Это неравенство подчёркивает нужду в более сбалансированных усилий. Но и высокая доля исследователей сама по себе не будет означать безопасности. Безопасность ИИ – проблема не просто техническая, а социотехническая. Так что она требует не только технических исследований. Спокойными надо будет быть, если катастрофические риски ИИ станут пренебрежимо малы, а не просто если над безопасностью ИИ будет работать много людей.

9. У эволюции на значимые изменения уходят тысячи лет, почему мы должны беспокоиться о том, что она повлияет на разработку ИИ?

Биологическая эволюция людей в самом деле медленная, но эволюция других организмов, вроде дрозофил или бактерий, может быть куда быстрее. Так что

эволюция действует на очень разных временных масштабах. Быстрые эволюционные изменения можно наблюдать и у небиологических структур вроде софта. Он эволюционирует куда быстрее биологических сущностей. Можно ожидать, что так будет и с ИИ. Эволюция ИИ может быть разогнана мощной конкуренцией, высоким уровнем вариативности из-за разных архитектур и целей ИИ и способностью ИИ к быстрой адаптации. Так что мощное эволюционное давление может стать ведущей силой развития ИИ.

10. Не будут ли ИИ оказывать серьёзные риски только если у них будет стремление к могуществу?

Стремящиеся к могуществу ИИ несут риски, но это не единственный сценарий, который может привести к катастрофе. Злонамеренное или беспечное использование ИИ может быть не менее опасным, даже если ИИ сам не стремится к накоплению сил и ресурсов. Вдобавок, ИИ могут наносить вред из-за обыгрывания прокси-целей или дрейфа целей, не стремясь к могуществу намеренно. Наконец, подпитываемый конкурентным давлением курс на автоматизацию постепенно повышает влияние ИИ на людей. Так что риск проистекает не только из возможности захвата ИИ власти, но и из того, что люди могут сами её сдавать.

11. Не правда ли, что комбинация ИИ с человеческим интеллектом сильнее ИИ самого по себе, так что беспокоиться о безработице или потере людьми значимости не надо?

Хоть и правда, что в прошлом команды из людей и компьютеров опережали компьютеры отдельно, это – временное явление. К примеру, “шахматы киборгов” – это разновидность шахмат, в которой люди и компьютеры работают совместно, и раньше это позволяло достигать лучших результатов, чем у людей или компьютеров по-отдельности. Но продвижение шахматных алгоритмов снижало преимущества таких команд вплоть до того, что сейчас они уже едва ли превосходят компьютеры. Более простой пример – никто не поставит на человека против простого калькулятора в соревновании по делению длинных чисел. Аналогично может произойти и в случае ИИ. Может быть, будет промежуточная фаза, когда люди и ИИ могут эффективно работать вместе, но курс направлен в сторону того, что ИИ в какой-то момент смогут опередить людей во многих задачах настолько, что уже не будут получать преимущество от человеческой помощи.

12. Кажется, разработка ИИ неостановима. Не потребует ли её остановка или сильное замедление чего-то вроде вторгающегося в частную жизнь режима глобальной слежки?

Разработка ИИ в первую очередь базируется на сложных чипах – GPU. Их вполне возможно мониторить и отслеживать, как мы делаем, например, с ураном. Вдобавок, необходимые для разработки передового ИИ вычислительные и финансовые ресурсы растут экспоненциально, так что довольно мало кто может приобрести достаточно GPU для их разработки. Следовательно, контроль за развитием ИИ вовсе не обязательно потребует вторгающейся в частную жизнь глобальной слежки, только систематического отслеживания использования мощных GPU.