Этические вопросы использования ИИ

Чем быстрее развивается ИИ, тем быстрее он вторгается практически во все аспекты нашей жизни. ИИ предоставляет больше возможностей, приносит больше удобства, позволяет улучшить жизнь человека. Но так ли все радужно на самом деле? Могут ли скрываться за нашим взаимодействием тревожные проблемы, которые с каждым новым витком развития ИИ будут привносить все больше опасности нарушения этических и гуманных аспектов?

В таких вопросах нужно рассматривать обе стороны взаимодействия: то, как ИИ «относится» к нам и то, как люди относятся к нему. В вопросе «отношения» ИИ к нам представляются несколько областей, которые вызывают обеспокоенность: наблюдение и конфиденциальность, предвзятость и дискриминация.

Алгоритмы ИИ могут анализировать поведение людей как в реальном мире, так и в онлайн-пространстве средствами мониторинга активности в социальных сетях, онлайн-поиска и моделей общения. Благодаря массовому использованию персональных данных для обучения систем ИИ эта технология становится во многом похожей на систему наблюдения, которая может «знать», что думают люди, и предсказывать, что им понравится, не понравится или что они могут сделать в данной ситуации. Каждый человек имеет право на неприкосновенность частной жизни, но в таких случаях он может просто не знать, что за ним «наблюдают».

Дискриминация отдельных лиц и групп может возникнуть из-за предвзятости в системах ИИ. Безусловно, ИИ улучшает процессы, например, здравоохранения, помогая в диагностике и рекомендации лечения человека. Но подобные приложения могут приводит и к получению несправедливых результатов.

Например, Поставщики медицинских услуг в США используют алгоритмы ИИ для принятия решений в области здравоохранения. Например, какие пациенты требуют особого ухода или медицинских привилегий. Исследователи из Калифорнийского университета выявили признаки расовой предвзятости в алгоритмах. Было замечено, что алгоритм назначает одинаковый уровень риска темнокожим пациентам, но более больным, чем пациентам со светлым цветом кожи. Пациентам со светлым цветом кожи были присвоены более высокие оценки риска, поэтому с большей вероятностью их отберут для оказания дополнительной помощи.

Предвзятость имеет тенденцию уменьшать более чем половину числа темнокожих пациентов, нуждающихся в дополнительной помощи, по сравнению с белыми. Основная причина этого заключается в том, что алгоритм использует стоимость лечения, а не болезнь для нужд здравоохранения. На темнокожих пациентов тратится меньше денег, но уровень потребностей у них одинаковый. Таким образом, алгоритм ошибочно

считает, что темнокожие пациенты более здоровы, чем белые пациенты с тем же заболеванием.

Сложно внедрить соображения «недискриминации» и справедливости в системы ИИ. Возможно, можно указать алгоритмам не учитывать чувствительные атрибуты, способствующие дискриминации, такие как пол или этническая принадлежность, на основе возникновения дискриминации в определенном контексте. Нам стоит понимать, какие задачи могут быть возложены на такого «помощника», а какие следует выполнять только людям, способным проявлять профессионализм, понимание и отзывчивость. Если мы полагаемся на ИИ в том, что он приведет нас в мир нового труда, безопасности и эффективности, мы должны убедиться, что машины ведут себе как запланировано.

Отличительная черта человека, позволяющая ему доминировать в мире, от всех остальных существ — способность мыслить и изобретать. Сможет ли однажды ИИ получить такое же преимущество перед нами? Этот вопрос вызывает много споров и еще многие годы будет оставаться открытым, ведь, с одной стороны, мы можем просто отключить машину, а с другой, подобная машина может выработать механизм защиты от этого выключения, защищаясь от источника угрозы — человека.

Следующая сторона рассмотрения вопроса — наше отношение к ИИ. Существует концепция сильного и слабого (общего) ИИ. Если слабый ИИ окружает нас уже практически повсеместно, то сильный пока остается в планах, пока его не существует. Мнения ученых расходятся в вопросе, возможно ли создание такой технологии, ведь она должна иметь сознание, способность к самообучаемости для выполнения принципиально новых задач. Этично ли будет наносить вред и мучить ИИ, если мы все же придем со временем к его сильной версии, которая будет способна чувствовать и переживать плохое отношение? Нормально ли будет оставить своего инновационного собеседника без общения надолго, если он будет способен чувствовать одиночество?

Ученые трудятся над созданием базовых механизмов принятия и вознаграждения у ИИ: сейчас они достаточно поверхностные, но в перспективе способны стать более сложными и «живыми». Для этого используются генетические алгоритмы, которые создают множество экземпляров системы одновременно. Из них «выживают» только наиболее успешные, чтобы объединиться и сформировать новое поколение экземпляров, а безуспешные удаляются. В какое момент с развитием ИИ можно рассматривать генетические алгоритмы как форму убийства?

Список вопросов можно продолжать до бесконечности. Привнося такую технологию в свою жизнь, мы должны понимать, что необходимо и установить правила обращения и общения с ней, а не инфантильно думать, что своим отношением мы не сможем привести мир к катастрофе. Эта тема

поднималась уже 100 лет назад в романе Карела Чапека «R.U.R.» (1920), в котором существовала Лига гуманности. Она декларировала, что роботы не должны подвергаться бесчеловечному обращению.

Боты на основе ИИ становятся все лучше в моделировании человеческого разговора и отношений. В 2015 году бот по имени Женя Густман выиграл конкурс Тьюринга впервые в истории. В нем люди посредством текстовых сообщений общались с неизвестной сущностью, а потом пытались угадать, беседовали они с человеком или машиной.

В то время, как люди ограничены во внимании и доброте, которую они могут расходовать на другого человека, боты могут тратить почти неограниченные ресурсы на построение отношений. В неправильных руках, как у создателей, так и у потребителей, такие инструменты могут оказаться не просто вредными, они могут нести разрушающий характер.

Например, существуют виртуальные «товарищи» в форме чат-ботов или питомцев, которые могут участвовать в разговорах и оказывать эмоциональную поддержку. Безусловно, это может быть полезно в моменте, когда человек остался совершенно один со своей проблемой, когда нужно просто выговориться. Но стоит понимать, что такие технологии могут быть созданы людьми, которые не имеют искреннего желания помочь вам. Они создали их, чтобы закрыть боль потребителя, его потребности, получив от этого максимальную прибыль. Поэтому подобные ИИ могут не нанести вред только в том случае, если человек в большей степени эмоционально стабилен и устойчив к таким состояниям. Если же он страдает одиночеством или имеет неустойчивую психику, подобные средства могут только усугубить социальную изоляцию и ситуацию.

Или, например, технология deepfake, позволяющая манипулировать аудио- и видеоконтентом, может вызывать обеспокоенность по поводу доверия и подлинности в отношениях, если используются с желанием навредить.

ИИ может использоваться не только в корыстных целях, но и на благо другого человека, организации или государства. Например, ИИ делает банковские системы более безопасными, используя специальные алгоритмы и сценарии для обнаружения и борьбы с мошенническими транзакциями, или позволяет обнаруживать нарушения в уплате налогов со стороны физических и юридических лиц. Также существуют, улучшаются и разрабатываются технологии с ИИ для людей с расстройствами аутистического спектра и другими когнитивными нарушениями для их обучения и поддержания жизнеспособности.

Крупнейшие ИТ-компании мира, такие как Microsoft, IBM, Сбер и Яндекс уже сформулировали свое видение этики ИИ в виде сводов правил, кодексов и наборов принципов. Однако самым известным кейсом считается Азиломарская конференция 2017 года, на которой были сформулированы «23

принципа искусственного интеллекта», некоторые из которых устанавливают правила этики:

1. **Безопасность:** системы ИИ должны быть безопасными и надёжными в течение всего срока эксплуатации, а также по возможности проверяемыми.
2. **Прозрачность отказов:** если система ИИ причиняет вред, должна быть возможность выяснить, почему это произошло.
3. **Ответственность:** разработчики и создатели продвинутых систем ИИ являются заинтересованными сторонами в моральных последствиях их использования, неправильного использования и любых действий, они несут ответственность и имеют возможность формировать эти последствия.
4. **Согласование ценностей:** высокоавтономные системы ИИ должны быть спроектированы таким образом, чтобы их цели и поведение могли быть гарантированно согласованы с человеческими ценностями на протяжении всей их работы.
5. **Человеческие ценности:** системы ИИ должны разрабатываться и эксплуатироваться таким образом, чтобы быть совместимыми с идеалами человеческого достоинства, прав, свобод и культурного разнообразия.
6. **Приватность личной жизни:** люди должны иметь право доступа, управления и контроля над данными, которые они генерируют, учитывая возможности систем ИИ анализировать и использовать эти данные.
7. **Свобода и частная жизнь:** Применение ИИ к личным данным не должно необоснованно ограничивать реальную или предполагаемую свободу людей.
8. **Общее благо:** технологии ИИ должны приносить пользу и расширять возможности как можно большего числа людей.
9. **Общее процветание:** экономическое процветание, созданное ИИ, должно быть широко распространено на благо всего человечества.
10. **Человеческий контроль:** люди должны выбирать, как им делегировать решения системам ИИ (и делать ли это вообще) для достижения целей, также определённых человеком.
11. **Отсутствие злоупотреблений властью:** власть, полученная благодаря контролю над высокоразвитыми системами ИИ, должна применяться с уважением, а также улучшать, а не подрывать социальные и гражданские процессы, от которых зависит здоровье общества.

ИИ имеет огромный потенциал, и ответственное и этичное «общение» с ним зависит, в первую очередь, от каждого человека. Стоит понимать, что нахождение баланса между использованием возможностей ИИ и сохранением глубины и подлинности человеческих связей имеет важное значение.