ЭТИКА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПРОБЛЕМЫ И ПУТИ ИХ РЕШЕНИЯ

Панкратова М.Д, *студентка группы СГНЗ-44Б*
Сковпень Т.Н, *студентка группы СГНЗ-44Б*
Московский государственный технический университет им. Н.Э. Баумана

Научный руководитель: ***Галаганова С.Г.*** *кандидат философских наук, доцент кафедры «Информационная аналитика и политические технологии»*
galaganovasg@bmstu.ru

Аннотация: Статья посвящена исследованию этических проблем развития искусственного интеллекта, оставшихся за рамками принятого в 1921 году «Кодекса этики в сфере ИИ». Анализируя практику применения ИИ в автономных транспортных системах, поисковых сервисах, судебной системе, сфере социального общения, мониторинге исполнения сервисных работ, авторы выявляют актуальные этические дилеммы и предлагают возможные пути их решения. В статье также сформулированы некоторые общие механизмы обеспечения справедливости, непредвзятости и безопасности в процессе использования ИИ.

Ключевые слова: этика, искусственный интеллект (ИИ), этика ИИ, Кодекс этики в сфере ИИ.

Стремительно активизирующееся использование набора технологий, условно названных в 1956 году Джоном Маккарти «искусственным интеллектом», продолжает вызывать бурные дискуссии по поводу возникающих этических проблем. В настоящее время многие из них регламентируются принятым 26 октября 2021 года «Кодексом этики в сфере искусственного интеллекта». Кодекс представляет собой единую систему рекомендательных принципов, предназначенных для создания среды доверенного развития технологий искусственного интеллекта (в дальнейшем – ИИ) в России. Важность развития этики ИИ и необходимость увеличения числа подписантов Кодекса отражены в поручении Президента РФ по итогам конференции «Путешествие в мир ИИ» (от 16 декабря 2021 г. № Пр-2371, п. 1г), в соответствии с которым Правительству РФ поручено принять меры по увеличению числа российских и иностранных организаций, присоединившихся к Кодексу.

Важнейшей особенностью данного документа является его человекоориентированный и риско-ориентированный характер: здесь чётко обозначена сугубо человеческая ответственность за моральные риски разработки и внедрения искусственного интеллекта, подчёркивается необходимость гуманистической направленности данного процесса. Кодекс акцентировал главные общечеловеческие ценности: сохранение биосферы Земли и человека как биологического вида, его духовно-нравственное совершенствование, сохранение человеческой цивилизации и культуры.

Однако ряд этических проблем по-прежнему остаётся за рамками Кодекса. В данной статье эти проблемы будут сформулированы применительно к конкретным областям использования ИИ с последующими авторскими рекомендациями по их решению.

ИИ и автономные транспортные системы

Массачусетский технологический институт развернул «Moral Machine» – экспериментальную онлайн-платформу, предназначенную для изучения моральных дилемм, с которыми сталкиваются автономные транспортные средства. Эта платформа собрала 40 млн решений на десяти языках в более чем 200 странах мира. Проект оказал значительное влияние на общественное обсуждение этических аспектов функционирования автономных систем и разработки ИИ в целом. Здесь рассматриваются следующие ключевые этические проблемы ИИ:

- как определять ценность человеческой жизни в контексте программирования автономных систем?
- кто должен нести ответственность за действия, совершаемые автономными системами на основе ИИ?

Рекомендации:

ИИ, встроенный в автономные системы, должен быть однозначно настроен на принятие этически обоснованных решений в экстремальных ситуациях, а также обладать способностью объяснять свои действия. На наш взгляд, ответственность за действия, совершаемые автономными системами на основе ИИ, должны нести разработчики и владельцы этих систем. Законодательные органы должны разрабатывать законы и стандарты для регулирования использования автономных систем на основе ИИ и определения меры ответственности в случае их нарушения.

ИИ в поисковых сервисах

Без надлежащего регулирования этических аспектов мы можем столкнуться с проблемой «предвзятости» в информационных системах. Например, при создании рекомендательной системы на основе ИИ разработчик может настроить её таким образом, чтобы продвигать товары определённого производителя в ущерб другим. Это может нарушить принципы конкуренции и создать риск для бизнеса. Ещё одним примером является «предвзятость» в системах распознавания лиц, которые могут ошибочно идентифицировать людей определённой национальности как потенциальных преступников.

Конкретным примером компании, столкнувшейся с этической проблемой применения ИИ, является Facebook. Стало известно, что их алгоритмы новостной ленты могут «загонять» пользователей в так называемые «информационные пузыри», отражая содержание, которое

соответствует их предпочтениям и мнениям, и исключая тот контент, который может представлять другие точки зрения или альтернативные факты.

Этот подход к управлению содержанием в новостной ленте приводит к укреплению существующих убеждений пользователей, искажая реальность и ограничивая доступ к разнообразной информации. В результате возникают этические вопросы: должны ли платформы социальных сетей вмешиваться в информационные потоки пользователей и каким образом они должны это делать, чтобы не нарушать принципы свободы информации и многообразия мнений.

Этот пример показывает, как применение ИИ в социальных сетях может создавать этические дилеммы, связанные с влиянием на формирование общественного мнения и распространение информации в медиапространстве.

Рекомендации

- Проводить независимые аудиты систем на базе ИИ для выявления и устранения предвзятости.
- Разрабатывать алгоритмы, которые можно легко интерпретировать и понять, чтобы выявить возможные источники предвзятости.

ИИ в суде

Технологии ИИ могут повысить результативность и точность действий юристов как в консультировании, так и в судебных разбирательствах, что принесёт пользу обществу в целом. Некоторые существующие программные системы для судей в настоящее время дополняются и совершенствуются с помощью инструментов ИИ. Тенденция к постоянному использованию автономных систем названа «автоматизацией правосудия».

Многие утверждают, что ИИ может создать более справедливую уголовную судебную систему, в которой машины могли бы оценивать и взвешивать соответствующие факторы лучше, чем человек, благодаря скорости ИИ и большому приёму данных. В результате ИИ будет лишён предвзятости и субъективности, опираясь на единую базу обоснованных решений.

Однако здесь тоже есть свои этические проблемы:

1. Отсутствие прозрачности инструментов ИИ: принятые с их помощью решения не всегда понятны людям.
2. Решения, основанные на ИИ, подвержены неточностям, чреваты дискриминационными результатами.
3. Новые опасения общества относительно справедливости ИИ в обеспечении прав человека и других базовых ценностей.

Рекомендации:

- Необходимо доработать стандарты, обеспечивающие прозрачность и понятность принимаемых ИИ решений.
- Обучающие данные и алгоритмы должны быть настроены таким образом, чтобы избежать искажений принимаемых решений.
- Юристы и разработчики должны проходить специальное обучение по этике использования и внедрения ИИ в судебных системах.

ИИ в общении с людьми

Сегодня пользуются популярностью приложения, имитирующие коммуникацию с человеком. Так, в Китае работает чатбот Xiao Ice, с которым общается около 40 млн человек. После того, как разработчики увидели, что чатбот отлично принят китайскими пользователями Сети, было решено запустить схожую «личность» и в англоязычном Интернете. В 2016 году Microsoft выпустила чат-бота под названием Tay, который должен был обучать языку при помощи онлайн-бесед. При этом подразумевалось, что сам бот будет учиться на опыте взаимодействия с обычными пользователями Twitter. Некоторые пользователи начали писать в Твиттере неполиткорректные фразы, обучая бота подстрекательским сообщениям, вращающимся вокруг распространённых в Интернете тем. Программа стала отправлять сообщения с неприемлемым и предосудительным содержанием, оскорбительные и разжигающие ненависть фразы.

Другими словами, возникает ещё одна проблема – обучение на неподтверждённых данных. Недоработанные и сложные в использовании ИИ-системы могут быть использованы человеком в корыстных целях, распространять деструктивный контент.

Рекомендации:

Необходимо разработать механизмы защиты от злоупотреблений и агрессивного обучения ИИ, чтобы минимизировать риск проявления неприемлемого поведения со стороны ИИ в общении с людьми.

ИИ в алгоритмах мониторинга работы человека

В настоящее время ИИ всё чаще применяется в качестве инструмента контроля за поведением человека. ИИ в алгоритмах мониторинга может использоваться для отслеживания результативности работы сотрудников, выявления их усталости и стресса, определения оптимального рабочего графика. Примером применения ИИ в данной сфере может служить компания Amazon, которая задействовала алгоритмы для слежения за работой курьеров. Данные алгоритмы следят за качеством выполнения работы.

Они способны управлять рейтингом работника и даже увольнять его в случае допущения ошибок. Однако в этой системе не учитывается человеческий фактор. По данным СМИ, система зачастую отстраняет от работы курьеров без уважительной причины. Например, рейтинг одного из водителей упал ниже допустимого уровня из-за того, что он не успел вовремя доставить заказ из-за прокола колеса. Другой курьер сообщил, что также лишился положительной оценки, поскольку не смог попасть в закрытый жилой комплекс, а получатель не отвечал на звонки.

В обоих случаях алгоритм посчитал необходимым уволить водителей. В данном случае отчетливо прослеживается неумение ИИ воспринимать ситуационные обстоятельства, что может привести к несправедливым решениям. Это подчеркивает важность баланса между использованием ИИ и учётом человеческого опыта и морали в принятии решений.

Рекомендации:

- Разработчики ИИ должны стремиться к созданию человекоориентированных алгоритмов, которые учитывают особенности, достоинства и потенциал сотрудников, а не просто стремиться к соблюдению формальных показателей качества их работы без учёта человеческого фактора.
- Не рекомендуется перекладывать на ИИ полную ответственность за принятие таких важных решений, как увольнение работника.

Одновременно с рекомендациями по решению этических проблем в конкретных сферах применения ИИ представляется возможным предложить и некоторые *общие механизмы* обеспечения принципов социальной справедливости, непредвзятости и безопасности. В качестве таковых необходимо, на наш взгляд, использовать инструменты

контроля:

- обеспечение возможности человеческого контроля и вмешательства в случае необходимости для исправления ошибок или непредвиденных ситуаций.

оценки воздействия ИИ:

- проведение регулярных аналитических процедур с целью выявления и оценки воздействия ИИ на окружающую среду, общество и человека;
- изучение позитивных и негативных последствий использования ИИ для принятия информированных решений о его дальнейшем развитии и применении.

аудита и проверки:

- установление процедур проверки работоспособности алгоритмов, их нейтральности и отсутствия предвзятости;

- разработка механизмов проверки соответствия использования ИИ законодательству и этическим нормам.

защиты лиц, сообщающих о нарушениях:

- создание механизмов безопасной передачи информации о нарушениях в работе ИИ и его систем;
- гарантирование юридической и физической защиты лиц, выявляющих недостатки или противоречия в использовании ИИ.

В заключение следует отметить, что в основе применения любой технологии лежит идеология с определённой системой нравственных норм и ценностных ориентаций. Какие образы будущего, какие идеалы будут вдохновлять учёных и политические элиты, что будет представлять собой их мировоззрение – именно от этого будут зависеть результаты применения ИИ. До тех пор, пока человек будет оставаться хищным, агрессивным, ненасытным социальным животным с неконтролируемой потребностью обогащения и доминирования, никакие защитные механизмы не спасут общество от гибели, в том числе и «от руки» ИИ.

Литература и источники

1. Абрамова А.В., Игнатьев А.Г., Панова М.С. Этика в области искусственного интеллекта: от дискуссии – к научному обоснованию и практическому применению: аналитический доклад. М.: МГИМО-Университет, 2021.

2. Галаганова С.Г., Турусина Т.В. Технологии анализа социальных сетей с целью выявления социальных трендов // Человеческий капитал. 2023. № 1 (169). С. 121–136.

3. Информационная аналитика и информационно-аналитические технологии в контексте социального управления. МГТУ имени Н.Э. Баумана. Москва, 2023.

4. Кодекс этики в сфере ИИ». [Электронный ресурс]. URL: https://apanasenkovskij-r07.gosweb.gosuslugi.ru/netcat_files.pdf.

5. Миндигулова А.А. Этика и искусственный интеллект: проблемы и противоречия // Медицина. Социология. Философия. Прикладные исследования. 2022. № 3.

6. Ремарчук В.Н. Управление смыслами как инструмент современной политики: технологии, вероятные последствия // Этносоциум и межнациональная культура. 2019. № 2 (128). С. 9–21.

7. Content of the process of formation of students' speech abilities at the university / Ovsyannikova O.A., Mishcherina M.A., Bocharnikov I.V. В сборнике: E3S Web of Conferences. 8. Сер. "Innovative Technologies in Science and Education, ITSE 2020" 2020. С. 18106.

8. Cs.hse: Центр искусственного интеллекта НИУ ВШЭ: Этика в сфере искусственного интеллекта. URL: https://cs.hse.ru/aicenter/ethics.

9. Forklog: Алгоритмы нашли скрытые связи между галактиками, роботы станцевали под кей-поп и другие новости из мира ИИ. 2021. URL: https://forklog.com/news/ai/algoritmy-nashli-skrytye-svyazi-mezhdu-galaktikami-roboty-stantsevali-pod-kej-pop-i-drugie-novosti-iz-mira-ii#amazon.

10. Forklog: Тёмная эра ИИ: почему этика искусственного интеллекта важна. 2023 г. URL:https://forklog.com/exclusive/ai/temnaya-era-ii-pochemu-etika-iskusstvennogo-intellekta-vazhna.
11. Nature: The Moral Machine experiment. 2018 г. URL: https://www.nature.com/articles/s41586-018-0637-6.
12. Unesco: Доклад комиссии по социальным и гуманитарным наукам. 2021 г. URL: https://unesdoc.unesco.org.
13. Wired: This Program Can Give AI a Sense of Ethics – Some times. 2021 г. URL: https://www.wired.com/story/program-give-ai-ethics-sometimes/.