УДК 004.8

ОБЗОР НАДЕЖНЫХ РЕКОМЕНДАЦИЙ ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

Барбахан И. (Университет ИТМО)
Научный руководитель – кандидат физико-математических наук, доцент Фильченков А.А.
(Университет ИТМО)

Эта работа сделана для обзора рекомендаций по использованию надежного искусственного интеллекта, рассмотрения того, почему эта концепция надежности становится необходимостью, и каковы выдающиеся рекомендации и как их достичь в реальной жизни.

Введение.
Надежный искусственный интеллект — это термин, используемый для описания искусственный интеллект, который является законным, соблюдает этические нормы и технически надежен [1]. Он основан на идее, что искусственный интеллект полностью раскроет свой потенциал, когда можно будет установить доверие на каждом этапе его жизненного цикла, от проектирования до разработки, развертывания и использования.

Основная часть.
Для создания надежного искусственный интеллект требуется несколько компонентов:

Конфиденциальность: в дополнение к обеспечению полной конфиденциальности пользователей, а также конфиденциальности данных также необходимы механизмы контроля доступа к данным. Они должны учитывать весь жизненный цикл системы, от обучения до производства, что означает персональные данные, изначально предоставленные пользователем, а также информацию, формируемую о пользователе в процессе его взаимодействия с системой [2].
Надежность: системы искусственный интеллект должны быть надежными и безопасными. Они должны быть точными, способными обрабатывать исключения, хорошо работать с течением времени и быть воспроизводимыми. Еще одним важным аспектом является защита от угроз и нападений со стороны противника. Атака искусственный интеллект может быть нацелена на данные, модель или базовую инфраструктуру. В таких атаках данные, а также поведение системы могут быть изменены, в результате чего система будет принимать другие или ошибочные решения и даже полностью отключится. Чтобы системы искусственный интеллект были надежными, они должны быть разработаны с учетом упреждающего подхода к рискам для минимизации и предотвращения вреда.

Модели искусственный интеллект и их решения часто называют «черными ящиками» из-за того, что даже экспертам трудно понять их внутреннюю работу. Заинтересованные стороны, участвующие в жизненном цикле системы искусственный интеллект, должны понимать, почему искусственный интеллект придумал решение и в какой момент оно могло быть другим. Глядя на пример использования оценки рисков и обнаружения мошенничества на основе искусственный интеллект при расследовании предупреждения о транзакции/входе в систему, необходимо открыть черный ящик и понять, почему событие было помечено как мошенничество таким образом, чтобы его мог интерпретировать человек. Объяснимый искусственный интеллект (XAI) предлагает новые методы определения важных функций, используемых в модели, которые влияют на оценку высокого риска. Это может помочь аналитикам мошенничества решить, следует ли предпринимать дальнейшие действия и со временем выявлять изменяющиеся модели или новые типы мошенничества, или знать, что сказать пользователю, которому было отказано в процессе оплаты. Также важно, чтобы объяснения были представлены в формате, подходящем для заинтересованных сторон, поскольку разным людям требуются разные уровни объяснений. По мере создания систем

объяснимого искусственный интеллект ИТ-специалисты должны решить, какой уровень понимания заинтересованных сторон необходим. Потребуется ли финансовым учреждениям сделать свои платформы искусственный интеллект понятными для инженеров, юристов, специалистов по соблюдению требований или аудиторов? По данным Gartner, к 2025 году 30% государственных и крупных корпоративных контрактов на покупку продуктов и услуг искусственный интеллект будут требовать использования искусственный интеллект, что понятно и этично.

Справедливость: системы искусственный интеллект должны быть справедливыми, беспристрастными и доступными для всех. Скрытые предубеждения в конвейере искусственный интеллект могут привести к дискриминации и исключению недостаточно представленных или уязвимых групп. Обеспечение справедливости систем искусственный интеллект и соответствующих мер защиты от предвзятости и дискриминации приведет к более равному обращению со всеми пользователями и заинтересованными сторонами.

Прозрачность: данные, системы и бизнес-модели, связанные с искусственный интеллект, должны быть прозрачными. Люди должны знать, когда они взаимодействуют с системой искусственный интеллект. Кроме того, соответствующие заинтересованные стороны и потенциальные пользователи должны понимать возможности и ограничения системы искусственный интеллект. В конечном итоге прозрачность будет способствовать более эффективному отслеживанию, аудиту и подотчетности.

Потенциальные риски в области искусственный интеллект требуют участия важнейших заинтересованных сторон в правительстве, промышленности и научных кругах для обеспечения эффективного регулирования и стандартизации. Ранее в этом году Европейская комиссия представила этические рекомендации для Trustworthy AI. Они включают в себя принципы, которые обеспечивают честность, безопасность, прозрачность и полезность систем искусственный интеллект для конечных пользователей. Кроме того, в США Национальный институт стандартов и технологий (NIST) работает над разработкой стандартов и инструментов для укрепления доверия к искусственный интеллект [3].

Выводы.

Системы искусственного интеллекта, способные достичь производительности экспертного уровня или даже превзойти ее, существуют здесь и сейчас, и все больше исследований демонстрируют их потенциал для улучшения результатов лечения пациентов, рабочего процесса медицинских работников и доступа к офтальмологической помощи. Надежный искусственный интеллект — это необходимый следующий шаг, который поможет сократить текущий разрыв между разработкой и интеграцией этих систем в офтальмологию.

Список использованных источников:

1. Abràmoff, Michael D., Danny Tobey, and Danton S. Char. "Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process." American journal of ophthalmology 214 (2020): 134-142.
2. Wing, Jeannette M. "Trustworthy ai." Communications of the ACM 64.10 (2021): 64-71.
3. Floridi, Luciano. "Establishing the rules for building trustworthy AI." Nature Machine Intelligence 1.6 (2019): 261-262.