Конспект лекции

С.С. Жумажанова, П.С. Ложников

Этика и правовые проблемы искусственного интеллекта

Методы оценки рисков, возникающих из­за внедрения ИИ. Передовые методы обнаружения и снижения воздействия ИИ

ДЛЯ ОБУЧЕНИЯ ИИ

МЕТОДЫ ОЦЕНКИ РИСКОВ, ВОЗНИКАЮЩИХ ИЗ­ЗА ВНЕДРЕНИЯ ИИ. ПЕРЕДОВЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ И СНИЖЕНИЯ ВОЗДЕЙСТВИЯ ИИ

Существует ряд юридических, политических, финансовых и регуляторных вопросов, которые необходимо решить, до безопасного использования систем и технологий на базе ИИ. Создание ИИ должно базироваться на специальных принципах (см. тему 2: Принципы ИИ), факт их нарушение может стать прямым источником рисков, которыми могут быть:

*- Автономное вооружение:* ИИ, в случае передачи ему контроля над вооружением, может убивать, например, с использованием ядерного и автономного вооружения государства, либо быть использовано наиболее негуманным сообществом лиц.

*- Манипулирование общественным мнением:* с помощью социальных сетей, благодаря автоматическим алгоритмам, возможно распространение пропаганды лицам, определенным с помощью алгоритмов и личных данных. ИИ может распространять любую информацию, которая потребуется в любом формате, который будет выглядеть наиболее убедительным, неважно, будет ли это правда или ложь.

*- Вмешательство в частную жизнь:* каждый современный человек в той или иной степени оставляет свой цифровой след, и, таким образом, можно отследить и проанализировать каждый шаг человека в интернете, его личную жизнь и рабочую деятельность. Это может привести к использованию данных цифрового следа, например, для взлома банковских счетов, оформления кредитов и т.д.

*- Несовпадение наших целей с целями машины:* важно правильно интерпретировать цели и задачи, которые выполняет для нас ИИ, так чтобы не было иных негативных последствий со стороны ИИ. Например, для быстрой доставки человека из одного пункта в другой, системе на базе ИИ нужно увеличить скорость и вследствие этого человек может пострадать. Важно очень четко оговаривать и цели, и безопасные способы их достижения.

*- Дискриминация*: Те данные, которые собирает ИИ для помощи конкретным группам людей, впоследствии могут быть использованы против них. Так как вместе с основной необходимой информацией всегда собирается какая-нибудь менее релевантная, с использованием которой ИИ может навести ущерб.

NIST SP 800-30: Риск является функцией вероятности того, что данный источник угрозы, использует потенциальную уязвимость организации и осуществит неблагоприятное воздействие на данную организацию

NIST FIPS 200: Риск - уровень воздействия на деятельность организации (включая предназначение, функции, имидж или репутацию), активы организации или людей, следующие из использования информационной системы, подвергаемой потенциальному воздействию угрозы, и вероятность появления угрозы.

ISO/IEC DIS 23894 Information technology — Artificial intelligence — Risk management: Риски ИИ должны быть идентифицированы, количественно или качественно описаны и расставлены по приоритетам в соответствии с критериями оценки рисков и целями организации.

*В качестве примеров источников риска применения ИИ в ГОСТе указаны*:

* неправильное использование или злоупотребление,
* последствия автоматизации,
* непрозрачность,
* непредсказуемость,
* новые угрозы безопасности,
* новые угрозы конфиденциальности,
* неправильное (некорректное или злонамеренное) использование,
* плохая спецификация,
* плохая реализация,
* субъективность и ее запечатление,
* угрозы правам человека.

*Согласно ГОСТу в качестве вариантов обработки рисков организация должна рассмотреть:*

* избегание риска путем принятия решения не начинать или не продолжать деятельность, которая порождает риск;
* принятие или повышение риска для того, чтобы воспользоваться возможностью;
* устранение источника риска;
* изменение вероятности;
* изменение последствий;
* хеджирование риска (например, через контракты, покупку страховки);
* сохранение риска путем принятия обоснованного решения.

*Принципы управления рисками ИИ:*

* неотъемлемость,
* структурированность и всеобъемлемость,
* персонализованность,
* инклюзивность,
* динамичность,
* высокое качество доступной информации,
* человеческие и культурные факторы,
* непрерывное совершенствование.

**Оценка свойств устойчивости систем ИИ.** В Proposal for a Regulation laying down harmonized rules on artificial intelligence (законопроект AIA, апрель 2021 г.) «Проект нормативно-правового акта Европейского парламента и Совета Европы, устанавливающего согласованные правила в отношении искусственного интеллекта (Закон об искусственном интеллекте) и вносящего изменения в некоторые законодательные акты Европейского Союза» используется соразмерный и основанный на оценке риска подход: чем выше риск применения конкретной системы ИИ, тем строже правила.

Выделено четыре уровня рисков:

**Неприемлемый риск**. К этой категории относятся СИИ, которые могут представлять угрозу для прав граждан и их безопасности, например, разработки, позволяющие манипулировать поведением пользователей. Примеры включают социальную оценку со стороны правительства (например, кредитную систему Китая), эксплуатацию уязвимых мест детей и использование подсознательных методов. Подобные системы будут запрещены.

**Высокий риск.** К таким системам относятся решения в области критической инфраструктуры (например, транспорта), медицины (в частности, роботизированная хирургия), образования, биометрическая идентификация, категоризация физических лиц, образование и профессиональная подготовка, трудоустройство, управление персоналом и доступ к самозанятости, доступ к основным частным и государственным услугам и льготам, данные правоохранительных органов, данные миграционных и пограничных служб, данные институтов отправления правосудия и демократических процессов и др. Отнесенные к этой категории разработки должны соответствовать строгим критериям безопасности (иметь четкую систему оценки рисков, контроль со стороны людей и др.).

**Умеренный риск.** Пользователи таких СИИ должны четко понимать, что взаимодействуют с машиной (например, чат-ботом), а не человеком. Системы ИИ с ограниченным (умеренным) риском будут подчиняться конкретным обязательствам по обеспечению прозрачности, чтобы пользователи могли принимать информированные решения, зная, что они взаимодействуют с машиной, и отключаться по желанию.

**Минимальный риск.** В эту категорию попадает большинство ИИ-систем: интеллектуальные спам-фильтры, видеоигры с поддержкой ИИ и так далее. К системам с минимальным риском новые правила не будут применяться, так как представляют минимальный или нулевой риск для прав или безопасности граждан. Компании и пользователи смогут использовать их бесплатно. По мнению членов Европейской комиссии, большинство приложений искусственного интеллекта попадут в эту категорию.

Компания AI Watch провела обзор стандартов по ИИ. Для каждой группы стандартов по ИИ в четырех исследуемых организациях (ETSI, ISO/IEC, ITU, IEEE) отмечена корреляция между стандартом/спецификацией и одним/всеми требованиями законопроекта. Эти требования включают:

* проверенный и качественный набор данных;
* наличие технической документации до выхода продукции на рынок;
* наличие механизма автоматической записи событий;
* прозрачность и доступность информации о системе ИИ для пользователей;
* возможности контроля ИИ человеком;
* точность, надежность и кибербезопасность;
* наличие внутренних проверок систем ИИ;
* наличие системы управления рисками.

В исследовании были представлены шесть стандартов, которые являются основными рабочими элементами для реализации законопроекта AIA, среди них:

* ISO/IEC 4213 ‑ Информационные технологии, ИИ, оценки классификации машинного обучения;
* SO/IEC 5338 – Информационные технологии, искусственный интеллект ‑ процессы жизненного цикла системы ИИ;
* ISO/IEC 23894-2 – Информационные технологии, ИИ, управление рисками;
* ISO/IEC 24027 – Информационные технологии – ИИ ‑ предвзятость в системах искусственного интеллекта и с использованием искусственного интеллекта;
* ISO/IEC 38507 ‑ Информационные технологии, управление ИИ, последствия использования систем ИИ;
* ISO/IEC 42001 ‑ Информационные технологии, ИИ, система управления.

**Список источников**

1. IEEE SA - IEEE 7010-2020 [electronic resource] // Sa main site. URL: <https://standards.ieee.org/ieee/7010/7718/>.
2. ISO/IEC DIS 23894 Information technology — Artificial intelligence — risk management [electronic resource] // ISO. URL: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/73/77304.html>.
3. ГОСТ Р 59276—2020 Системы искусственного интеллекта. Способы обеспечения доверия [electronic resource]. URL: <https://docs.cntd.ru/document/1200177291>.
4. ISO/IEC DTR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns [electronic resource] // ISO. URL: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/85/78507.html>.
5. ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence [electronic resource] // ISO. URL: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77608.html>.
6. ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in ai systems and ai aided decision making [electronic resource] // ISO. URL: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77607.html>.
7. Этика и «цифра»: от проблем к решениям / под ред. Е.Г. Потаповой, М. С. Шклярук. — М.: РАНХиГС, 2021. — 184 с.