**Доверенный искусственный интеллект**

# Управление рисками искусственного интеллекта

А.Е. Самотуга, П.С. Ложников

Конспект лекции

# Управление рисками искусственного интеллекта

Как мы уже говорили, построения доверенных систем ИИ невозможно без внедрения процесса управления рисками ИИ. Любое управление рисками должно иметь комплексный подход, который должен соответствовать потребностям организации. Основные принципы позволяют организации определить приоритеты и принять решения о наиболее эффективном достижении своих целей.

* 1. ***Принципы управления рисками.***

В стандарте ИСО 31000:2018 описываются общие принципы управления рисками, приведем их применение в процессе разработки и использовании ИИ.

Принципы включают: Неотъемлемость, структурированность и комплексный подход, персонализированность, инклюзивность, динамичность, доступная информация, человеческие и культурные факторы, Постоянные улучшения

* 1. ***Обязательства по управлению рисками. Распределение ролей, полномочий и ответственности по управлению рисками.***

Высшему руководству и надзорным органам (в применимых случаях) следует демонстрировать постоянную приверженность управлению рисками. Это может быть реализовано посредством политики, программного заявления или иным способом, четко отражающим цели и приверженность организации риск-менеджменту. Приверженность риск-менеджменту должна включать, но может не ограничиваться следующим:

- нацеленность организации на управление рисками и связи с ее целями и другими политиками;

- укрепление потребности в интеграции риск-менеджмента в общую культуру организации;

- проведение процедур интеграции риск-менеджмента в основные виды деятельности и процессы принятия решений;

- определение полномочий, ответственности и подотчетности;

- обеспечение доступа к необходимым ресурсам;

- способы решения конфликтных задач;

- измерение показателей эффективности организации и отчетность по ним;

- пересмотр и улучшение. Приверженность риск-менеджменту следует подобающим образом распространять внутри организации и доводить до сведения заинтересованных сторон.

Высшему руководству и надзорным органам (в применимых случаях) следует обеспечивать, чтобы полномочия, ответственность и подотчетность для соответствующих ролей в отношении управления рисками определялись и доводились до сведения соответствующих лиц на всех уровнях организации, а также необходимо:

- акцентировать внимание на том, что риск-менеджмент является основной обязанностью;

- определить лиц, подотчетных и имеющих полномочия по управлению рисками (владельцы рисков).

Высшее руководство и/или контролирующие органы (где таковые имеются), должны распределить ресурсы и назначить лица:

- уполномоченные устранять риски ИИ;

- ответственные за создание и отслеживание процессов устранения рисков ИИ.

* 1. ***Процесс управления рисками. Критерии рисков ИИ. Оценка рисков ИИ. Идентификация рисков ИИ. Анализ рисков ИИ. Классификация рисков ИИ.***

Поскольку процесс риск-менеджмента может применяться на разных уровнях (например, стратегическом, операционном, программном, проектном или др.), важно четко понимать затрагиваемую риск-менеджментом сферу охвата, соответствующие цели, которые необходимо учитывать, и их согласованность с целями организации. При планировании подхода необходимо учитывать следующее:

- цели и решения, которые необходимо принять;

- ожидаемые результаты от шагов, предпринимаемых в рамках этого процесса;

- время, местоположение, определенные допущения и исключения;

- соответствующие инструменты и методы оценки рисков;

- требуемые ресурсы, зоны ответственности и документацию (отчетность);

- взаимное влияние с другими проектами, процессами и действиями.

Организациям следует учитывать современное техническое состояние при выборе и применении соответствующих инструментов и методов оценки рисков в отношении своих целей.

Риски ИИ должны быть идентифицированы, количественно или качественно описаны и расставлены по приоритетам в соответствии с критериями оценки рисков и целями организации.

Оценку рисков следует рассматривать как комплексный процесс, который следует применять на протяжении всего жизненного цикла ИИ-услуги или ИИ-продукта. Поэтому организации, занимающиеся разработкой или предоставлением систем ИИ, должны соотносить свою деятельность по оценке управления рисками с жизненным циклом системы. Различные методы оценки рисков могут применяться на разных этапах жизненного цикла.

Для идентификации применяется пункт 6.4.2 руководства, представленного в ИСО 31000:2018 и дополнительные рекомендации приведем ниже.

При определении рисков систем искусственного интеллекта следует учитывать различные источники рисков, в зависимости от характера рассматриваемой системы и контекста ее применения.

Автономность-это степень, в которой система искусственного интеллекта работает независимо от человеческого надзора и контроля.

Уровень автономности системы ИИ определяет, какой объем информации о фактическом поведении системы доступен организации. Системы с высоким уровнем автономности могут демонстрировать неожиданное поведение, которое может быть трудно обнаружить и контролировать. Высокоавтономные системы могут создавать риски, связанные с их безопасностью и надежностью.

Степень прозрачности принятия решений определяет, какой объем информации о внутренних процессах системы искусственного интеллекта доступен организации. Системы с низкой степенью прозрачности принятия решений могут создавать риски в отношении их объективности, прозрачности и подотчетности. С другой стороны, высокая степень прозрачности решений может привести к путанице. Важно найти соответствующий уровень прозрачности принятия решений, который поможет целевой аудитории понять результаты работы системы и доверять им.

Сложность среды системы ИИ определяет диапазон потенциальных ситуаций, с которыми система ИИ должна справляться в своем рабочем контексте. Как правило, более сложные среды могут привести к ситуациям, которые не были учтены на этапе проектирования системы искусственного интеллекта. Следовательно, сложные среды могут создавать риски, связанные с надежностью и безопасностью системы искусственного интеллекта.

Системная инженерия включает в себя проектирование, спецификацию, внедрение, а также проверку и валидацию системы искусственного интеллекта. Неадекватные или недостаточные методы и процессы, применяемые на каждом из этих этапов, могут привести к низкому качеству, следовательно, к небезопасным, небезопасным, ненадежным или предвзятым системам.

Технология искусственного интеллекта все больше зависит от пользовательских микросхем ASIC, а не от стандартного оборудования. Это может привести к появлению новых категорий угроз безопасности, особенно в условиях, когда доступ к оборудованию для машинного обучения предоставляется внешним пользователям (например, в приложениях облачных вычислений). Пользователи, имеющие доступ к чипам ускорителя, потенциально могут использовать этот доступ для взлома системы, кражи интеллектуальной собственности или повреждения оборудования.

Технологическая готовность показывает, насколько зрелой является данная технология в данном контексте применения. Менее зрелые технологии, используемые при разработке и применении систем искусственного интеллекта, могут создавать риски, которые неизвестны организации или трудно поддаются оценке. С другой стороны, для зрелых технологий может быть доступно большее многообразие данных, что облегчает выявление и оценку рисков. Однако, если технологии являются зрелыми, существует также риск самоуспокоенности и технического долга, поэтому это зависит от постоянного мониторинга рисков и надлежащего технического обслуживания.

Для систем искусственного интеллекта, управляемых данными, процессы, используемые для сбора данных, являются источником рисков, которые особенно трудно диагностировать и обнаружить.

Анализ рисков может проводиться с различной степенью детализации в зависимости от критичности ценностей или ранее выявленных активов организации, а также степени неблагоприятных и положительных воздействий, которые уже могут быть известны, и любых накопленных знаний, опыта, включая инциденты (историю), связанные с организацией

Любая методология анализа рисков может быть качественной или количественной, или их комбинацией, в зависимости от того, что подходит организации. Возможно, будет целесообразно сначала использовать качественный анализ для получения общего представления об уровне риска и выявления основных рисков. Позже может потребоваться провести более конкретный или количественный анализ основных рисков, поскольку, как правило, качественный анализ менее сложен и менее затратен, чем количественный.

Каждый выявленный источник риска должен быть оценен с использованием всей соответствующей информации или данных.

Консервативно, для рисков или опасных ситуаций, которые не могут быть ранжированы (например, вероятность наступления последствий не может быть оценена), следует определить и проанализировать возможные последствия.

Организациям следует рассмотреть возможность проведения априорного анализа, а также анализа на ходу. Априорный анализ включает, например:

─ Строгость в подготовке данных, включая проверку на предвзятость (т. е. объективность), включение различных точек зрения и релевантности, достаточный объем и распространение проверенных данных по мере необходимости для целей проблемы;

─ обоснованность критериев, определяющих выбор алгоритмов, которые будут применяться, и их гибридизацию;

─ наличие экспертных знаний в области применения и использования алгоритмов.

В ходе анализа механизмов ИИ рассматриваются риски, связанные с проектированием систем, их обслуживанием, включая адаптацию, использование и переподготовку. Аспекты, как риски, связанны с:

─ Риском объяснимости;

─ рисками прозрачности;

─ рисками вероятности.

Апостериорный анализ, касается рисков, связанных с принятием решений.

Варианты обработки рисков, определенные организацией, должны быть разработаны таким образом, чтобы снизить риски до приемлемого уровня. Если выяснится, что требуемое снижение риска не может быть достигнуто путем применения различных вариантов обработки рисков, организация должна провести анализ рисков и выгод для остаточных 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.

В дополнение к вариантам устранения рисков, следует изучить примеры конкретных контрмер/средств контроля рисков ИИ, которые организация должна рассмотреть в целях устранения источника риска.

Как только меры по устранению рисков будут реализованы в соответствии с планом устранения рисков, любой остаточный риск должен быть оценен в соответствии с зарегистрированными критериями риска. Если остаточные риски считаются неприемлемыми с использованием этих критериев, следует применять дополнительные меры по контролю рисков. Для остаточных рисков, оцененных как приемлемые, организация может решить, какие остаточные риски должны быть включены в сопроводительные документы продукта.

**Список источников**

1. ISO 31000:2018 [Электронный ресурс]. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html (дата обращения: 12.04.2022).

2. ISO/IEC DIS 23894 [Электронный ресурс]. URL:<https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/73/77304.html> (дата обращения: 12.04.2022).