Об управлении рисками искусственного интеллекта:
руководство Международной организации
по стандартизации (ISO)
и основы управления рисками Национального института
по стандартам и технологиям
Департамента торговли США (AI RMF NIST)

Международная организация по стандартизации (ISO) выпустила руководство ISO/IEC 23894:2023 по управлению рисками искусственного интеллекта (ИИ)[1].

Руководство направлено на те организации, которые разрабатывают, внедряют или используют системы искусственного интеллекта в своей деятельности.

Руководство состоит из трех частей:

- основополагающие принципы риск-менеджмента;
- структура управления рисками;
- процессы управления рисками.

В ISO/IEC 23894 используются международные стандарты управления рисками ISO 31000:2018, которые содержат установленные общие рекомендации для риск-менеджмента[2]. При подходе к определению рисков ИИ и управлению ими разработчики ISO/IEC 23894 делают ссылку на общие принципы менеджмента рисков и признают излишним выработку новых принципов управления рисками ИИ.

Несмотря на то, что можно полностью полагаться на общие принципы риск-менеджмента, существуют факторы риска, характерные исключительно для функционирования ИИ. Системы ИИ работают на гораздо более сложном уровне, чем другие технологии, и это приводит к большему количеству

[1] https://www.iso.org/standard/77304.html.
[2] https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en.

источников риска. ISO/IEC 23894 закрепляет данные факторы в *Приложении С «Управление рисками и жизненный цикл системы ИИ»*.

Согласно документу, менеджмент рисков ИИ должен быть направлен на обеспечение следующих характеристик:

- компетентность систем ИИ;
- доступность и качество тестирования данных;
- воздействие на окружающую среду;
- справедливость;
- пригодность к ремонту;
- конфиденциальность;
- прочность;
- надежность;
- безопасность;
- прозрачность и объяснимость.

Основными факторами риска, в соответствии с документом, являются:

- недостаточная прозрачность и объяснимость;
- уровень автоматизации;
- факторы риска, связанные с машинным обучением;
- проблемы системной аппаратуры;
- проблемы жизненного цикла системы;
- технологическое развитие.

В январе 2023 года Национальным институтом по стандартам и технологиям **Департамента торговли США** выпущены основы риск-менеджмента искусственного интеллекта (The Artificial Intelligence Risk Management Framework (**AI RMF**)[3].

[3] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Данный документ издан ранее руководства ISO/IEC 23894, поэтому его основой стали более ранние руководства ISO, в частности, вышеупомянутый документ ISO 31000:2018.

Американский риск-менеджмент ИИ закрепляет «**характеристики надежного ИИ**», к которому должны стремиться все процессы по внедрению и использованию ИИ:

- действительность и надежность;
- безопасность;
- защищенность и устойчивость;
- подотчетность и прозрачность;
- объяснимость и интерпретируемость;
- обеспечивание конфиденциальности
- справедливость к пагубной предвзятости.

Как и ISO/IEC 23894:2023, документ AI RMF прямо закрепляет отличия рисков ИИ от традиционных рисков программного обеспечения (ПО) (*Приложение В*). Среди них, например: непредвиденные изменения во время машинного обучения; высокая степень сложности прогнозирования побочных эффектов; расширенные возможности агрегирования данных.

Согласно документу, управление рисками характерно для таких стадий функционирования ИИ, как: (1) проектирование (2) разработка (3) внедрение (4) оценка (5) использование.

Менеджмент рисков ИИ в документе разделен на 4 стадии: управление, планирование, измерение, менеджмент рисков.

Управление

В рамках данного аспекта руководящими органами организаций, проектирующих, разрабатывающих, внедряющих, проводящих оценку или приобретающих системы ИИ, могут быть приняты следующие меры:

- выработка политики управления рисками;
- прозрачное описание процессов прогнозирования, выявления и управления рисками;

- оценка потенциального воздействия;
- определение ответственных лиц;
- и др.

Планирование

Данная функция во многом связана с анализом каждой стадии жизненного цикла ИИ. В рамках данного направления выделяются следующие направления:

- установление контекста определения рисков, связанных с системой ИИ;
- категоризация систем ИИ;
- оценка возможностей ИИ, его целевого использования, ожидаемых выгод и затрат по сравнению с соответствующими контрольными показателями;
- сопоставление рисков и преимуществ для всех компонентов системы ИИ, включая стороннее программное обеспечение и данные;
- характеристика воздействия на отдельных лиц, группы, сообщества, организации и общество в целом.

Измерение

В контексте данного направления деятельности по управлению рисками используются количественные, качественные или смешанные инструменты, методы и методологии для анализа, оценки и мониторинга рисков ИИ и связанных с ним воздействий. В рамках этой функции используются данные, собранные в ходе процесса Планирования, а результаты, полученные на выходе, становятся основой для следующего направления – Менеджмента рисков.

Измерение включает в себя отслеживание метрик вышеуказанных «характеристик надежного ИИ», социального воздействия и взаимодействия человека с ИИ.

Все процессы тестирования, оценки, проверки, включая используемые метрики, методы и методологии, документируются. Каждые действия в

рамках процедуры Измерения должны соответствовать научным, правовым и этическим нормам и осуществляться в рамках открытого и прозрачного процесса.

Соблюдение данных требований в организациях может быть сопряжено с выработкой новых типы качественных и количественных измерений.

Менеджмент рисков

Информация, полученная в результате консультаций с экспертами и вклада всех лиц, задействованных в рамках функции Управления, и обработанная в ходе Планирования, используется на этой стадии для снижения вероятности сбоев системы ИИ и иных негативных последствий. Практика систематического документирования подкрепляет усилия по управлению рисками ИИ и повышает прозрачность и подотчетность.

В рамках данной функции выполняются:

- приоритезация рисков ИИ;
- выработка стратегий максимизации преимуществ ИИ и минимизации негативных последствий;
- контроль рисков и преимуществ ИИ от сторонних организаций;
- обработка рисков.

Таким образом, второй документ более детально закрепляет направления деятельности по управлению рисками, связанными с ИИ, на каждой стадии его жизненного цикла. Предполагается, что AI RMF будет постоянно дополняться и обновляться, поэтому в нем остаются нерассмотренными следующие риски:

- проблема вредоносной предвзятости в системах ИИ;
- проблемы безопасности, связанные с машинным обучением;
- учет нарушений безопасности;
- риски, связанные с использованием сторонних технологий ИИ (вне организации).

В российской экспертной практике рекомендации по управлению рисками ИИ содержатся в **Кодексе этики искусственного интеллекта**[4]. В частности, субъектам, использующим системы ИИ, рекомендуется проводить оценку потенциальных рисков применения ИИ, в том числе – с помощью независимого аудита, и выработать соответствующие методики оценки рисков.

Кроме того, Кодексом рекомендуется разработка методик, руководств, «чек-листов» и иных методических материалов.

С учетом того, что искусственный интеллект, равно как и традиционные программные технологии, подвержен быстрым инновациям, Кодексом этики предусмотрена возможность создания сводов наилучших/наихудших практик решения вопросов, возникающих в жизненном цикле ИИ.

[4] https://ethics.a-ai.ru/assets/2022/12/30/%D0%9A%D0%BE%D0%B4%D0%B5%D0%BA%D1%81_%D1%8D%D1%82%D0%B8%D0%BA%D0%B8_%D0%B2_%D1%81%D1%84%D0%B5%D1%80%D0%B5_%D0%98%D0%98_%D1%84%D0%B8%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B8.pdf.