Анализ рисков контента, созданного искусственным интеллектом

Коптев Дмитрий Владимирович
Студент гр. 4116, ГУАП, dmitrii.koptev76@gmail.com

Аннотация. Рассматриваются риски и проблемы распространения нейросетей, способных создавать комплексный контент. При анализе рисков были выделены несколько главных проблем, которые впоследствии были подробно разобраны.

Ключевые слова: искусственный интеллект, нейросеть, машинное обучение. контент, генерация, обучение, распространение.

Risk analysis arising from the use of content created by AI

Koptev Dmitrii Vladimirovich
Student 4116, SUAI, dmitrii.koptev76@gmail.com

Annotation. The risks and issues of widespread distribution of neural networks capable of creating complex content have been reviewed. Several main problems have been identified during the risk analysis, which were subsequently thoroughly examined.

Keywords: artificial intelligence, neural network, machine learning, content, generation, training, distribution.

Введение. Искусственный интеллект (ИИ) - это область компьютерных наук, которая занимается созданием программ и систем, выполняющих задачи, требующие обычно человеческого интеллекта. Нейросети, способные генерировать изображения, аудио, тексты и более сложный контент, стали широко распространены среди обычных пользователей. Это происходит благодаря развитию методов глубокого обучения искусственного интеллекта в целом, и возникновению простых для пользователя инструментов взаимодействия. Можно выделить ряд рисков и проблем использования ИИ пользователями.

Нарушение конфиденциальности и безопасности. Ястреб Н. А. в своей статье рассматривает проблему размытия границ персональных данных в современном мире [6]. ИИ во время обучения и использования оперирует огромными объёмами данных, в том числе и персональными данными пользователей. И крупная фирма, и обычный пользователь могут обмениваться с нейросетью важными документами, расшифровками переговоров, адресами, изображениями, записями голоса и иной подобной информацией. Все эти данные хранятся и обрабатываются на серверах используемой нейросети и могут легко попасть в открытый доступ, далее оказаться в руках злоумышленников или быть использованы третьими лицами без согласия хозяина.

Этические проблемы. Ивлев Д. В. в своей статье разделяет этических проблем использования ИИ на два уровня: на систему взаимодействия ИИ и общества и систему взаимодействия ИИ и человека [3]. Первая система включает в себя вопросы автоматизации, роботизации, концепции умного

города и дома. Во вторую систему входят вопросы общения и психологического комфорта пользователя при взаимодействии с нейросетями.

Рассмотрим один из аспектов взаимодействия ИИ и человека. С нейросетью могут взаимодействовать люди в разном психическом состоянии, в том числе и глубоко подавленные люди с депрессией. У таких людей могут возникнуть губительные для себя или окружающих намерения. Пользователь сформулирует соответствующий запрос к нейросети (например, создание запрещённых химических соединений, оружия или нанесение себе физических увечий). Подобные задачи нейросеть не отличит от обычных безвредных запросов и выдаст исчерпывающий ответ на них.

Развитие киберпреступности. Минбалеев А. В. в своей работе рассматривает потенциальное участие ИИ в киберпреступности [4]. Среди проблем наиболее специфичны дистанцирование злоумышленника и жертвы, облегчение рутинных задач преступника и повышение степени его анонимности.

Рассмотрим основные сценарии взаимодействия злоумышленника и нейросети:

1. Злоумышленник, используя нейросеть, может создать виртуального двойника реального человека. Существует широкий спектр средств для генерации реалистичных изображений и голосов людей. Так, получив запись речи близких пользователя или достаточный набор фотографий (например, из соцсетей), злоумышленник может на основе этих данных генерировать неотличимые от настоящих изображения и голоса друзей и родственников жертвы, создать правдоподобный образ с целью втереться в доверие. Подобные технологии создают практически безграничные возможности для шантажа и махинаций.
2. Обучающая выборка или готовый запрос составляется человеком в соответствии со своими вкусами и предпочтениями. Нейросети же действуют в соответствии с запросом, заданными настройками и обучающим массивом данных. Модели ИИ могут делать случайно или заведомо неверные генерации.

Такой контент попадает на онлайн хостинги. Любой веб-браузер ищет тексты и изображения по ключевым фразам и словам. Браузер находит сгенерированный ИИ контент, не отличает его от реального и представляет пользователю в первых строчках выдачи. Эти ситуации вводят обычного пользователя в заблуждение. Наиболее опасно выглядит ситуация преднамеренного лоббирования таких изображений в поисковой выдаче. Это может быть использовано, например, в политических целях, чтобы понизить или повысить чей-то рейтинг соответствующими материалами.

3. Злоумышленники часто используют такое средство, как DDoS-атака на веб-серверы с целью их отказа. В таких операциях огромное значение играют искусственно созданные пользователи – боты. Технологии детекции таких ботов и борьбы с ними всё время улучшаются, но нейросети позволяют создавать безграничные числа ботов, которые будут неотличимы от реальных пользователей. Такие боты могут содержать привязанные аккаунты на различных площадках, у них будет страничка в социальных сетях и

правдоподобный цифровой след и поведение в сети. Таких ботов крайне сложно отследить и отличить от реальных людей. Также нейросети могут быть обучены и использованы для создания фишинговых сайтов, писем, компьютерных вирусов.

Недостаточность правовой базы. Наиболее остро стоит вопрос о юридическом статусе ИИ. Как утверждает Поздеева В. С., проблема состоит в разграничении зон ответственности между разработчиком, пользователем ИИ и самим ИИ [5]. Можно рассмотреть два подхода к данной проблеме:

- ИИ необходимо рассматривать как самостоятельное юридическое лицо;
- человек-оператор и нейросеть должны быть связаны общей ответственностью перед законом.

Barfield W. в своей книге поднимает вопрос о неопределённости ответственности за действия ИИ [1]. Рассмотрим данную проблему в контексте авторских прав уже сгенерированного контента. Например, разработчик обучил нейросеть на работах известного современного художника. Этой нейросетью по запросу пользователя были сгенерированы изображения с целью их коммерческого использования. Происходит размытие ответственности, потому что нельзя с уверенностью сказать кому принадлежит авторское право в рассматриваемой ситуации. Это может быть либо художник, либо разработчик, либо пользователь.

Любая нейросеть должна быть обучена на некотором наборе данных. Обучающие выборки формируют разработчики ИИ. Данные для выборки состоят из контента, созданного реальным человеком, будь то изображения, текст или аудио. Разработчик, обучая нейросеть на таких данных, пользуется чужой интеллектуальной собственностью. Некоторые крупные компании не раскрывают массивы обучающих данных, используемые для тренировки ИИ, специально, чтобы скрыть факт отсутствия разрешения на использования контента от истинных владельцев данных. Это приводит к тому, что контент, полученный нейросетью, может быть использован в коммерческих целях, а также выдан за результат чужого интеллектуального труда без ведома автора.

Экономические проблемы. Дадашев З. Ф. и Устинова Н. Г. в своей статье формулируют проблему внедрения ИИ в мировой рынок труда [2]. Распространение ИИ в рабочем процессе приведёт к миграции спроса у работодателей от кандидатов, выполняющих рутинный труд, к кандидатам, занимающимся социальной, познавательной или творческой деятельностью, а также к тем, чью деятельность трудно или невозможно автоматизировать. В первое время такое данная тенденция может привести к росту безработицы.

Данную проблему можно проиллюстрировать сейчас ситуацией в индустрии изобразительного искусства и дизайна. Безусловно, потребность в художниках останется, но рутинная часть данного ремесла, будь то создание графического дизайна, интерфейса приложения или иллюстраций будет частично выполняться нейросетями.

Данный процесс основан на способности ИИ обучаться на огромном массиве данных и выполнять генерации контента, ничем не отличимые от создаваемого реальным человеком. Всё, что потребуется - это обучить

нейросеть на работах художника, чей стиль удовлетворяет заказчика, и дальше такая нейросеть способна практически бесконечно выдавать похожие изображения по запросу. Для функционирования такой системы потребуется разработчик, настроивший данную нейросеть и, оператор-редактор, который будет формулировать запросы и вносить правки в созданный контент.

Вывод. Развитие и свободное распространение нейросетей может сопровождаться множеством проблем. Человечество стоит перед проблемой регулирования и ограничения искусственного интеллекта, но в условиях свободного интернета и рынка отсутствует возможность контроля и блокировки популяризации и расширения ИИ. Это вызвано тем, что технологии ИИ не монополизированы и могут свободно распространятся с открытым исходным кодом, что означает безграничное копирование и развитие текущих моделей крупными фирмами и независимыми разработчиками. В связи с этим, пользователям придётся привыкнуть к новой реальности сосуществования с искусственным интеллектом, в которой человечество оказалось после создания нейросетей.

Список литературы:

1. ***Barfield W.*** Research Handbook on the Law of Artificial Intelligence – Edward Elgar Publishing, 2018, 736 С.
2. ***Дадашев З. Ф.***, Устинова Н. Г. Влияние искусственного интеллекта на экономику // Эпоха науки. - 2019. - №18. - С. 53 – 57.
3. ***Ивлев Д. В.*** Искусственный интеллект и проблемы этики // Право и практика. - 2023. - №4. - С. 263 – 267.
4. ***Минбалеев А. В.*** Проблемы использования искусственного интеллекта в противодействии киберпреступности // Вестник ЮУрГУ. Серия: Право. - 2020. - №4. - С. 116 – 120.
5. ***Поздеева В. С.*** Правовое регулирование и область применения искусственного интеллекта // Вопросы российской юстиции. - 2022. - №22. - С. 249 – 260.
6. ***Ястреб Н. А.*** Как проблема персональных данных меняет этику искусственного интеллекта? // Философские проблемы информационных технологий и киберпространства. - 2020. - №1 (17). - С. 29 – 44.