Институт статистических
исследований и экономики знаний

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

ТРЕНДЫ | РАЗРАБОТКИ | ЧЕЛОВЕЧЕСКИЙ КАПИТАЛ
ИНФРАСТРУКТУРА | ИСПОЛЬЗОВАНИЕ | ЭФФЕКТЫ

№ 8 / 2024

Институт статистических исследований и экономики знаний (ИСИЭЗ) НИУ ВШЭ анализирует вызовы в сфере цифровой безопасности, связанные с применением решений на основе искусственного интеллекта (ИИ), и рассматривает доступные их разработчикам и пользователям методы противодействия кибератакам.

Данный выпуск новой серии информационно-аналитических материалов подготовлен в рамках проекта «Мониторинг технологического развития искусственного интеллекта в Российской Федерации» тематического плана научно-исследовательских работ, предусмотренных Государственным заданием НИУ ВШЭ.

КАК ПОВЫСИТЬ БЕЗОПАСНОСТЬ РАЗРАБОТКИ И ИСПОЛЬЗОВАНИЯ ИИ?

В последние годы многократно возросла актуальность вопросов безопасности в цифровой среде на фоне общей активизации киберпреступности и появления новых способов воздействия на информационные системы, реализуемого в том числе посредством ИИ-решений.

Справочно: За 2023 г. количество утечек информации ограниченного доступа в мире превысило показатели предыдущего года более чем на 61%. При этом более 47 млрд записей персональных данных были скомпрометированы (стали известны посторонним лицам). Одна из причин роста киберинцидентов – распространение сервисов на основе ИИ. К примеру, в США 75% специалистов по кибербезопасности крупных компаний (численностью более 1 тыс. сотрудников) относят использование персоналом генеративного ИИ к факторам увеличения числа и скорости кибератак.

Согласно результатам обследования 2.5 тыс. организаций 20 отраслей экономики (обрабатывающая промышленность, торговля, финансы и страхование, транспорт и логистика, ИТ-отрасль, телекоммуникации и др.), проведенного ИСИЭЗ НИУ ВШЭ в конце 2023 г., регулирование вопросов информационной безопасности при использовании продуктов и услуг на основе ИИ находится на шестом месте среди стимулов, наиболее значимых для расширения внедрения этого класса технологий. Необходимость повышения требований к кибербезопасности, устанавливающих правила *применения* ИИ-систем, отметила почти каждая четвертая из опрошенных организаций – пользователей подобных решений (более 23%).

При этом в случае *разработки* ИИ аналогичные требования законодательного урегулирования вопросов информационной безопасности фигурируют в ответах лишь 8.6% респондентов. Недостаточная заинтересованность в этом вопросе на этапе разработки может повлечь серьезные последствия и ущерб для компаний, ведь **функция безопасности должна закладываться на каждом этапе создания и использования ИИ-решений и гарантироваться пользователю «по умолчанию»**, наряду с базовым функционалом конкретного продукта. Для этого на каждом из этапов жизненного цикла технологий ИИ могут применяться разные методы (табл. 1).

Таблица 1. Методы противодействия атакам на ИИ-решения на этапах их разработки и использования

ЭТАПЫ: ⚙ разработка 👤 использование / МЕТОДЫ:	⚙ Подготовка обучающих данных	⚙ Формирование классификатора	⚙ Обучение ИИ-модели в основе ИИ-решения	⚙ 👤 Использование ИИ-решения
Анализ подозрительной активности				
Использование нескольких классификаторов				
Ограничение использования данных с открытых платформ				
Определение алгоритмов безопасного обучения				
Очистка и сравнение данных с исходными				
Очистка обучающей выборки от отравляющих атак				
Применение антивирусного ПО и межсетевого экрана				
Своевременное обновление ОС и ПО				
Состязательное обучение				
Статистический и динамический анализ кода				
Фильтрация входных данных				

ИСИЭЗ НИУ ВШЭ на основе докладов Kaspersky, Federal Office for Information Security (Germany).

Особое внимание разработчики уделяют **обучающим данным** для классификатора машинного обучения, от которого зависит корректное функционирование ИИ-решения. Специалисты по кибербезопасности призывают проводить очистку данных, сравнивать их с исходными, производить переобучение на примерах тренировочных наборов данных, отслеживать нацеленные на порчу обучающей выборки отравляющие атаки (призваны спровоцировать ошибочность дальнейшей работы классификатора). На этапе эксплуатации ИИ-решения применяется фильтрация входных данных, включающая предварительную проверку обрабатываемой информации на предмет наличия в ней вредоносного содержимого. Специалисты по кибербезопасности также рекомендуют на разных этапах разработки и эксплуатации ИИ-решения осуществлять статистический и динамический анализ кода (статистический подразумевает выявление ошибок программы без ее запуска, динамический – при ее выполнении).

Крайне важно при **формировании классификатора** опираться на безопасные алгоритмы, разработанные с учетом специфики отрасли компании и обрабатываемых данных. Один из методов инициирования злоумышленником ошибок нейросети – состязательные атаки (в т.ч. ранее упомянутые отравляющие). Напротив, **состязательное машинное обучение** помогает распознать действия злоумышленника по манипуляции данными с помощью намеренного провоцирования ошибок в поведении нейронной сети.

На всех этапах разработки ИИ-решения рекомендуется **ограничить применение данных с открытых платформ**, а также проводить дополнительную проверку данных от третьих сторон. Согласно некоторым оценкам, крупнейшая облачная платформа для размещения проектов и совместной разработки GitHub на начало 2024 г. содержала около 100 тыс. зараженных репозиториев (хранилищ данных). Основным типом атаки на GitHub является практика подмены репозиториев под схожими именами. В результате пользователи могут по ошибке загрузить с платформы вредоносную версию репозитория вместо заслуживающей доверия исходной и подвергнуться целому ряду неправомерных действий, включая передачу учетных данных для входа в приложения, паролей и других конфиденциальных данных.

Вне зависимости от этапа разработки и использования ИИ-решения базовыми остаются **традиционные средства защиты и правила цифровой гигиены**, в т.ч. использование лицензированного ПО, установка надежных паролей, резервное копирование данных, посещение только проверенных веб-сайтов и др. По мере распространения ИИ интенсивность атак и ущерб от них будут стремительно расти. Это делает особенно важным проведение регулярного мониторинга и повышение уровня цифровой грамотности работников и их корпоративной лояльности.

В России вопросам безопасности в связи с развитием ИИ уделяется все большее внимание. В июле 2024 г. принят закон, содержащий положения об обязательном страховании участниками экспериментального правового режима ответственности за вред, причиненный жизни, имуществу и здоровью граждан вследствие использования технологий ИИ. Ранее в этом году при поддержке Минцифры России был создан консорциум, нацеленный на изучение возможностей по обеспечению безопасности технологий ИИ. Дальнейшая дискуссия вокруг данной тематики должна учитывать отраслевую специфику применения решений на основе ИИ, оценку общих и специфических рисков, их потенциальный ущерб и влияние на субъектов экономической деятельности. Помимо государственных инициатив безопасность ИИ должна стать одним из ключевых приоритетов разработчиков и пользователей, а ее обеспечение – включать технические, организационные, нормативные и социальные инструменты.

Авторы: **А. И. Фокина, Ю. В. Туровец**