**УДК 004.056.53**

# ПОДХОД К ОБНАРУЖЕНИЮ ВНУТРЕННЕГО НАРУШИТЕЛЯ ЗАЩИЩЕННОСТИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

**Асадуллин А.Я., Менисов А.Б.** (ВКА им. А.Ф.Можайского)
**Научный руководитель – докторант, кандидат технических наук, Менисов А.Б.**
(ВКА им. А.Ф.Можайского)

В исследовании приводится подход к обнаружению внутреннего нарушителя защищенности систем искусственного интеллекта. Рассматриваются способы воздействия внутреннего нарушителя на системы искусственного интеллекта, предлагается метод обнаружения злонамеренных действий.

**Введение.** В настоящее время рыночная доля информационных решений на основе искусственного интеллекта составляет в 119,78 млрд долларов. Прогнозируется, что к 2030 он достигнет 1 597,1 млрд долларов при зарегистрированном среднегодовом темпе роста 38,1% в период с 2022 по 2030 год. Нельзя игнорировать тот факт, что искусственный интеллект содержит особенные уязвимости характерные только для компонентов систем искусственного интеллекта. Следствием эксплуатации этих уязвимостей является колоссальный ущерб организации, особенно объектов критической информационной инфраструктуры [1], таких как: информационные системы, информационно-телекоммуникационные сети, автоматизированные системы управления субъектов критической информационной инфраструктуры.

**Основная часть.** Внутренний нарушитель в системе искусственного интеллекта может нанести ущерб путем неправильного обращения со статистическими данными, эксплуатацией или физическим повреждением библиотеки и платформы машинного обучения, фальсификацией данных, неправильным использованием модели машинного обучения, неправильной конфигурации модели и др. [2].

Подход к обнаружению внутреннего нарушителя заключается в следующем:

1. Сбор данных действий пользователей с компонентами систем искусственного интеллекта (анализ метрик и показателей, анализ эффективности внедренного функционала, анализ больших данных и др.)
2. Построение вектора действий пользователей по категориям: пользователь систем искусственного интеллекта, пользователь систем искусственного интеллекта-нарушитель, разработчик, разработчик-нарушитель.
3. Создание конвеера с моделями машинного обучения: метод k-ближайших соседей, метод опорных векторов, логистическая регрессия, наивный байес.
4. На основании работы моделей машинного обучения определяется важность признаков действий внутреннего нарушителя [3, 4].

**Выводы.** Практическая значимость подхода обнаружения внутреннего нарушителя защищенности систем искусственного интеллекта заключается в возможности его применения при обосновании и разработке технических решений информационной безопасности.

**Список использованных источников**:

1. Федеральный закон от 26 июля 2017 года № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации».

2. Ломако А. Г., Менисов А. Б. Ландшафт угроз безопасности информации технологий искусственного интеллекта // Сборник 31-й научно-технической конференции методы и технические средства обеспечения безопасности информации. – 2022. – С. 49.

3. Melis M. et al. secml: A python library for secure and explainable machine learning //arXiv preprint arXiv:1912.10013. – 2019.

4. Kumar R. S. S. et al. Adversarial machine learning-industry perspectives //2020 IEEE Security and Privacy Workshops (SPW). – IEEE, 2020. – С. 69-75.