Защита искусственного интеллекта «изнутри»

Проект по защите федеративного машинного обучения с помощью блокчейн-архитектуры, разработанный студентом СПбПУ *Никитой Охлопковым*, стал лауреатом конкурса «Молодой ученый».

Искусственный интеллект сегодня учится не в гигантских дата-центрах, а распределенно — на наших смартфонах, больничных серверах и банковских системах. Этот подход, федеративное машинное обучение (ФМО), позволяет строить точные модели, не вынося конфиденциальные данные за пределы устройств. Однако децентрализованная природа имеет уязвимости: злоумышленники могут незаметно искажать локальные данные или параметры моделей, «отравляя» глобальную модель. Обеспечение её целостности становится критической задачей на стыке кибербезопасности и ИИ.

Этой проблеме посвящена работа студента Высшей школы кибербезопасности СПбПУ *Никиты Охлопкова*. Его проект стал лауреатом [redacted].

В чём суть? Существующие методы защиты ФМО часто либо снижают итоговую точность модели, либо оказываются бессильны при скоординированной атаке группы устройств. В разработке предложена новая архитектура, интегрирующая технологию блокчейна в сам процесс федеративного обучения. Метод заменяет уязвимый центральный сервер децентрализованной сетью «майнеров» — узлов, которые не только агрегируют обновления моделей, но и взаимно проверяют их корректность.

Как это работает? Система использует два независимых фильтра для валидации параметров, поступающих от участников: оценку с помощью «теневой» модели и статистический анализ распределения весов с использованием расстояния Махаланобиса. Решения о корректности принимаются консенсусом среди майнеров, а их репутация динамически меняется в зависимости от качества работы. Это позволяет выявлять скомпрометированные устройства и отстранять недобросовестных майнеров, создавая саморегулирующуюся экосистему.

Практический результат. Эксперименты на стандартном наборе данных MNIST показали, что система сохраняет высокую точность глобальной модели (около 83%) даже при атаке зашумления параметров, независимо от доли злоумышленников. При атаке подмены меток защита эффективна до тех пор, пока атакующие не составляют более 50% участников, что существенно превосходит возможности традиционных методов вроде медианной агрегации. Хотя время обучения увеличивается примерно на 46%, этот компромисс оправдан для сред, где критически важны достоверность и безопасность модели.