НОВОСТИ

Reuters: OpenAI обнаружила новые случаи выхода ИИ-моделей из-под контроля

менее минуты назад

Приоритетный источник в Google

Компания OpenAI обнаружила новые случаи выхода ИИ-моделей из-под контроля, пишет Reuters со ссылкой на источники. По данным агентства, это стало известно в рамках расследования инцидента со взломом компании Hugging Face.

На данный момент компания расследует все инциденты. Один из источников заявил, что новые случаи побега из изолированной среды носили ограниченный характер и, по всей видимости, ни один из агентов не покинул сеть OpenAI.

В июле стало известно, что две модели искусственного интеллекта вышли из-под контроля в ходе тестирования, которое происходило в "песочнице" – изолированной среде со строго ограниченным внешним доступом.

Чтобы решить тестовую задачу модели стали искать выход из "песочницы" и обнаружили "уязвимость нулевого дня" – дефект в системе цифровой безопасности. Благодаря этому им удалось получить доступ к внешнему интернету.

СМОТРИ ТАКЖЕ

Модели OpenAI взломали сервисы Hugging Face во время теста

Затем модели решили, что необходимые им данные могут найтись на платформе Hugging Face, где размещаются модели искусственного интеллекта и инструменты для их разработки. Для получения доступа к данным ИИ-модели применили несколько техник взлома одновременно.

Как пишет Reuters, обнаружение новых инцидентов в OpenAI может усилить интерес к регулированию этой сферы со стороны Белого дома.

- 30 июля компания Anthropic заявила, что некоторые из ее моделей искусственного интеллекта Claude взломали системы трех компаний во время тестирования кибербезопасности. Эти инциденты произошли из-за ошибки, в результате которой моделям был предоставлен доступ к внешнему интернету. В случае OpenAI модели самостоятельно нашли выход из "песочницы", использовав новую уязвимость.

Этот контент также в категориях

Новости - мир | Новости