Как искусственный интеллект учится манипулировать людьми и что с этим делать.

OpenAI — это компания, которая занимается исследованиями и разработкой в области искусственного интеллекта. Она была основана в 2015 году и создана с целью сделать искусственный интеллект более доступным и безопасным для людей. Компания разрабатывает и использует нейронные сети и другие методы искусственного интеллекта для решения различных задач, включая анализ данных, генерацию текста, голоса, изображений и т.д." data-html="true" data-original-title="OpenAI" >OpenAI, создатель ChatGPT, сообщила, что обнаружила использование своих инструментов искусственного интеллекта для создания и распространения дезинформации в операциях, связанных с Китаем, Ираном и Израилем. В условиях обострения информационной войны на фоне выборов, технологии становятся мощным оружием.

Компания из Сан-Франциско в своем отчете заявила, что несколько скрытых операций использовали её модели ИИ для генерации текста и изображений в больших объемах с меньшим количеством ошибок, чем ранее. Это включало создание комментариев или ответов на собственные публикации. Политика OpenAI запрещает использование её моделей для обмана или введения в заблуждение.

Содержание дезинформации касалось таких тем, как «конфликт в Газе, выборы в Индии, политика в Европе и США, а также критика китайского правительства со стороны китайских диссидентов и иностранных правительств», говорится в отчете OpenAI.

Сети также использовали ИИ для повышения собственной продуктивности, применяя его для отладки кода и исследования активности в социальных сетях.

Платформы социальных сетей, включая Meta и YouTube от Google, стараются сдерживать распространение дезинформационных кампаний, особенно после выборов президента США в 2016 году, когда была выявлена попытка манипуляции голосами неизвестной тролль-фабрикой.

Растет давление на компании, занимающиеся искусственным интеллектом, такие как OpenAI, поскольку быстрые достижения в их технологиях делают создание реалистичных дипфейков и манипуляцию медиа дешевле и проще. В условиях, когда около 2 миллиардов человек готовятся к выборам в этом году, политики настоятельно призывают компании внедрять и соблюдать соответствующие меры безопасности.

Бен Ниммо, главный исследователь по разведке и расследованиям в OpenAI, на звонке

с журналистами отметил, что кампании не смогли «значительно» увеличить свое влияние с помощью моделей OpenAI. Однако, он добавил: «Сейчас не время для самоуспокоения. История показывает, что операции по влиянию, которые годами терпели неудачи, могут внезапно добиться успеха, если за ними не следить».

OpenAI, поддерживаемая Microsoft, заявила, что привержена выявлению таких дезинформационных кампаний и разрабатывает собственные инструменты на базе ИИ для более эффективного обнаружения и анализа. В отчете также говорится, что системы безопасности компании уже затрудняют работу злоумышленников, так как модели отказываются генерировать требуемый текст или изображения.

OpenAI также раскрыла, что несколько известных государственных дезинформационных акторов использовали её инструменты. Среди них например китайская сеть Spamouflage, продвигающая интересы Пекина за рубежом. Кампания использовала модели OpenAI для создания текста или комментариев на нескольких языках перед публикацией на платформах, таких как X.

Кроме того, была отмечена ранее не сообщавшаяся операция под названием Bad Grammar, использовавшая модели OpenAI для отладки кода и создания коротких политических комментариев, которые затем публиковались в Telegram.

Также сообщается, что OpenAI пресекла прокитайскую дезинформационную кампанию, якобы управляемую из Тель-Авива компанией STOIC, которая использовала модели для генерации статей и комментариев в в различных соцсетях.