OpenAI, лидер в области исследований ИИ, недавно представила новый классификатор изображений. Этот инструмент может идентифицировать изображения, созданные их собственным мощным генератором изображений DALL-E 3, с точностью 98%. Однако его эффективность в обнаружении контента, созданного другими моделями ИИ, в настоящее время ограничена.

Чтобы решить эту проблему, OpenAI поддерживает стандарт C2PA, который встраивает устойчивые к взлому метаданные в контент, созданный ИИ. Эти метаданные действуют как цифровой ярлык, указывая источник и, возможно, инструмент, использованный для создания контента. OpenAI надеется, что широкое внедрение C2PA повысит точность обнаружения.

Однако эксперт по безопасности Стив Гробман из McAfee утверждает, что водяных знаков, подобных метаданным C2PA, недостаточно. Хитрые злоумышленники могут просто удалить такие маркеры. Вместо этого McAfee разрабатывает детектор дипфейков, который анализирует уже аудио на предмет признаков манипуляций ИИ. Их инструмент использует сам искусственный интеллект для выявления подозрительных паттернов в речи, таких как необычная лексика или отредактированные аудиофрагменты.

Хотя детектор McAfee многообещающий, он также не является надежным. Он может генерировать ложные срабатывания, ошибочно отмечая реальные аудиозаписи как поддельные. И OpenAI, и McAfee активно совершенствуют свои инструменты, причем OpenAI обращается за помощью к общественным исследованиям, а McAfee фокусируется на уменьшении количества ложных срабатываний.