Методика базируется на идее о том, что любую нейросетевую модель можно изменить так, чтобы она реагировала на определенные цифровые «водяные знаки». Важно, чтобы эти знаки сохраняли свою эффективность даже после модификации модели, отметили в пресс-службе.

Исследователи проверили свой подход на популярной нейросети ResNet34, используемой для классификации изображений. Они разработали набор «водяных знаков» и проверили, сохранит ли модель способность реагировать на них после модификаций.

Эксперименты показали, что новый подход позволяет выявить неправомерное использование модели в 73−100% случаев, превосходя существующие методики. Ученые надеются, что их разработка поможет разработчикам ИИ эффективнее защищать свои проекты, и уже выложили ее в открытый доступ.