Компьютер и человек – кто кого?

> «Компьютер делает не то, что вы хотите, а то, что вы ему приказываете»
>
> Автор неизвестен

Продолжение. Начало в №№ 7, 8, 9/2023

Арсений Ворошилов, по материалам Reuters

Похоже, человечество сейчас переживает настоящий бум, связанный с развитием и использованием искусственного интеллекта. Достаточно сказать, что словарь Коллинза (Collins Dictionary) признал словом года словосочетание Artificial Intelligence, то есть искусственный интеллект.

И здесь, как и практически всегда, когда появляется новая технология, ее развитие и внедрение опережает формирование соответствующей нормативно-правовой базы, что неизбежно вызывает проблемы. Одной из наиболее очевидной и острой применительно к обучению моделей искусственного интеллекта для создания неких творческих произведений стала проблема авторских прав. А конкретнее, авторских прав на те данные, которые используются для обучения моделей AI. Ведь в результате модель AI начинает копировать, к примеру, характер живописи того или иного художника, музыкальный стиль конкретного исполнителя и/или композитора, манеру пения и др.

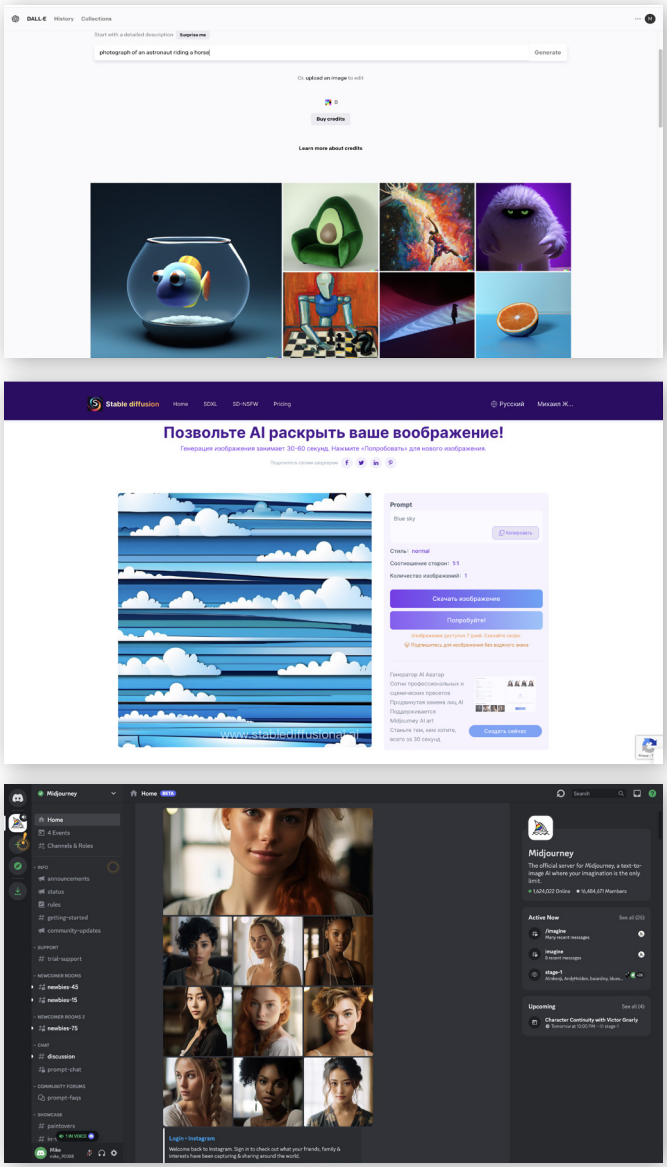
В настоящее время нет никаких конкретных правовых норм, определенным образом регулирующих подобные действия, но уже есть масса недовольных (и справедливо недовольных) правообладателей, не согласных с тем, что плоды их творчества бесконтрольно и без разрешения применяются для обучения моделей AI. И есть разработчики программного обеспечения, разделяющие это недовольство и готовые предложить собственные решения, пока законодатели разных стран заняты чем-то другим. Оказалось, что отравить можно не только живое существо, но и цифровой визуальный образ, причем выглядеть для зрителя он будет без изменений, а весь «яд» достанется модели искусственного интеллекта, если этот «отравленный» образ применить для ее обучения.

Искусственный интеллект: яд и противоядие – разница в дозировке

Борьба с использованием данных для обучения моделей искусственного интеллекта становится все более ожесточенной.

Новый инструмент, получивший название Nightshade («Паслен»), предназначен для того, чтобы пользователи как бы присоединяли его к результатам своей творческой деятельности, чтобы тот искажал – или отравлял – данные, в которых это произведение используется для обучения моделей AI. В конечном счете это может разрушить будущие модели платформ искусственного интеллекта, такие как DALL-E, Stable Diffusion и Midjourney, поскольку лишит их возможности создавать изображения.

Суть действия Nightshade заключается в том, что пиксели в части цифрового художественного произведения подвергаются невидимым изменениям. Когда это произведение используют для обучения модули AI, введенный в изображение цифровой «яд» использует уязвимости в системе безопасности, чтобы сбить модель с толку, после чего она больше не может распознавать изображение достоверно. Например, изображение автомобиля больше не воспринимается как автомобиль, а вместо этого распознается моделью как корова или иной объект, не имеющий ничего общего с автомобилем.



AI-платформы DALL-E, Stable Diffusion и Midjourney

Один из создателей Nightshade, профессор Университета Чикаго Бен Чжао надеется добиться выравнивания

баланса и устранения его перевеса в сторону AI-компаний, которые не стесняются использовать защищенные авторским правом данные для обучения своих моделей. Утверждается, что данные, используемые для обучения AI-моделей, преобразующих текст в изображение, уязвимы для атак того типа, которые свойственны этому новому средству.

Даже сами создатели Nightshade удивились тому, что даже небольшое число атак, организованных с помощью этого средства, могут дестабилизировать общие функции генеративной модели преобразования текста в речь, эффективно блокируя ее способность создавать значимые образы.

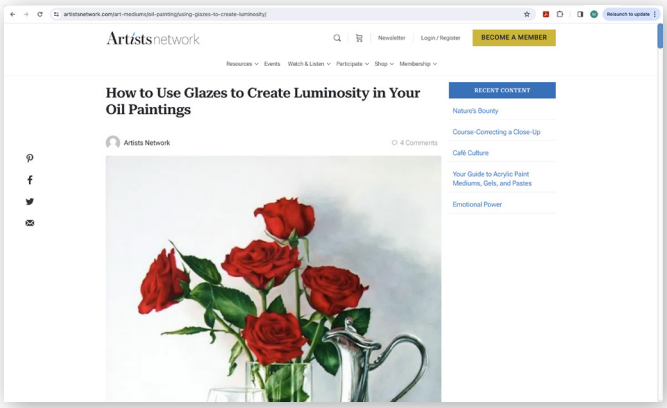
Вот что утверждается в научной работе «Реагирующие на запрос отравляющие атаки на модели преобразования текста в изображение» (Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models), авторы которой Шон Шен, Вэньсинь Динь, Джозефин Пассананти, Хайтао Чжэн и Бен И. Чжао: *«Атаки по отравлению данных манипулируют данными для обучения, чтобы вызвать непредсказуемое поведение моделей, подвергаемых машинному обучению, в процессе этого обучения. Для генеративных моделей преобразования текста в изображение, для которых применяются огромные объемы наборов данных, нынешнее понимание отравляющих атак предполагает, что для успешной атаки потребовалось бы ввести миллионы отравляющих образцов в конвейер обучения моделей. Мы же показываем, что отравляющие атаки могут быть успешными и даже в ограниченных масштабах они делают модели уязвимыми к атакам, сформированным по конкретному запросу, что опирается на способность модели реагировать на тот или иной запрос.*

Мы разработали Nightshade – оптимизированный инструмент отравляющей атаки по конкретному запросу, где отравляющий образец выглядит визуально идентичным к отравляемым им изображениям, которые соответствуют текстовым запросам.

Отравляющие образцы Nightshade также оптимизированы по эффективности и могут повредить запрос Stable Diffusion SDXL с помощью не более 100 отравляющих образцов. Отравляющие воздействия Nightshade «просачиваются» в схожие концепции, а многочисленные атаки можно объединять в реакции на один и тот же запрос.

Мы предлагаем использовать Nightshade и аналогичные средства как последнюю защиту создателей контента от инструментов web-скрейпинга, которые игнорируют директивы opt-out/do-not-crawl».

У тех же, кто создает различные визуальные произведения и хочет воспрепятствовать их неправовому использованию, есть возможность загрузить плод своей работы в Glaze – на специальный портал для художников, и там применить инструмент Nightshade, который маскирует художественный стиль автора произведения. В частности, картина, созданная в стиле классической реалистичной живописи, в результате может быть распознана моделью AI как нечто в стиле кубизма. Nightshade в составе Glaze позволяет пользователям выбрать, что именно они хотят получить, – использовать «таблетку яда» или просто сделать так, чтобы модель AI не смогла имитировать их личный художественный стиль.



Портал Glaze – здесь вскоре появится возможность применять Nightshade

В настоящее время проблемы авторских прав применительно к созданному с помощью AI контенту и обучающим данным остаются в серой зоне, поскольку отсутствует правовая база, регулирующая все эти аспекты. В судах рассматривается множество соответствующих исков. Но пока что методы противодействия использованию данных из Интернета для обучения моделей AI без разрешения авторов этих данных ограничиваются блокированием доступа для скрейперов. Такие компании как Adobe планируют маркировать то, что сгенерировано средствами AI, параллельно информируя, кому принадлежат авторские права на исходный контент.

Некоторые из первых судебных исков против платформ генеративного AI касались защищенного авторским правом материала и его применения в качестве обучающих данных для моделей AI, особенно для платформ преобразования текста в образ. В начале нынешнего года уже три художника подали в суд на Stable Diffusion, Midjourney и художественный сайт DeviantArt, обвиняя их в том, что модели этих ресурсов обучались на произведениях данных художников без получения на это разрешения. Ресурс Getty Images также подал иск против Stable Diffusion, прежде чем создал свой собственный AI-генератор изображений, обученный на принадлежащих Getty Images лицензированных образах.

Но есть и апологеты генеративного AI даже в его нынешней – «дикой» – форме. IT-гиганты Google и Microsoft заявили, что готовы принять на себя судебную ответственность, если на некоторых клиентов подадут в суд за нарушение авторских прав при использовании их решений для генеративного AI. Правда, подавляющее большинство этих решений пока не предназначено для создания образов и ограничивается только работой с текстами.

В общем, борьба за «чистоту спорта» в сфере искусственного интеллекта только начинается. Возможности AI, несомненно, будут стремительно расти, вскоре они сделают возможным имитацию не только статичных образов, но и видеоизображений. И тогда несложно будет научить AI генерировать видеоряд «под Тарантино», «под Фрейзера», «под Аронофски» и так далее. Как всегда, яд и противоядие идут рука об руку, представляя собой, по сути, один и тот же препарат, эффект от использования которого определяется в большинстве случаев только дозировкой.

Продолжение следует