Игнорирование общепринятых протоколов ИИ-компаниями может привести к хаосу в интернете.

Компания Perplexity, позиционирующая свой продукт как «бесплатную поисковую систему на базе искусственного интеллекта», оказалась в центре скандала. После обвинений Forbes в краже материалов и их перепубликации на различных платформах, издание Wired сообщило, что Perplexity игнорирует протокол исключения роботов (robots.txt) и осуществляет несанкционированный сбор данных с сайтов Wired и других изданий медиахолдинга Condé Nast. Технологический сайт The Shortcut также выдвинул аналогичные обвинения.

Теперь, по данным Reuters , Perplexity — не единственная компания, игнорирующая robots.txt и сканирующая сайты для получения контента, который затем используется для обучения их технологий. Агентство ссылается на письмо от TollBit, стартапа, который помогает издателям заключать лицензионные сделки с компаниями, использующими ИИ. В письме сообщается, что «ИИ-агенты из множества источников (не только одной компании) выбирают обход протокола robots.txt для извлечения контента с сайтов».

Robots.txt – это простой, но эффективный инструмент, с помощью которого владельцы сайтов управляют индексацией поисковыми роботами. Несмотря на то, что его использование носит рекомендательный характер, он применялся с 1994 года.

TollBit не указал конкретные компании, однако Business Insider сообщил, что 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 и Anthropic — это компания, которая разрабатывает искусственный интеллект с целью создания более устойчивого и разумного будущего. Специалисты Anthropic разрабатывают алгоритмы и модели, которые могут упростить управление такими сложными системами, как экономика, политика и т.д. Компания также работает над созданием более демократических и прозрачных систем принятия решений на основе искусственного интеллекта." data-html="true" data-original-title="Anthropic" >Anthropic — создатели чат-ботов ChatGPT и Claude соответственно — также игнорируют сигналы robots.txt. Оба этих разработчика ранее заявляли о соблюдении инструкций «не сканировать», указанных в robots.txt файлах.

В ходе собственного расследования Wired обнаружил, что машина на сервере Amazon, «определенно управляемая Perplexity», обходила инструкции robots.txt на сайте издания. Для подтверждения того, что Perplexity сканирует их контент, Wired предоставил инструменту заголовки своих статей и краткие описания материалов. В результате он выдал тексты, «сильно напоминающие» статьи Wired и «практически без указания авторства».

В интервью Fast Company генеральный директор Perplexity Аравинд Сренивас отрицал преднамеренное игнорирование robots.txt. Он объяснил, что компания использует сторонние веб-сканеры в дополнение к своим собственным, и что сканер, выявленный Wired, был одним из них. На вопрос Fast Company о том, сказал ли Perplexity поставщику сканера прекратить сканирование сайта Wired, он ответил лишь, что «это сложно».

Сренивас попытался оправдать действия компании, заявив, что протокол исключения роботов – «не юридическая база», и предложил издателям и ИИ-компаниям выстраивать новые взаимоотношения. Он также намекнул, что Wired намеренно использовал подсказки, чтобы заставить чат-бот Perplexity вести себя определенным образом, и обычные пользователи не получат таких же результатов. Относительно неточной информации, сгенерированной инструментом, Сренивас сказал: «Мы никогда не утверждали, что никогда не галлюцинируем».