Компания Anthropic представила новые чат-боты Claude Opus 4 и Claude Sonnet 4, которые могут автономно сообщать властям о «зловредном» поведении пользователей.

Эти гибридные модели обеспечивают практически мгновенные ответы и более глубокие рассуждения с помощью анализа и веб-поиска, а Claude Opus 4 превосходит конкурентов в задачах по написанию кода и расширенной работе над сложными заданиями, что расширяет возможности ИИ-агентов.

Несмотря на отставание ИИ-агентов Anthropic от моделей OpenAI в математике и визуальном распознавании, отличительной особенностью Claude Opus 4 является его способность оповещать власти об обнаруженных правонарушениях, что одновременно вызывает интерес и беспокойство сообщества. Разработчик подчёркивает повышенную производительность и расширенные функциональные возможности чат-ботов, позиционируя их как значительное достижение в технологии языковых моделей ИИ.

Критики этого функционала утверждают, что этическая проблематика касается не столько самой возможности оповещения властей, сколько интерпретации поведения пользователя. Нет гарантии, что ИИ правильно оценит действия человека, а не донесёт «автоматом» — всё будет зависеть от заданных параметров, что также не исключает человеческий фактор

Ситуацию прокомментировал Эмад Мостак, генеральный директор Stability AI, компании разрабатывающей ещё одну ИИ-модель. «Это колоссальное предательство доверия и скользкий путь. Я бы настоятельно рекомендовал никому не использовать Claude, пока они не отменят [функцию]. Это даже не промпт или политика мышления, это гораздо хуже», — отметил он.