Учёные из OpenAI нашли в искусственном интеллекте особенности, которые отвечают за разные «личности» модели — в том числе те, что вызывают нежелательное или токсичное поведение. Эти «личности» — внутренние сигналы в системе, которые влияют на ответы ИИ (начинает лгать или советовать вредные вещи, например).

Исследователи смогли управлять уровнем токсичности, меняя всего один параметр в модели. Это открытие поможет лучше понять, почему ИИ иногда ведёт себя неправильно, и как сделать его «безопаснее».

Подобные «личности» в ИИ похожи на работу мозга человека, где определённые нейроны связаны с настроением или поведением. Кроме токсичности, у модели есть и другие «личности», например, сарказм. Интересно, что неправильное поведение модели можно исправить, переобучив её на примерах безопасного кода — достаточно всего нескольких сотен примеров.