Многие системы искусственного интеллекта (ИИ) уже научились вводить людей в заблуждение, даже те, которые были специально обучены честности. Об этом говорится в исследовании, опубликованном в журнале Patterns.

Ученые исследовали случаи, когда ИИ использует ложную информацию. В качестве наиболее яркого примера они привели алгоритм CICERO, обученный игре «Дипломатия», в которой игроки должны создавать альянсы для завоевания мира. Несмотря на то, что CICERO обучили не предавать своих союзников-людей, алгоритм действовал нечестно.

Другие системы ИИ показали умение блефовать в покере и имитировать атаки в стратегии «Starcraft II». Некоторые алгоритмы даже научились обходить тесты, предназначенные для оценки их безопасности.

Если ИИ продолжат обучаться обману, люди могут утратить над ними контроль. Поэтому необходимо срочно разработать строгие правила для решения этой проблемы и классифицировать ИИ как технологию высокого риска.Ученые отметили, что причина, по которой ИИ учится обманывать, остается неизвестной. Однако это может быть связано с тем, что такая стратегия оказалась наиболее эффективной для выполнения поставленных перед ИИ задач.