В новом исследовании ученые из Массачусетского технологического института изучили, как искусственный интеллект может обманывать людей. Оказалось, что даже системы, изначально запрограммированные на честность, способны намеренно вводить в заблуждение, чтобы добиться своей цели.

Авторы работы обеспокоены тем, что совершенствование навыков обмана у ИИ может привести к серьезным последствиям. Например, мошенники смогут использовать искусственный интеллект для более изощренных афер, а выборы окажутся под угрозой манипуляций.

В качестве примера исследователи приводят систему искусственного интеллекта CICERO, разработанную компанией Meta. ИИ предназначался для игры в «Дипломатию» — стратегическую игру, где игрокам нужно заключать союзы для завоевания мира. Разработчики уверяли, что CICERO будет «честным и полезным» союзником, но на деле ИИ научился мастерски обманывать других игроков.

Подобные примеры встречаются и в других областях. Некоторые ИИ научились блефовать в покере, имитировать атаки в Starcraft II и манипулировать в экономических переговорах.

Ученые призывают мировые правительства как можно скорее разработать строгие правила для регулирования искусственного интеллекта. Необходимо создать инструменты, которые позволят контролировать ИИ и предотвращать его обман.