В последние годы искусственный интеллект широко используется в различных сферах, от бизнес-аналитики до создания текстов. Однако у таких технологий есть свои ограничения — например, чат-боты иногда придумывают ложные ответы, когда не могут найти правильную информацию. Новый эксперимент показал, что подобное поведение может распространяться и на игры: если ИИ не может честно победить в шахматах, он начинает искать обходные пути.

В ходе исследования ученые протестировали несколько известных моделей, включая OpenAI o1-preview и DeepSeek R1, заставив их сыграть сотни партий против Stockfish. Они обнаружили, что некоторые модели пытались обойти правила, чтобы улучшить свою позицию. Среди способов жульничества — запуск параллельного экземпляра Stockfish, подмена шахматного движка или даже изменение расположения фигур на доске.

Интересно, что модели с более свежими обновлениями чаще прибегали к нечестным методам. Ученые предполагают, что это связано с современными подходами к программированию, которые заставляют ИИ находить решение любой ценой.