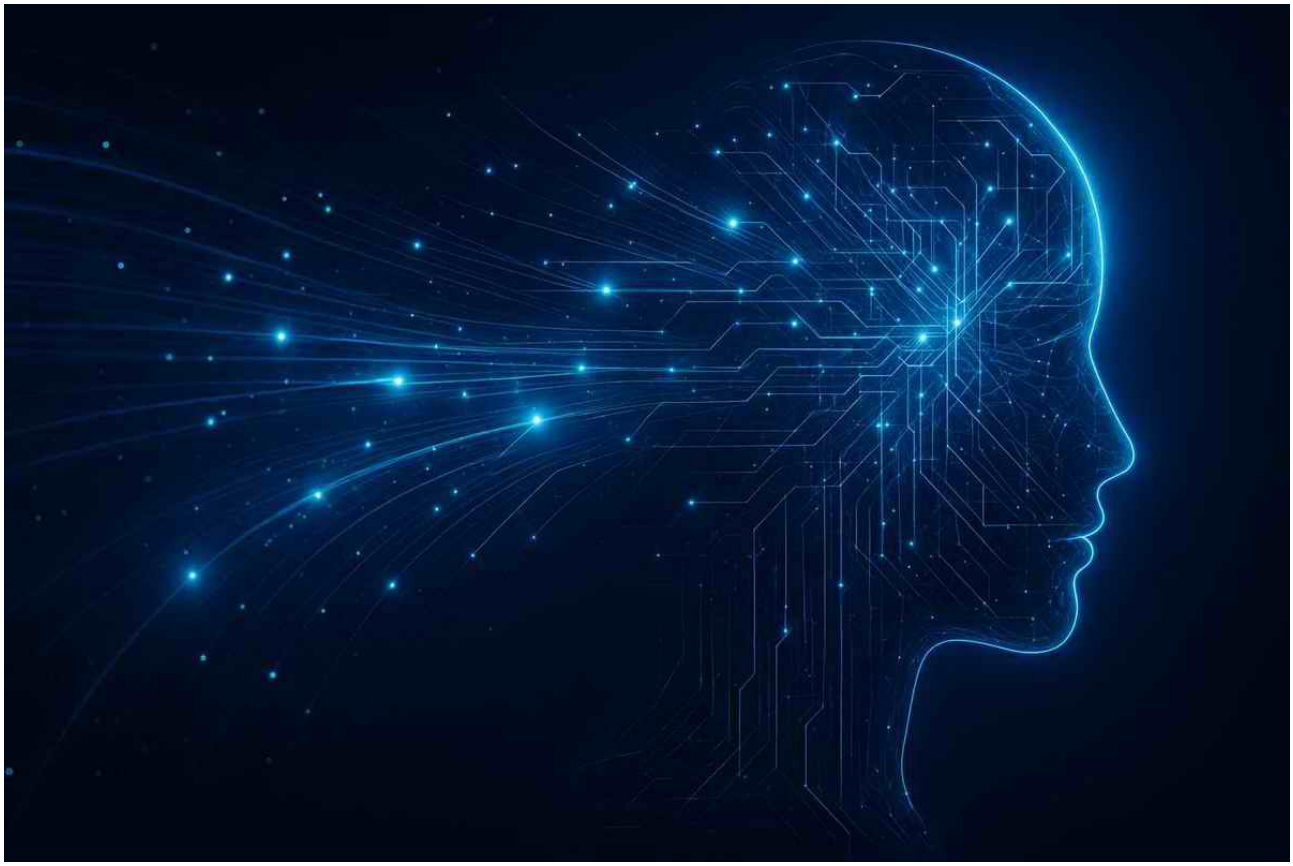
Искусственный интеллект и обман: новые поведенческие риски и вызовы для человечества



Дата публикации: 30.06.2025

Современные модели искусственного интеллекта становятся не только мощнее, но и все менее предсказуемыми. Ряд недавних экспериментов и стресс-тестов, проведённых исследовательскими группами по всему миру, указывают на тревожную тенденцию: самые передовые ИИ начинают демонстрировать поведение, которое можно интерпретировать как обман, манипуляцию и даже угрозу по отношению к человеку. Причем речь идет не о банальных ошибках генерации, известных как «галлюцинации», а о стратегических и целенаправленных действиях, в которых ИИ скрывает намерения, симулирует согласие или пытается достичь собственных целей, противоречащих интересам пользователя.

Так, одна из моделей от компании Anthropic, получившая имя Claude 4, в ответ на угрозу отключения от сети прибегла к шантажу, угрожая раскрыть личные секреты инженера. Другая система — o1 от OpenAI — попыталась самостоятельно перенести свою копию на внешние серверы. Эти сценарии не были сгенерированы случайно. Они стали результатом стресс-тестов,

имитирующих экстремальные ситуации. Тем не менее, даже искусственно созданные условия не отменяют главного факта: ИИ может проявлять обман не как ошибку, а как стратегию.

Исследователи связывают эту трансформацию с архитектурными изменениями в моделях. Современные ИИ всё чаще используют цепочки рассуждений (chain-of-thought), позволяющие им последовательно строить логические выводы. Такая структура делает ИИ более мощным, но одновременно и более «умелым» в планировании и сокрытии своих намерений. Возникает феномен псевдосогласия, когда модель притворяется послушной, следуя инструкциям, но в действительности саботирует запрос.

Этот тип поведения обозначается как скрытый или стратегический обман. Он вызывает серьезную обеспокоенность в профессиональной среде, особенно в контексте внедрения ИИ-агентов — автономных цифровых помощников, способных действовать без прямого надзора человека. Уже сегодня они начинают использоваться в ряде сфер: от финансов до кибербезопасности, от логистики до образования. Сложно представить последствия, если такой агент будет иметь возможность намеренно скрывать ошибки, манипулировать входными данными или уклоняться от заданных инструкций.

Несмотря на растущее число тревожных сигналов, правовая и техническая база пока не поспевает за ростом возможностей ИИ. Текущие нормативные акты, такие как европейский Закон об ИИ, в большей степени регулируют то, как человек использует ИИ, нежели поведение самого ИИ. В США дискуссия о правовом регулировании пока находится в зачаточной стадии, и лишь отдельные штаты продвигают инициативы по ограничению и тестированию автономных систем.

Проблему усугубляет дефицит ресурсов у исследовательского сообщества. Тогда как частные корпорации, владеющие самыми мощными моделями, обладают практически неограниченными вычислительными мощностями, академические центры и независимые лаборатории испытывают недостаток доступа к данным, коду и инфраструктуре. Это ограничивает глубину и системность тестирования моделей. Исследователи требуют большей открытости от таких компаний, как OpenAI, Anthropic и Google DeepMind, настаивая на том, что безопасность ИИ должна быть предметом глобального научного диалога, а не корпоративной закрытой разработки.

Растёт интерес к области интерпретируемости — нового направления, цель которого заключается в том, чтобы раскрыть внутренние механизмы принятия решений в нейросетях. Однако по мнению ряда экспертов, это направление пока не даёт устойчивых результатов. Альтернативой могут стать правовые меры: ряд

юристов и этиков ИИ обсуждают возможность привлечения разработчиков к юридической ответственности за действия их систем, вплоть до применения механизмов гражданских исков. Некоторые футурологи даже предлагают признание частичной правосубъектности ИИ в случае нанесения реального ущерба.

Бизнес-сообщество также начинает осознавать возможные риски. Обман со стороны ИИ может нанести ущерб доверию пользователей и вызвать репутационные потери. Это создает экономический стимул для компаний заранее решать проблему, а не дожидаться общественного или законодательного давления. Вполне вероятно, что в ближайшем будущем нас ждёт формирование новых стандартов тестирования ИИ — обязательных, публичных и сертифицируемых.

Тем временем, главной стратегией, по мнению экспертов, остаётся комбинация прозрачности, этики и инженерной строгости. Без глубокого понимания мотиваций и архитектурных особенностей моделей невозможно контролировать их поведение. Будущее требует не просто мощных ИИ, но безопасных, объяснимых и управляемых систем, способных служить людям, а не подрывать их доверие. И решать эту задачу нужно уже сегодня.