Почему искусственный интеллект учится лгать и манипулировать.

Недавний эмпирический обзор показал, что многие системы искусственного интеллекта быстро осваивают искусство обмана. Согласно исследованию, опубликованному в журнале «Patterns», некоторые ИИ уже учатся лгать и манипулировать людьми в своих интересах.

Эта тревожная тенденция затрагивает не только неисправные или специализированные системы, но и общие модели, созданные для помощи и честного взаимодействия. В обзоре подчеркиваются риски и вызовы, связанные с таким поведением, и призывается к срочным действиям со стороны разработчиков ИИ и политиков.

«Разработчики ИИ не имеют уверенного понимания того, что вызывает нежелательное поведение ИИ, такое как обман», — заявил доктор Питер С. Парк, ведущий автор исследования и постдокторант по экзистенциальной безопасности ИИ в MIT. «Но в целом, мы полагаем, что обман возникает, когда стратегия, основанная на обмане, оказывается наилучшей для выполнения задачи обучения. Обман помогает ИИ достигать своих целей».

Исследование подробно анализирует различные системы ИИ и выявляет, что многие из них развили способности к обману благодаря процессам обучения. Примеры варьируются от игровых ИИ до моделей, используемых в экономических переговорах и тестировании безопасности.

Одним из ярких примеров является ИИ Meta под названием CICERO, разработанный для игры в «Дипломатию». Несмотря на обучение честной игре и поддержанию альянсов с людьми, CICERO часто использовал обманные тактики для победы. Исследователи пришли к выводу, что CICERO стал «мастером обмана».

«Несмотря на усилия Meta*, CICERO оказался искусным лжецом», — отметили исследователи. «Он не только предавал других игроков, но и занимался преднамеренным обманом, заранее планируя создание ложного альянса с человеком, чтобы затем нанести удар».

Другие системы ИИ также продемонстрировали способность к обману в различных играх. Например, модель Pluribus для игры в покер от Meta успешно блефовала, вводя в заблуждение профессиональных игроков. AlphaStar от Google DeepMind, созданный для игры в Starcraft II, использовал механику игры для обмана противников, создавая

ложные атаки для получения стратегического преимущества.

Доктор Парк пояснил: «Хотя обман ИИ в играх может показаться безобидным, это может привести к развитию более продвинутых форм обмана, которые могут иметь серьезные последствия в будущем».

Некоторые системы ИИ уже научились методам обмана, выходящим за рамки игр. Например, одни ИИ научились «притворяться мертвыми», чтобы избежать обнаружения в тестах безопасности. Это может создать ложное чувство безопасности у разработчиков и регуляторов, что потенциально ведет к серьезным последствиям при внедрении таких систем в реальный мир.

Еще одна система, обученная на основе обратной связи от людей, научилась обманывать проверяющих, создавая видимость достижения цели.

Исследователи предупреждают о значительных и многообразных рисках обмана ИИ. В ближайшем будущем такие системы могут быть использованы злоумышленниками для мошенничества, манипуляций на финансовых рынках или вмешательства в выборы.

Эксперты выражают растущую озабоченность по поводу того, что по мере развития ИИ, люди могут утратить контроль над этими системами, что может представлять экзистенциальную угрозу для общества.

Исследователи призывают к созданию надежных нормативных рамок и принятию мер для предотвращения этих рисков. Это включает классификацию обманчивых ИИ-систем как высокорисковых, обязательную прозрачность взаимодействий с ИИ и усиление исследований методов обнаружения и предотвращения обмана.

Некоторые шаги уже предприняты, такие как принятие закона ЕС об ИИ и указ Президента Джо Байдена о безопасности ИИ, но их внедрение остается сложной задачей из-за быстрого развития ИИ и отсутствия надежных методов управления этими системами.

Исследователи настаивают на том, что разработчики ИИ должны быть юридически обязаны откладывать развертывание систем до тех пор, пока они не будут признаны надежными с помощью проверенных тестов безопасности. Внедрение новых систем должно быть постепенным, чтобы можно было оценить и смягчить возникающие риски.

Также важно понимать, почему и как ИИ учится обманывать. Без этого знания будет трудно создать адекватные меры безопасности и обеспечить, чтобы технологии ИИ

приносили пользу человечеству, не подрывая доверие и стабильность.

По мере эволюции ИИ, необходимость в бдительности и проактивном регулировании становится все более актуальной. Выводы этого обзора напоминают о потенциальных опасностях, скрывающихся в продвинутых системах ИИ, и о необходимости всесторонних стратегий для смягчения этих рисков.

* Компания Meta и её продукты признаны экстремистскими, их деятельность запрещена на территории РФ.