Новое исследование раскрывает, насколько сложно отличить человека от ИИ в онлайн-общении.

В наше время взаимодействие с искусственным интеллектом (ИИ) онлайн происходит не только чаще, чем когда-либо, но и незаметнее для пользователей. Исследователи решили проверить, могут ли люди отличить ИИ от человека, проведя эксперимент с участием одного человека и трех различных ИИ-моделей.

«Тест Тьюринга», впервые предложенный как «игра в имитацию» компьютерным ученым Аланом Тьюрингом в 1950 году, оценивает способность машины демонстрировать интеллект, неотличимый от человеческого. Чтобы пройти этот тест, машина должна убедить собеседника в том, что она человек.

Ученые решили воспроизвести этот тест, предложив 500 участникам пообщаться с четырьмя респондентами: человеком, программой ELIZA 1960-х годов, а также моделями GPT-3.5 и GPT-4 (Generative Pre-trained Transformer 4) — это чётвертая версия модели глубокого обучения, разработанная компанией OpenAI. Основное преимущество GPT-4 по сравнению с предыдущими версиями заключается в его способности к более глубокому пониманию контекста и генерации более качественных и связных ответов. GPT-4 может обрабатывать и анализировать более сложные запросы, а также продолжать начатые тексты с сохранением смысла и стиля." data-html="true" data-original-title="GPT-4" >GPT-4, которые работают на базе ChatGPT. Каждое общение длилось пять минут, после чего участники должны были определить, говорили ли они с человеком или ИИ. Мы ранее рассказывали об исследовании , в котором ученые выяснили, что GPT-4 была признана человеческой в 54% случаев.

ELIZA, система с заранее запрограммированными ответами, но без больших языковых моделей (LLM) или нейронной архитектуры, была признана человеческой лишь в 22% случаев. GPT-3.5 набрала 50%, в то время как человек получил 67%.

Нелл Уотсон, исследователь ИИ в Институте инженеров электротехники и электроники (IEEE), отметила: «Машины могут создавать правдоподобные объяснения, как это делают люди. Они могут подвергаться когнитивным искажениям, быть сбитыми с толку и манипулируемыми, становясь все более обманчивыми. Все эти элементы делают ИИ системами, похожими на человека, что значительно отличает их от предыдущих подходов с ограниченным набором готовых ответов».

Исследование, которое основывается на десятилетиях попыток заставить ИИ пройти тест Тьюринга, подчеркивает распространенные опасения, что системы ИИ,

признанные человеческими, будут иметь «широкие социальные и экономические последствия». Ученые также отметили, что существует обоснованная критика упрощенности теста Тьюринга, утверждая, что «стилистические и социально-эмоциональные факторы играют большую роль в прохождении теста Тьюринга, чем традиционные представления об интеллекте». Это предполагает, что подход к поиску машинного интеллекта нуждается в пересмотре.

Уотсон добавила, что исследование представляет собой вызов для будущего взаимодействия человека и машины, и что люди будут становиться все более подозрительными к природе таких взаимодействий, особенно в чувствительных вопросах. Она подчеркнула, что исследование демонстрирует, как ИИ изменился в эпоху GPT.

«ELIZA была ограничена готовыми ответами, что значительно ограничивало ее возможности. Она могла бы обмануть кого-то на пять минут, но вскоре ее ограничения становились очевидными,» — сказала она. «Языковые модели невероятно гибки, способны синтезировать ответы на широкий круг тем, говорить на определенных языках или социолектах и изображать себя с характерными личностями и ценностями. Это огромный шаг вперед по сравнению с тем, что программируется вручную, независимо от того, насколько умело и тщательно это сделано.»

Современные языковые модели искусственного интеллекта, такие как GPT-4, демонстрируют поразительную способность имитировать человеческий интеллект и речь, что ставит под сомнение традиционные представления о машинном интеллекте. В эксперименте, воспроизводящем тест Тьюринга, GPT-4 была распознана как человек в 54% случаев, что значительно превосходит показатели более ранних систем ИИ. Это свидетельствует о существенном прогрессе в развитии ИИ и его способности генерировать правдоподобные, гибкие и контекстные ответы, сравнимые с человеческими.

Однако такая высокая степень человекоподобия ИИ также вызывает опасения относительно возможных социальных и экономических последствий, когда люди не могут отличить взаимодействие с ИИ от общения с человеком. Это требует пересмотра подходов к оценке машинного интеллекта и выработки новых критериев и методов различения человеческого и искусственного интеллекта. В будущем людям придется быть более осторожными и критичными при взаимодействии с ИИ, особенно в чувствительных вопросах, чтобы избежать манипуляций и неправильных суждений.

На перекрестке науки и фантазии — наш канал