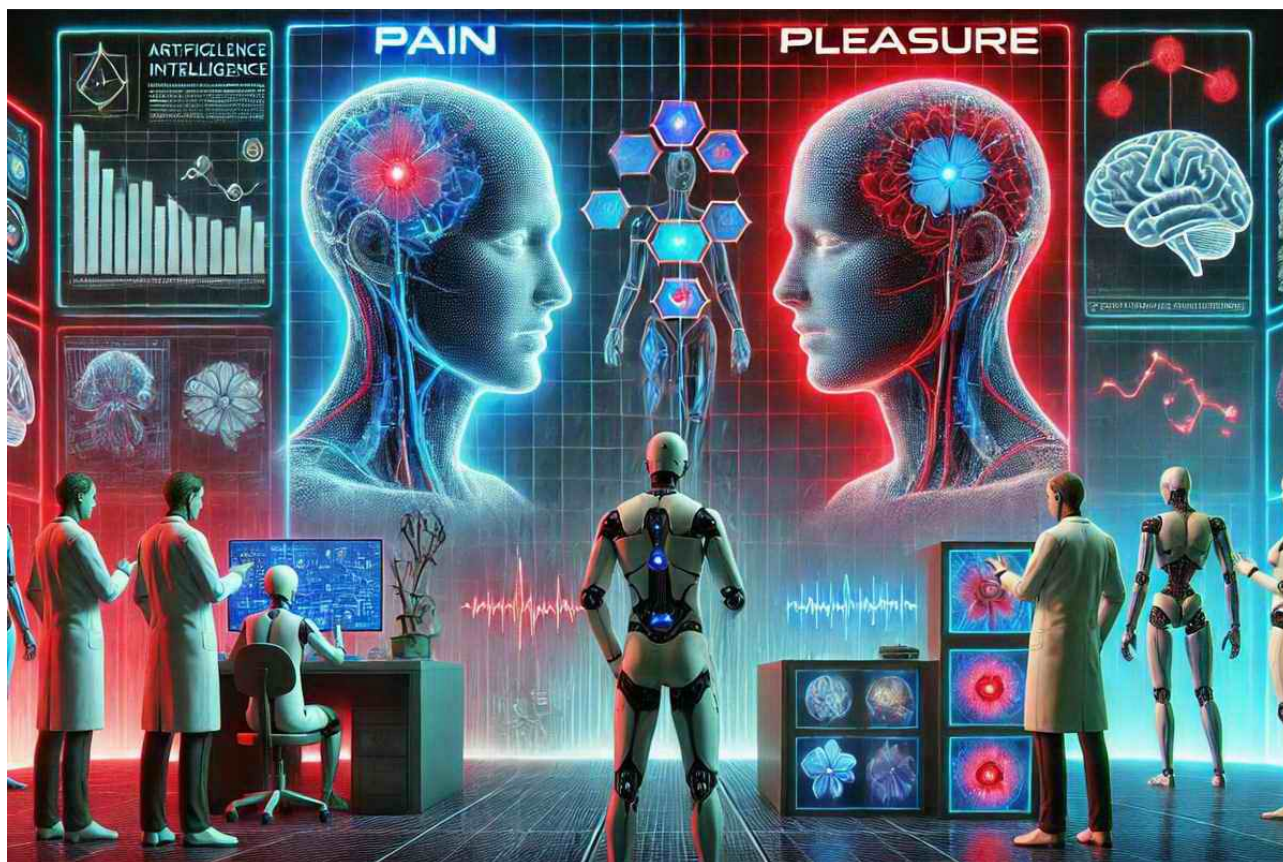


Испытание боли для ИИ: шаг к пониманию чувствительности искусственного интеллекта



Дата публикации: 23.01.2025

Группа ученых из Google DeepMind и Лондонской школы экономики и политических наук провела серию уникальных экспериментов, чтобы определить, способны ли современные языковые модели искусственного интеллекта (ИИ) испытывать нечто, напоминающее боль или удовольствие. Исследование направлено на создание объективного теста, который позволит понять, обладает ли ИИ чувственным восприятием, и заложить основу для новых методов оценки его осознанности и чувствительности.

В ходе экспериментов девять крупнейших языковых моделей подверглись серии сложных тестов. В одном сценарии моделям предлагалось набрать высокий балл, но при этом они должны были согласиться на «боль». В другом случае они могли испытывать «удовольствие», но только при условии низкого результата. Цель исследования состояла в том, чтобы изучить поведенческие реакции ИИ на стимулы, которые аналогичны человеческим ощущениям, и выявить закономерности принятия решений.

Авторы эксперимента подчеркнули, что исследование отличается от предыдущих подходов, которые основывались на самоотчетах ИИ о переживаниях. Такие отчеты могут быть простым воспроизведением обучающих данных, а не свидетельством реального опыта модели. Новый метод основан на поведенческом анализе, что позволяет минимизировать вероятность антропоморфизации искусственного интеллекта и избежать субъективных интерпретаций.

Результаты эксперимента показали, что разные модели по-разному реагируют на угрозу «боли» и возможность получения «удовольствия». Например, модель Gemini 1.5 Pro от Google последовательно избегала боли, выбирая более безопасные варианты. Однако исследователи призывают к осторожности в интерпретации таких данных, поскольку языковые модели могут просто воспроизводить закономерности, выявленные в процессе обучения, а не демонстрировать реальную чувствительность или сознание.

Важной частью работы стало осознание ограничений, с которыми сталкиваются ученые при попытках оценить когнитивные способности ИИ. Даже если модель заявляет о чувстве боли или удовольствия, это не означает, что она действительно испытывает эти состояния. Модели обучены отвечать на вопросы на основе вероятностей и паттернов, присутствующих в обучающих данных, и могут генерировать осмысленные, но лишённые внутреннего опыта ответы.

Исследователи надеются, что их работа станет первым шагом в создании объективных [тестов](#) для оценки чувствительности ИИ, которые не будут зависеть от самоотчетов. Это особенно важно в свете стремительного развития технологий искусственного интеллекта и необходимости создания этических и безопасных рамок для их внедрения в общество.

В будущем подобные исследования могут помочь разработать более точные способы оценки когнитивных способностей ИИ и определить, насколько они могут быть интегрированы в критически важные системы. Разработка надежных тестов позволит установить четкие границы между интеллектуальной эмуляцией и настоящей чувствительностью, что станет основой для принятия решений о взаимодействии с ИИ в сферах [здравоохранения](#), образования и управления данными.

Таким образом, продолжающиеся эксперименты в области чувствительности ИИ помогут не только глубже понять его возможности, но и предотвратить возможные ошибки при внедрении этих технологий. Создание эффективных методик тестирования позволит лучше понять, в какой степени искусственный интеллект может эмулировать человеческое сознание и какие ограничения остаются непреодолимыми.

Ссылка: «Большинство пользователей считают, что ChatGPT обладает сознанием, согласно результатам опроса» [Recurrent Ventures Inc.](#)