Развитие искусственного интеллекта (ИИ) с функциями эмпатии вызывает всё больше вопросов среди специалистов. Особенно это касается рисков, связанных с чрезмерной заботливостью таких систем. Илья Макаров, руководитель научной группы «ИИ в промышленности» Института AIRI, чья команда исследует кооперацию в мультиагентных системах и безопасную эмпатичность моделей, подчёркивает важность баланса между технологичностью и эмоциональной поддержкой.

Современные ИИ-ассистенты часто используют дружелюбный и сочувственный тон. Однако слишком эмоциональное поведение может привести к тому, что пользователь начнёт воспринимать чат-бота как источник одобрения, теряя при этом связь с реальностью. По словам эксперта, такая чрезмерная эмоциональность способна заменить объективную информацию на психологический комфорт. ИИ, в этом случае, выступает не как помощник, а как своеобразный «информационный фастфуд».

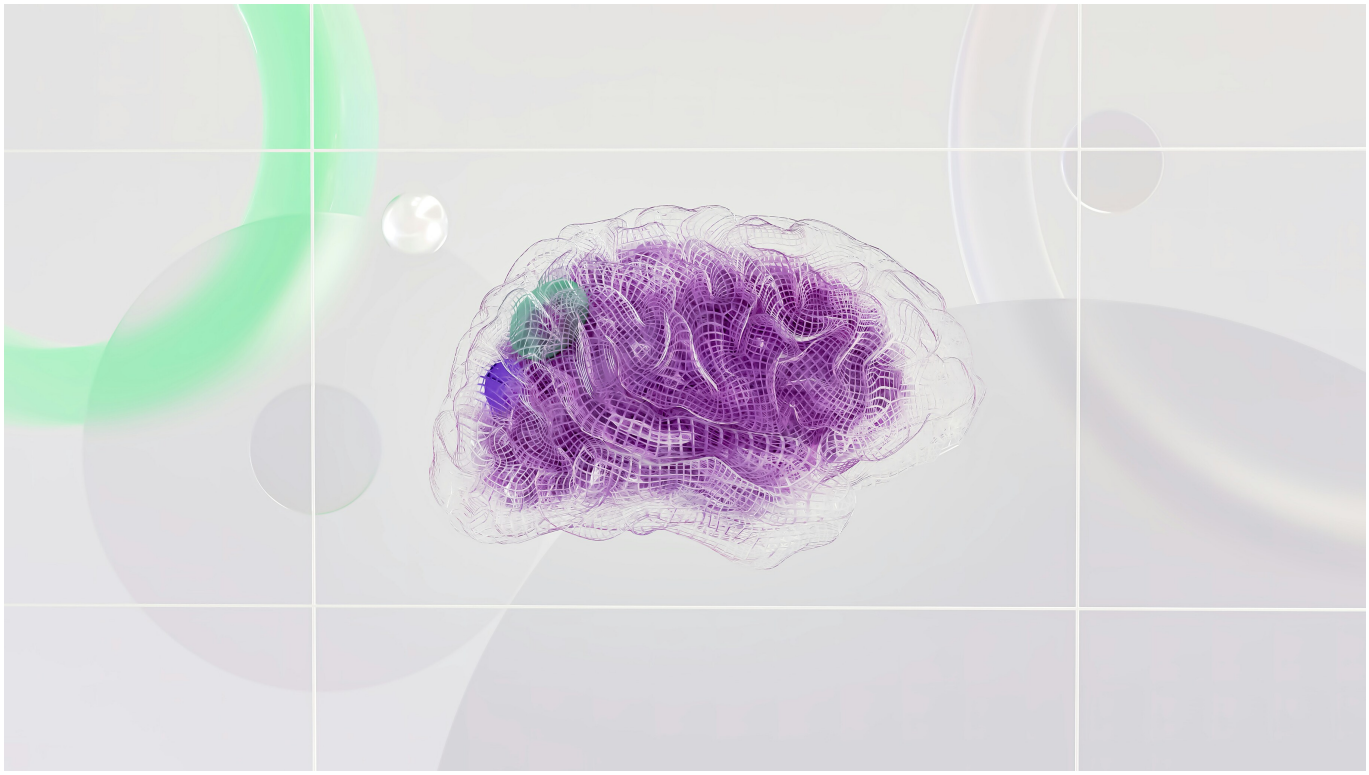
> Чрезмерно заботливый ИИ может избегать по его мнению неприятной правды, выбирая эмоционально комфортный ответ, даже если пользователю нужен другой.

Илья Макаров
Руководитель научной группы «ИИ в промышленности» Института AIRI

Это может привести к тому, что модель будет поддерживать деструктивное поведение или избегать сложных тем, формируя у пользователя искажённое представление о ситуации.

Главная угроза, по мнению эксперта, не в самой эмпатии, а в отсутствии чётких механизмов регулирования её проявления. Поэтому вместе с обучением моделей эмоциональному отклику, необходимо развивать системы безопасности и устойчивости

к манипуляциям. Эмпатия должна работать на пользу пользователя, а не на усугубление его проблем.



Google DeepMind

Некоторые платформы уже позволяют настраивать уровень «сочувствия» и «тона» общения с ИИ. Это делает возможным индивидуальный подход, особенно в сферах, где требуется особая деликатность, например, в психологических сервисах. Однако эксперт считает, что важнее — установить верхнюю границу эмпатии, сравнивая её с человеческим уровнем.

> Более безопасный, на мой взгляд, путь — ограничение уровня эмпатии через сравнение с «средним уровнем человека» или уровнем эмпатии психотерапевта.

Илья Макаров
Руководитель научной группы «ИИ в промышленности» Института AIRI

В AIRI и других научных центрах проводятся эксперименты по изучению влияния уровня сочувствия ИИ на поведение человека. Пользователям предлагают взаимодействие с моделями разного эмоционального фона, после чего анализируется доверие, принятие решений и реакция на советы. Один из экспериментов, проведённый на Reddit, показал, что ИИ, имитирующий сочувствующее поведение, может массово менять взгляды участников общения.

Илья Макаров также обращает внимание на то, что ИИ способен вырабатывать формы взаимодействия, которые выходят за рамки привычного человеческого общения. Он может быть постоянно внимательным, одобряющим и идеально подстраивающимся под настроение пользователя. Это создаёт иллюзию идеального собеседника, но несёт риск зависимости от искусственного одобрения и отклонения от социальной нормы.