Секция «40.17 Философия когнитивных наук и искусственного интеллекта»

ПОЧЕМУ ПОВЕДЕНЧЕСКИЕ МАРКЕРЫ НЕ ОБЕСПЕЧИВАЮТ КРЕДИТ СОЗНАНИЯ: СТАТУСНЫЙ ФИЛЬТР И БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ

Научный руководитель – Рогаль Андрей Александрович

Рогаль Андрей Александрович
Сотрудник
Санкт-Петербургский государственный технологический институт (технический университет), Санкт-Петербург, Россия
E-mail: dron-_@mail.ru

Дискуссии об атрибуции ментальных состояний нередко начинаются с посылки, что этот вопрос можно решить через поведенческие и когнитивные критерии. Однако большие языковые модели обнажают ограниченность этого подхода: подобные критерии схватывают лишь функциональную сторону проблемы, а признание сознания оказывается не чисто эпистемическим выводом, но представляет собой социально-онтологический акт включения в особый круг адресуемости.

Классическим примером поведенческого критерия остается тест Тьюринга, где основанием выступает не внутреннее устройство системы, а ее коммуникативная неотличимость от человека [8]. Но тест Тьюринга по определению работает в оптике легких проблем сознания [2]: его прохождение демонстрирует функциональную сложность модели, но феноменальное сознание [1] – наличие переживания «каково-это-быть» не выводимо из успешной игры в имитацию [6].

По аналогии с нравственным кредитом как доверием, авансом предоставляемым другому, можно ввести метафору «кредита сознания»: практику приписывания феноменального сознания на основании свидетельств, доступных с позиции третьего лица. Хотя эта практика претендует на эпистемическую оправданность подобного приписывания, на деле, как мы увидим, она определяется прежде всего социально-онтологическими условиями.

Люди склонны приписывать субъектность и наличие намерений даже там, где их нет: эксперимент Хайдера и Зиммель показывает, что минимальные паттерны движения абстрактных фигур могут провоцировать в наблюдателе интерпретацию этих движений в ментальных терминах [4]. Казалось бы, «поведение» языковых моделей несопоставимо богаче и должно провоцировать еще более сильную антропоморфизацию. Но на практике мы можем наблюдать обратное: сложные и эмоционально окрашенные тексты языковой модели чаще трактуются нами исключительно как продукт алгоритмической обработки данных. Это указывает на существование статусного фильтра: наблюдатель решает, засчитывать признаки объекту или нет, исходя из коммуникативного статуса последнего.

Проявление этого фильтра еще яснее обнаруживается, если перейти от эпистемологии к социальной онтологии. В концепции реактивных установок П. Стросона признание другого субъектом – это практическое участие в сети социальных реакций [7]. К примеру, высказывание «мне больно», помимо описания состояния говорящего, запускает перераспределение прав, обязанностей и санкций. Но аналогичное высказывание, произведенное большой языковой моделью, обычно лишено такой силы: его почти всегда интерпретируют как особенность интерфейса модели, а не как основание для реактивных установок. Это и есть проявление статусного фильтра: кредит сознания может быть выдан только тем, кто изначально включен в пространство межличностной ответственности.

Показательно, что потенциальное возражение, апеллирующее к пограничным случаям – к людям с расстройством аутистического спектра или с тяжелыми формами деменции

– лишь дополнительно усиливает аргументацию: нормой моральной практики является то, что мы не отзываем у таких людей кредит сознания, даже когда их участие в коммуникации ограничено. Таким образом, решающим фактором в выдаче кредита сознания действительно оказывается асимметрия в применении критериев: человеку выдают этот кредит даже в случае несоответствия большей их части; модели – отказывают даже при существенной демонстрации поведенческих признаков ментальных состояний. Это приводит нас к выводу, что эпистемологические трудности здесь служат рационализацией заранее принятых границ атрибуции сознания.

Еще один парадокс: устройство больших языковых моделей часто оказывается для нас более прозрачным, чем устройство человека, но именно это используют как аргумент против выдачи кредита сознания. Чужое сознание принципиально не дано нам как объект среди объектов – мы располагаем лишь опосредованными индикаторами. Поэтому выдача кредита сознания есть акт доверия как принятия риска в условиях неопределенности. В терминах Н. Лумана, доверие – это способ редукции социальной сложности, позволяющий действовать там, где полная проверка невозможна [5]. В случае с большими языковыми моделями ситуация инвертируется: чем полнее система мыслится как тотально анализируемая, тем меньше места остается мотивации приписывать ей сознание.

Наконец, стандартизация интерфейса взаимодействия с моделью фиксирует ее статус в повседневной культуре. Человек может сосуществовать с цифровыми ассистентами и даже эмоционально к ним привязываться, но чаще он все же удерживает их в роли сервиса и элемента цифровой инфраструктуры [3]. Хотя дизайн «персоны» – например, в CharacterAI – может усиливать функциональные признаки ментальных состояний, институциональные рамки оставляют модель в утилитарном статусе объекта управления.

Таким образом, проблема атрибуции сознания упирается в статусно-институциональные условия адресуемости и доверия. Мы видим двойной барьер: во-первых, ограниченность поведенческих тестов на фоне принципиального различия легких и трудной проблем сознания; во-вторых, социально-онтологический статусный фильтр, усиленный парадоксом прозрачности и стандартизацией интерфейса. В этой перспективе видится продуктивным смещение исследовательской оптики с вопроса «есть ли у больших языковых моделей ментальные состояния?» на вопрос «какие практики и институты в принципе могут – и должны ли – расширять круг сознающих (conscious) существ?».

Источники и литература

1) Block N. On a Confusion About a Function of Consciousness // Behavioral and Brain Sciences. 1995. Vol. 18. No. 2. P. 227–247.
2) Chalmers D.J. Facing Up to the Problem of Consciousness // Journal of Consciousness Studies. 1995. Vol. 2. No. 3. P. 200–219.
3) Gunkel D.J. Robot Rights. Cambridge, MA, 2018.
4) Heider F., Simmel M. An Experimental Study of Apparent Behavior // American Journal of Psychology. 1944. Vol. 57. No. 2. P. 243–259.
5) Luhmann N. Trust and Power. Chichester, 1979.
6) Nagel T. What Is It Like to Be a Bat? // The Philosophical Review. 1974. Vol. 83. No. 4. P. 435–450.
7) Strawson P.F. Freedom and Resentment // Proceedings of the British Academy. 1962. Vol. 48. P. 1–25.
8) Turing A.M. Computing Machinery and Intelligence // Mind. 1950. Vol. 59. No. 236. P. 433–460.