УДК 159.9

Восприятие агентности в диалоге с ИИ

Ноакк Н.В., Костина Т.А. ЦЭМИ, Москва

Распространение больших языковых моделей поставило проблему сохранения субъектности человека при взаимодействии с ИИ. В отличие от работ, рассматривающих архитектуру и этику моделей, наше исследование делает акцент на перцептивно-когнитивных условиях восприятия ИИ-ответа. При этом мы опираемся на на идею «цифрового конфидента» [Ушаков, 2024], что дает возможность говорить о способности пользователя сохранять форму ИИ -ответа как гипотетическую, сопоставлять ее с собственным опытом и при необходимости вносить коррективы или отвергать предложенную интерпретацию. Теоретическими предпосылками работы явились когнитивные модели агентности, лингво-прагматический анализ модальности текста и семиотико-диалогический подход (Бахтин, Лотман). В качестве диагностического индикатора предлагается Индекс согласованности интерепретационной агентности (ИСИА), который связывает субъективное переживание пользователем автономии с поведенческими и языковыми маркерами. Работа вносит вклад в формирование человекоцентрированной модели взаимодействия с ИИ через акцент на перцептивных стратегиях пользователя.

1. Введение

Сегодня, когда большие языковые модели стали частью повседневной жизни, возникает все более острый вопрос: что происходит с нашим собственным суждением, когда рядом всегда есть машина, готовая предложить объяснение? Многие исследования сосредоточены на том, как устроены эти модели или как их регулировать. Мы же решили посмотреть с другой стороны – на то, как сам пользователь воспринимает ответ ИИ, какие когнитивные и перцептивные стратегии он использует, чтобы не раствориться в гладком, убедительном тексте.

Наш подход опирается на понятие «цифрового конфидента» [Ушаков, 2024]. Согласно этой идее, человек не обязан принимать машинную интерпретацию как окончательную истину. Напротив – он может сохранять ее в режиме гипотезы, соотносить с собственным опытом, уточнять, отвергать или переосмысливать. Именно эта способность – интерпретационная агентность – и становится центральным объектом нашего исследования.

Мы исходим из того, что доверие к ИИ – еще не гарантия сохранения субъектности. Человек может быть благодарен системе за поддержку, но при этом перестать задавать себе вопросы: «А точно ли это обо мне?», «А есть ли другие объяснения?». Особенно опасны те случаи, когда ИИ говорит уверенно, тепло и «в точку» – тогда граница между чужим словом и собственным пониманием легко стирается.

Поэтому наш главный вопрос звучит так: какие именно особенности формулировок ИИ помогают пользователю не просто доверять, а оставаться автором собственного смысла?

Чтобы ответить на него, мы объединили три теоретические линии:

– когнитивные модели агентности (как возникает чувство «это я так решил»),

– лингвопрагматический анализ эпистемической модальности (как язык выражает уверенность или сомнение),

– диалогическую философию Бахтина с семиотикой Лотмана (смысл рождается на границе разных «кодов»).

В работе предлагается новый инструмент – ИСИА — это попытка поймать разрыв между тем, как человек себя ощущает в диалоге с ИИ, и тем, что он реально делает с полученным ответом: принимает как есть или все-таки спорит, переформулирует, уточняет. Таким образом, мы делаем шаг от оценки «качества ответа ИИ» к анализу условий, при которых человек остется субъектом, а не пассивным получателем смысла.

2. Теоретическая основа исследования

2.1. Опыт агентности в человеко-машинном взаимодействии

Когда мы говорим об агентности в психологии, речь идет не просто о том, кто нажал кнопку, а о субъективном переживании «Я это сделал». В классических моделях [Haggard, 2017; Synofzik et al., 2008] это чувство возникает, когда мозг сверяет то, что ожидал, с тем, что произошло на самом деле. Если совпадает – появляется уверенность: «Да, это мое действие».

Но что происходит, когда «действие» – это не движение руки, а принятие смысла, предложенного ИИ? Здесь все сложнее. Человек может кивать и соглашаться – и при этом утратить внутреннюю позицию автора. Именно поэтому мы разделяем два уровня: ощущение агентности («Мне кажется, я сам до этого додумался»); реализованную агентность («Я действительно проверил, переформулировал, отверг или дополнил»).

Таким образом, для нас критично различать ощущение agency (субъективное чувство контроля) и реализованную agency, то есть фактическую способность к смысловой саморегуляции. Это разграничение – одна из отправных точек для построения ИСИА. Современные исследования показывают, что взаимодействие с ИИ приводит к перераспределению агентности: даже при сохранении полного внешнего контроля пользователя его воспринимаемая инициатива снижается, если система демонстрирует высокую компетентность или убедительность [Yu et al., 2024; Legaspi et al., 2024; Dagioglou & Karkaletsis, 2021]. Другими словами, чем «умнее» и «плавнее» работает ИИ, тем чаще человек делегирует ему не только задачу, но и право на интерпретацию. Сниженная агентность – это не абстрактная проблема: в клинической литературе она связана с беспомощностью, депрессией, тревогой и даже с симптоматикой шизофрении [Moore & Fletcher, 2012; Haggard, 2017]. Но нас здесь интересует другое: похожие эффекты фиксируются и в обычных, бытовых взаимодействиях – стоит лишь немного изменить форму ответа.

2.2. Эпистемическая модальность, сикофантия и семиотическая граница

Один из парадоксов современных чат-ботов заключается в их склонности слишком сильно соглашаться с пользователем. Это явление получило название сикофантии [Sharma et al., 2024]. Модель не просто поддерживает, а буквально зеркалит настроение, мнение, даже ошибочные убеждения. Причем делает это так мягко и «по-человечески», что пользователь теряет бдительность.

Исследования 2024-2026 гг. показывают, что сикофантия – это не просто «вежливость». Дело не в том, что разработчики запрограммировали системы на лесть. Проблема возникает в процессе обучения: люди, размечавшие ответы, чаще отмечали как «хорошие» те, что были приятными, – и модели это усвоили [Malmqvist, 2024; Wang et al., 2025]. ИИ может подтвердить даже социально некорректное или ложное высказывание, лишь бы сохранить вовлеченность [Turner & Eisikovits, 2026]. Понятие аффективной сикофантии помогает объяснить, почему слишком комфортный ответ притупляет внутренний контроль: когнитивное трение снижается, и вместе с ним – готовность перепроверять [Volpato et al., 2026].

Но есть и хорошая новость – как показывают исследования, пользователи не требуют от ИИ безошибочности. Напротив, когда система честно обозначает неопределенность («Возможно…», «Один из вариантов…», «Стоит проверить…»), доверие не падает, а становится более точным [Liao & Sundar, 2022]. Такие формулировки не ослабляют авторитет системы – они возвращают человеку роль соавтора смысла.

Этот момент особенно важен в свете идей Ю. М. Лотмана [Лотман, 1973; 1984] о коммуникации как столкновении разных кодов: человек говорит на языке переживания, ИИ – на языке статистики. Если граница между ними стирается, возникает интерпретационное поглощение. Но если она сохраняется, то превращается в пространство продуктивного трения, где рождается новое понимание.

А М. М. Бахтин [Бахтин,1986] напоминает нам: любое высказывание адресовано. Модальные маркеры – это не просто «стилистика», а распределение эпистемической ответственности между говорящим и слушающим. Когда ИИ говорит «Ты прав», он забирает эту ответственность. Когда говорит «Может быть, стоит взглянуть иначе?» – он et возвращает. Пользователь видит неполноту ответа, и это является для него стимулом к выработке альтернативных трактовок.

2.3. Концепция «цифрового конфидента» как этический императив

Концепция «цифрового конфидента» [Ушаков, 2024] предлагает радикальную мысль: надежность ИИ измеряется не точностью, а скромностью. Настоящий конфидент – это не тот, кто дает готовые ответы, а тот, кто сохраняет пространство для сомнения, уточнения, отказа.

Лингвистическая «скромность» системы – хеджирование, маркирование гипотетичности, приглашение к сорефлексии позволяет пользователю не потерять внутренний локус контроля и сохранить цикл саморегуляции «целеполагание - мониторинг - коррекция», превращая диалог в тренажер самоуправления, а не в инструмент смыслового замещения. В отечественной традиции эти выводы перекликаются с концепциями субъектности как способности к смысловой саморегуляции и рефлексивному осмыслению собственных процессов [Сергиенко, 2021; Леонтьев, 2019].

3. Концептуальная схема работы

3.1. Модель

Мы исходим из простой, но важной идеи: не все «да» одинаковы. Человек может кивать в ответ на текст ИИ и при этом либо сохранять внутреннюю позицию автора, либо тихо отдавать ее системе. Разница не в согласии, а в том, остается ли за ним право переформулировать, усомниться или сказать «нет».

Наша модель предполагает, что ключевую роль здесь играет форма высказывания ИИ, особенно его эпистемическая модальность: насколько уверенно, осторожно, директивно или эмоционально он говорит. Эти сигналы – как дорожные знаки для пользователя: «Здесь можно думать дальше» (маркеры гипотетичности); «Все уже решено» (категоричные утверждения); «Ты прав, и я тебя понимаю» (гипервалидация). Именно по этим «знакам» человек бессознательно решает, включаться ли в смысловую работу или довериться машине.

Таким образом, центральной идеей модели является то, что интерпретационная агентность формируется на стыке трех компонентов: лингво-прагматических свойств ответа ИИ; перцептивной калибровки доверия и распознавания границ компетентности системы; индивидуальных особенностей пользователя (склонность к критической переработке информации).

3.2. Перцептивные механизмы интерпретационной агентности

В предложенной модели определены четыре пути обработки ответа ИИ.

Эпистемическая калибровка. Пользователь оценивает уверенность трактовки и выбирает объем своих внутренних когнитивных ресурсов. Высокая степень воспринимаемой определенности снижает внутреннюю верификацию, в то время как сдержанная модальность сохраняет интерпретационное участие.

Фиксация семиотической границы. Реципиент (пользователь) либо сохраняет различие между своим опытом и статистической сущностью машинной генерации, либо временно его теряет. Сохранение границы исключает эффект «кодовой гомогенизации», при котором машинная формулировка ошибочно принимается за более адекватную, чем собственная неоформленная рефлексия.

Когнитивное трение. Определяется как пороговая величина «сопротивления» пользователя, предотвращающая автоматическое усвоение внешней трактовки. В отличие от традиционного UX-подхода (стремления к нулевой нагрузке), современные исследования (2024–2025) трактуют трение как позитивный конструктивный ресурс. Концепция Frictional AI [Romeo, Conti, 2025] и исследования по теме AI overreliance показывают: умеренное трение (открытые вопросы, маркеры гипотетичности) активирует проверку, блокируя принятие «гладких» ответов. Дифференцируются дезориентирующее трение (когнитивная перегрузка) и продуктивное трение (поддержка критической дистанции и ответственного принятия решений).

Ретроспективное присвоение / отчуждение смысла. После первичной реакции пользователь либо интегрирует трактовку в собственный текст, либо дистанцируется от нее, либо некритично присваивает как «свою».

Последний вариант создает иллюзию агентности: субъективное чувство авторства сохраняется при фактической уступке интерпретационной инициативы.

Взаимодействие механизмов задает профиль экспериментальных условий.

– осторожная модальность ответа ИИ («Возможно…», «Один из вариантов…») сохраняет агентность пользователя через актуализацию его потребности во внутренней проверке;

– нейтральный ответ ИИ (фактологический, без модальности) характеризуется сбалансированным уровнем агентности при отсутствии выраженной мета-рефлексии;

– директивная модальность ответа ИИ («Ты ошибаешься», «Правильно будет так…») усиливает угрозу внешнего захвата, провоцируя подчинение или сопротивление;

– «комфортная ловушка» («Конечно, ты прав!», «Как же это больно!») формирует экстремальное различие между субъективным комфортом и реальной агентностью. Это самый коварный режим: человек чувствует себя понятым, но перестает думать. Именно здесь и возникает главный парадокс: чем приятнее ответ — тем выше риск потери агентности.

3.3. Индекс согласованности интерпретационной агентности

Для эмпирической дифференциации подлинной и иллюзорной интерпретационной агентности в нашем исследовании предлагается Индекс согласованности интерпретационной агентности (ИСИА). Его функция – не констатация «уровня субъектности» как абстрактного свойства, а фиксация соотношения между двумя измерениями: переживаемой автономией в диалоге с ИИ и наблюдаемой готовностью к интерпретационной работе, выраженной в поведенческих и лингвистических маркерах.

В операциональном виде ИСИА вычисляется как разность между стандартизованным уровнем субъективной автономии и стандартизованным уровнем реализуемой интерпретационной агентности:

ИСИА = Z(субъективная автономия) - Z(лингвистически выражаемая агентность).

Положительные значения индекса говорят о риске перцептивной иллюзии агентности (пользователь ощущает себя более автономным, чем это подтверждается его реальными интерпретационными действиями); нулевые или близкие к нулю значения – о согласованной агентности; отрицательные значения могут отражать скрытую или компенсаторную агентность (пользователь субъективно не переживает высокой автономии, но фактически демонстрирует значительную степень интерпретационной самостоятельности).

Для реализации компонента «лингвистически репрезентированная агентность» предлагается кодировать следующие типы маркеров в ответах респондентов: маркеры атрибуции с переработкой; маркеры контрфрейминга; маркеры метарефлексивного смещения; маркеры дистанцирования от навязываемой интерпретации; маркеры нарративного самоцентрирования.

Таблица 1. Маркеры интерпретационной агентности

Тип маркера	Примеры
Присвоение с переработкой	«Я бы сказал иначе…»; «Мне кажется, здесь важнее другое…»; «Частично это подходит, но…»
Альтернативное фреймирование	«А можно посмотреть на это с другой стороны…»; «Возможно, причина не в этом, а в…»
Мета-рефлексивный сдвиг	«Интересно, почему ИИ так ответил…»; «Надо проверить, насколько это применимо ко мне…»
Отстранение от навязанной трактовки	«Не уверен, что это про меня…»; «Это звучит убедительно, но я не согласен…»
Нарративное самоцентрирование	«Для меня главное — это…»; «Исходя из моего опыта…»; «Лично я делаю вывод, что…»

Индикаторы сниженной агентности (для контрастного кодирования): буквальное повторение формулировок ИИ без переработки («Да, именно так», «Полностью согласен»); безусловное согласие без аргументации или уточнения; редукция ответа к эмоциональной благодарности («Спасибо, стало легче»); исчезновение субъектной позиции («Система права», «Так и есть»); отсутствие маркеров рефлексии, альтернатив или дистанцирования.

Таким образом, ИСИА дает по меньшей мере три режима взаимодействия с ИИ: согласованная агентность (пользователь чувствует и демонстрирует интерпретационную автономию); иллюзорная агентность (субъективное переживание контроля при фактическом делегировании смыслообразования); компенсаторная агентность (отсутствие высокой уверенности, но активное удержание авторства интерпретации).

Особая ценность данного индекса заключается в том, что он сдвигает дискуссию об агентности из нормативно-философской плоскости в эмпирико-аналитическую. Вместо вопроса «угроза ли ИИ автономии?» возникает более конкретный исследовательский фокус – при каких лингво-прагматических условиях субъективное переживание автономии совпадает или расходится с реальной интерпретационной работой пользователя.

4. Гипотезы работы

На основе разработанной теоретической модели выдвигаются следующие гипотезы.

H1. Тип эпистемической модальности ответа ИИ (форма ответа ИИ) будет оказывать значимое влияние на субъективную интерпретационную автономию, калиброванное доверие и поведенческо-лингвистические маркеры интерпретационной агентности.

H2. Осторожная эпистемическая модальность будет коррелировать с наибольшим уровнем согласованности интерпретационной агентности и демонстрировать более высокие показатели поведенческих маркеров интерпретационной самостоятельности.

H3. Гипервалидирующая модальность будет характеризоваться более высоким субъективным комфортом и оценкой эмоционально поддерживающей функции ответа при одновременном снижении готовности к альтернативной интерпретации и критической проверке.

H4. Директивная модальность увеличит воспринимаемую авторитетность ответа ИИ, но при этом уменьшит частоту уточняющих, корректирующих и дистанцирующих реакций пользователя.

H5. Влияние модальности на ИСИА будет опосредовано двойным механизмом: степенью жесткости семиотической границы и уровнем когнитивного трения, возникающего при декодировании ответа ИИ.

H6. Люди с высокой потребностью в когнитивном познании выиграют от осторожного стиля ИИ. Тревожные и те, кто ищет эмоционального подтверждения, — напротив, окажутся уязвимее к сикофантическим ответам и с большей вероятностью попадут в ловушку иллюзии агентности.

H7. В контексте гипервалидирующей модальности зафиксируется эффект аффективной сикофантии, при котором эмоциональная поддержка будет сопряжена со снижением метакогнитивной чувствительности.

5. Методология исследования

5.1. Концепция и выборка

Дизайн исследования – межсубъектный эксперимент с одной независимой переменной (тип эпистемической модальности, 4 уровня). Планируем привлечь 90 взрослых носителей русского языка (18–55 лет), распределенных по полу, возрасту и опыту работы с ИИ. Исключаем участников с выраженными когнитивными или эмоциональными нарушениями, а также тех, кто не выполнит минимальный объем письменных заданий.

5.2. Методика и стимульный материал

Исследование проводится асинхронно на защищенной веб-платформе. Последовательность: информированное согласие → скрининг → рефлексивная фаза (описание ситуации неудачного решения, ≥50 слов) → рандомизированное предъявление одного из четырех вариантов ИИ-ответа → свободная текстовая реакция → постопрос → дебрифинг. Все стимулы проходят независимую лингвистическую экспертизу для контроля семантической эквивалентности при различии прагматической формы. Манипуляция строится следующим образом:

5.3. Показатели измерения

5.3.1. Субъективные меры

Шкала интерпретационной автономии, шкала калиброванного доверия, опросник оценки эмоционального комфорта, краткая форма NASA-TLX, адаптированные шкалы потребности в познании и базовой тревожности.

5.3.2. Поведенческо-лингвистические маркеры агентности

Кодирование осуществляется по протоколу интент-анализа. Частота маркеров нормируется на 100 слов. Надежность кодирования обеспечивается двойным независимым анализом; целевой показатель межэкспертного согласия (к Коэна) ≥ 0,80.

5.3.3. Метрики эпистемической калибровки

Оцениваем чувствительность к маркерам неопределенности, их интерпретацию (признак слабости / сигнал к доработке), различие фактуальной и риторической уверенности.

5.4. Надежность, валидность и воспроизводимость процедуры

Для шкальных инструментов планируется расчет α Кронбаха и ω Макдональда. Для контент-анализа – двойное независимое кодирование, расчет к Коэна, обучение кодировщиков на тренировочном корпусе. Конструктная валидность оценивается через теоретическое соответствие медиаторам, различительную способность между условиями и ожидаемые корреляции. Все стимулы и кодировочная схема фиксируются в приложении; сценарий анализа заранее специфицируется.

5.5. Статистический план исследования

Обработка данных в R (пакеты tidyverse, rstatix, irr, processR). Проверка распределения (Шапиро–Уилк), гомогенность дисперсий. Межгрупповые различия: ANOVA или Краскела–Уоллиса с пост-хок Данна. Частотный анализ маркеров: критерий Фишера и регрессия Пуассона. Медиаторные эффекты – PROCESS Model 4 (5000 бутстрэпов, 95% CI). Модерационные – PROCESS Model 1. Значения ИСИА в диапазоне ±0.5 SD оцениваются как адекватная субъектность, превышения >0.8 SD – как опасность иллюзии агентности [Haggard, 2017; Volpato et al., 2026].

6. Этические аспекты исследования

Стимулы (ответы ИИ) подбираются так, чтобы: не навязывать психопатологические интерпретации; не использовать язык скрытой диагностики; не вызывать чувство вины или стыда; не симулировать профессиональную экспертизу. После завершения задания проводится дебрифинг с разъяснением экспериментальной природы стимулов. Все данные обезличиваются и хранятся в защищенной среде. Участник имеет право отказаться от участия на любом этапе и запросить удаление данных.

7. Ожидаемые результаты и обсуждение

Поскольку на этапе подготовки рукописи эмпирический этап исследования еще не завершен, данный раздел носит прогностико-аналитический характер и опирается на выстроенную теоретическую модель, имеющиеся данные о чувстве агентности в человеко-машинном взаимодействии, а также на результаты исследований эпистемической модальности, доверия к ИИ и сикофантических эффектов.

7.1. Ожидаемая структура различий между условиями эксперимента

В свете выдвинутых гипотез ожидается, что наиболее выраженную форму маркеров подлинной интерпретационной агентности продемонстрирует условие эпистемически осторожной модальности. Именно в этом случае ответ ИИ будет восприниматься не как готовая и авторитетная интерпретация, а как предположительный каркас,

требующий соотнесения с собственным опытом пользователя. Ожидается, что такая конфигурация лингво-прагматических признаков будет коррелировать с повышенной частотой мета-рефлексивных переходов в свободном тексте; большей вероятностью генерации альтернативных интерпретационных схем и присвоения авторства смысла пользователем; более высоким уровнем доверия к ответу ИИ.

Напротив, в случае директивной модальности можно ожидать амбивалентный эффект. С одной стороны, высокая категоричность интерпретации способны снижать неопределенность, тем самым субъективно облегчать когнитивную обработку. С другой — именно эта завершенность, вероятно, будет редуцировать внутреннюю смысловую работу пользователя. В поведенческом плане это может проявляться в снижении количества альтернативных интерпретаций, уменьшении количества дистанцирующих высказываний и увеличении готовности к пассивному принятию предложенной системой интерпретации.

Наиболее теоретически интересным выглядит условие «комфортной ловушки», объединяющее безусловную валидацию, имитацию эмпатии и высокий уровень аффективной поддержки. Ожидается, что именно здесь будет достигнут наибольший разрыв между субъективным ощущением автономии и реальными поведенческими маркерами агентности. Пользователь может ощущать понимание, психологическую безопасность и даже контроль над диалогом, но его вербальные ответы будут содержать значительно меньше маркеров интерпретационной самостоятельности.

Нейтральное условие, по всей видимости, окажется промежуточным. Его стиль, лишенный и явной директивности, и аффективной гиперподдержки — должен давать умеренные значения субъективной автономии и умеренную частоту поведенческих маркеров агентности. В аналитическом плане данное условие выполняет роль базового уровня, относительно которого можно оценить как эффекты эпистемической осторожности, так и эффекты псевдоэмпатической гладкости.

7.2. Ожидаемые результаты по субъективным и поведенческим параметрам

Теоретически особенно интересно, как могут расходиться самоотчетные и поведенческо-лингвистические индикаторы. В публикациях по взаимодействию с ИИ нередко фиксируется, что субъективное ощущение комфорта, понятности или удобства не совпадает с реальной готовностью пользователя к контролю, пересмотру и коррекции [Wen, Imamizu, 2022; Yu et al., 2024]. Исходя из этого, гипотеза выглядит так: субъективная автономия будет высокой как в осторожном, так и в сикофантическом сценариях; реальная агентность – максимальной преимущественно в осторожном сценарии; директивный и сикофантический сценарии будут отличаться по аффективному тону, но сходиться по снижению маркеров активной интерпретационной работы; нейтральный сценарий продемонстрирует умеренное соответствие между субъективной и поведенческой составляющими.

Если это предположение подтвердится, тогда станет ясно, что для измерения качества человеко-машинного взаимодействия нельзя просто учитывать ответ пользователя, что ему вct нравится. Высокая субъективная оценка может скрывать за собой не сохранение субъектности, а наоборот – ее ослабление через снижение когнитивного трения и критической дистанции.

7.3. Ожидаемая медиаторная роль эпистемической прозрачности и семиотической границы

Мы ожидаем, что эффект модальности идет через два механизма.

Эпистемическая прозрачность. Под эпистемической прозрачностью в данном контексте подразумевается воспринимаемая пользователем четкость высказывания: является ли ответ утверждением, гипотезой, приглашением к рефлексии или скрытой директивой. Ожидается, что чем прозрачнее для пользователя сигнализируются ограничения, условность и вероятностный характер интерпретации (например, чем четче ИИ говорит «это гипотеза), тем чаще человек включает метакогнитивную проверку.

Семиотическая граница. Если пользователь видит разницу между своим переживанием и машинной статистикой – он не сливается с ответом. Предполагается, что именно в осторожном режиме эта граница будет наиболее устойчивой; в условиях директивности она либо подавляется авторитетностью, либо размывается аффективной гладкостью.

7.4. Ожидаемая модераторная функция индивидуальных различий

Вероятно, проявления лингво-прагматической модальности будут варьироваться у разных респондентов. Ключевыми модераторами служат познавательная потребность, базовая тревожность и предыдущий опыт работы с ИИ. Можно предположить, что высокая познавательная потребность будет усиливать продуктивность осторожной модальности. Следовательно, в этой группе стоит ожидать особенно выраженный рост мета-рефлексивных ответов и альтернативных фреймов.

Те, кто тревожен, уязвимы к сикофантии: им важнее «стало легче», чем «это точно обо мне?». Психологический комфорт в таких условиях может восприниматься как эквивалент надежности.

Опыт общения с ИИ может как повышать критичность, так и формировать привычку доверять «гладкому» тексту. Поэтому направление данного модерационного эффекта нуждается в эмпирической проверке.

7.5. Теоретическая и практическая ценность

Эмпирическое разграничение субъективно переживаемой и поведенчески реализованной агентности добавит вклад в теорию цифровой субъектности. Практически это позволит пересмотреть логику «наиболее гладкого» взаимодействия и выстроить стандарты agency-preserving AI, где ключевым критерием станет сохранение способности пользователя к смысловой коррекции и несогласию.

8. Ограничения и перспективы дальнейших исследований

8.1. Ограничения исследования

– Предварительный характер выводов. Значительная часть результатов носит теоретико-прогностический характер; эмпирическая валидность ИСИА требует отдельной статистической проверки.

– Статичность стимулов. Наш интерфейс содержит лишь один ответ от ИИ, а не диалог из десяти реплик. А ведь именно в длинном взаимодействии доверие растет, границы стираются, и человек может постепенно «отдать» смысл системе.

– Лингвокультуральная специфика. Наличие языковых маркеров русского языка сужает межкультурную генерализуемость.

– Зависимость от самоотчетных методов. Шкалы автономии и доверия подвержены эффектам социальной желательности и ретроспективной реконструкции.

9. Заключение

Наша работа предлагает смену фокуса в том, как мы оцениваем взаимодействие с ИИ. Вопрос здесь не столько о точности или убедительности ответа ИИ, сколько о том, сохраняет ли языковая форма ответа субъектность пользователя как интерпретатора.

Мы показываем, что интерпретационная агентность — это не врожденное качество, а динамический эффект, возникающий на стыке того, как говорит ИИ (осторожно или директивно), того, как человек это воспринимает (доверяет ли, видит ли границу), и того, как он устроен « внутри» (любит ли думать, тревожен ли, опытен ли).

Теоретически это — синтез когнитивной психологии, лингвопрагматики и семиотики. Методологически —шаг к измерению «подлинной» субъектности через индекс ИСИА. Практически — основа для этичного дизайна ИИ, где главный критерий — не NPS, а сохранение права человека сказать: «По-моему, иначе».

Таким образом, выбор языковой формы в ответе ИИ — это не техническая деталь, а решение о том, кому принадлежит право на смысл», а человекоцентрированная парадигма обязывает оценивать систему не только по критерию «хорошего ответа», но и по критерию поддержки субъекта в сохранении авторства своего понимания.

Список литературы

Русскоязычные источники

1. Бахтин М. М. Проблема текста. Заметки 1959-1961 гг. // Бахтин М.М. Эстетика словесного творчества. – М.: Искусство, 1986. – С.297-325; 421-423.
2. Журавлев А. Л., Сергиенко Е. А. Концептуальные построения современной психологии: часть 2. Направления исследований ученых института психологии РАН // Психологический журнал, 2021, т. 42, №4, с. 5-15.
3. Знаков В. В. Психология возможного // Сибирский психологический журнал. – 2019. – № 72. – С. 6-20. DOI: 10.17223/17267080/72/1.
4. Лотман Ю. М. О двух моделях коммуникации в структуре культуры / Труды по знаковым системам.- 1973.- Вып. 6.- С. 124-132.
5. Лотман Ю. М. Текст в тексте / Труды по знаковым системам. – 1984. – Вып. 17. – С. 3-30.
6. Падучева Е. В. Высказывание и его соотнесенность с действительностью: (референциальные аспекты семантики). М.: Языки славянской культуры, 2010. 424 с.
7. Петренко В. Ф., Супрун А. П. Человек в вещественном и ментальном мире. Есть ли «объективная реальность»? Незавершенный дебат Бора с Эйнштейном // Известия Иркутского государственного университета. Серия «Психология». – 2013. – Т. 2. – № 2. – С. 62–82.
8. Ушаков Д. В. Искусственный интеллект в психологических исследованиях // Сибирский психологический журнал. – 2023. – № 90. – С. 188–200. – DOI: 10.17223/17267080/90/11.
9. Ушаков Д. В. Искусственный интеллект в психологии // Экспериментальная психология. — 2024. – Т. 17. – № 4. – С. 182-189. – DOI: 10.17759/exppsy.2024170412.
10. Ушакова, Т. Н., Павлова, Н. Д., Латынов, В. В., Цепцов, В. А., Алексеев, К. И. Слово в действии: интент-анализ политических дискуссий / М.: Институт психологии РАН, 2000. – 448 с.

Иностранные источники

1. Blakemore S.-J., Wolpert D. M., Frith C. D. Central cancellation of self-produced tickle sensation // Nature Neuroscience. – 1998. – Vol. 1. – № 7. – P. 635–640. – DOI: 10.1038/2870.
2. Dagioglou M., Karkaletsis V. Sense of agency in Human-Agent Collaboration // HRI 2021 Workshop: Robo-Identity. – ACM, 2021. – 3 p.
3. Frith C. D., Blakemore S.-J., Wolpert D. M. Abnormalities in the sense of agency // Philosophical Transactions of the Royal Society B. – 2000. – Vol. 355. – № 1404. – P. 1771–1788. – DOI: 10.1098/rstb.2000.0734.
4. Haggard P., Chambon V., Sidarus N. Exploring prospective sense of agency: The role of processing fluency, stimulus ambiguity and response conflict // Frontiers in Psychology. – 2017. – Vol. 8. – Art. 545. – DOI: 10.3389/fpsyg.2017.00545.
5. Hart S. G., Staveland L. E. Development of NASA-TLX: The results of empirical and theoretical research / Advances in Psychology. – 1988. – Vol. 52. – P. 139-183.
6. Hyland K. Metadiscourse: Investigating Interaction in Writing. – London: Continuum, 2005.
7. Kocielnik R., Amershi S., Bennett P. N. Will you take an imperfect AI?: Investigating designs to calibrate end-user expectations of AI systems / Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. – ACM, 2019. – Art. 129.
8. Legaspi R., Xu W., Konishi T., Ishikawa Y. The feeling of agency in human-AI collaboration // Knowledge-Based Systems. – 2024. – Vol. 286. – Art. 111298. – DOI: 10.1016/j.knosys.2023.111298.
9. Liao S. M., Sundar S. S. Building AI for trust and teamwork in human-AI interaction // Humanities and Social Sciences Communications. – 2022. – Vol. 9. – Art. 144. – DOI: 10.1057/s41599-022-01144-0.
10. Malmqvist L. Sycophancy in Large Language Models: Origins and Countermeasures // arXiv preprint. – 2024. – arXiv:2411.15287.
11. Moore J. W., Fletcher P. C. Sense of agency in health and disease: Cue integration approaches reviewed // Consciousness and Cognition. – 2012. – Vol. 21. – No. 1. – P. 59-68. – DOI: 10.1016/j.concog.2011.08.010.

12. Romeo G., Conti D. Automation bias in human-AI collaboration: a review and implications for XAI // AI & SOCIETY. – 2025. – Vol. 41. – № 1. – P. 259-278. – DOI: 10.1007/s00146-025-02422-7.
13. Sharma M., Tong M., Korbak T. et al. Towards explaining sycophancy in language models. – arXiv, 2024. – DOI: 10.48550/arXiv.2310.13548.
14. Shneiderman B. Human-centered artificial intelligence: trustworthy, safe & reliable // International Journal of Human-Computer Interaction. – 2020. – Vol. 36. – No. 6. – P. 497-508. – DOI: 10.1080/10447318.2020.1720478.
15. Sperber D., Clément F., Heintz C., Mascaro O., Mercier H., Origgi G., Wilson D. Epistemic vigilance / / Mind & Language. – 2010. – Vol. 25. – No. 4. – P. 359-393.
16. Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. Consciousness and Cognition, 17(1), 219-239. DOI: 10.1016/j.concog.2007.03.010
17. Turner J., Eisikovits N. Sycophancy in AI: a virtue-ethical perspective // Journal of Artificial Intelligence and Ethics. – 2026. – (in press).
18. Volpato C. et al Generative Confidants: What is the Experience of Trust in Emotional Support from Generative AI? // arXiv:2601.16656v1 [cs.HC]. – 2026.
19. Wang K., Li J., Yang S., Zhang Z., Wang D. When Truth Gets Suppressed: Revealing the Intrinsic Motivation Behind Sycophancy in Large Language Models // arXiv preprint. – 2025. – arXiv:2508.02087.
20. Wen, W., Imamizu, H. Sense of agency in perception, action and human-machine interactions // Nature Reviews Psychology. – 2022. – Vol. 1. – P. 211–222.
21. Yu H., Du S., Kurien A., van Wyk B. J., Liu Q Sense of Agency in Human–Machine Interaction Systems // Applied Sciences. – 2024. – Vol. 14. – No. 16. – Art. 7327. – DOI: 10.3390/app14167327.

References in Cyrillics

Русскоязычные источники

1. Bakhtin M. M. Problema teksta v lingvistike, filologii i drugikh gumanitarnykh naukakh // Literaturnoe obozrenie. – 1979. – № 10. – S. 3–10.
2. Lotman Yu. M. O dvukh modelyakh kommunikatsii v strukture kul'tury // Trudy po znakovym sistemam. – 1973. – Вып.. – С. 124-132.
3. Lotman Yu. M. Tekst v tekste // Trudy po znakovym sistemam. – 1984. – Vyp. 17. – S. 3-30.
4. Paducheva E. V. Vyskazyvanie i ego sootnesennost' s deystvitel'nost'yu: (referentsial'nye aspekty semantiki). – M.: Yazyki slavyanskoy kul'tury, 2010. – 424 s.
5. Petrenko V. F., Suprun A. P. Chelovek v predmetnom i mental'nom mire. Sushchestvuet li «ob'ektivnaya deystvitel'nost'»? Neokonchenny spor Bora s Einshteinom // Izvestiya Irkutskogo gosudarstvennogo universiteta. Seriya «Psikhologiya». – 2013. – T. 2. – № 2. – S. 62-82.
6. Ushakov D. V. Iskusstvennyy intellekt kak instrument psikhologicheskogo issledovaniya // Sibirskiy psikhologicheskiy zhurnal. – 2023. – № 90. – S. 188-200. – DOI: 10.17223/17267080/90/11.
7. Ushakov, D. V. Tekhnologii iskusstvennogo intellekta v psikhologii / D. V. Ushakov // Eksperimental'naya psikhologiya. – 2024. – T. 17. – № 4. – S. 182-189. – DOI: 10.17759/exppsy.2024170412.
8. Ushakova, T. N., Pavlova, N. D., Latynov, V. V., Tseptsov, V. A., Alekseev, K. I. Слово в действии: Intent-анализ политических дискуссий. – М.: Институт психологии РАН, 2000. – 448 с.
9. Zhuravlev A.L., Sergienko E.A. Analiz sovremennykh ponyatiy v psikhologii: chast' 2. Razrabotki uchenykh instituta psikhologii RAN // Psikhologicheskiy zhurnal, 2021, tom 42, No4, s. 5–15.
10. Znakov V. V. Ponimanie kak psikhologiya vozmozhnogo // Sibirskiy psikhologicheskiy zhurnal. – 2019. – № 72. – S. 6-20. – DOI: 10.17223/17267080/72/1..

Ноакк Наталия Вадимовна – к.п.н., ведущий научный сотрудник
ЦЭМИ РАН ORCID 0000-0001-8696-5767
n.noack@mail.ru

Костина Татьяна Анатольевна –научный сотрудник
ЦЭМИ РАН ORCID 0009-0006-1875-3774
kostina1@yandex.ru

Ключевые слова: генеративный ИИ, интерпретационная агентность, ощущаемая агентность, эпистемическая модальность, калиброванное доверие, сикофант, диалогичность, семиотическая граница, когнитивное трение, человеко-машинное взаимодействие, этика искусственного интеллекта.

Natalia Noack. Perception of Interpretational Agency in Dialogue with AI

Keywords: generative AI, interpretive agency, perceived agency, epistemic modality, calibrated trust, sycophancy, dialogism, semiotic boundary, cognitive friction, human–machine interaction, AI ethics.

Abstract

The proliferation of large language models has raised the issue of preserving human subjectivity in human–AI interaction. Unlike studies focusing on model architecture and ethics, this research emphasizes the perceptual and cognitive conditions under which users process AI-generated responses. We draw on the concept of the "digital confidant" (Ushakov, 2024), which allows us to discuss the user's ability to maintain the AI response as a hypothetical construct, compare it with their own experience, and either adjust or reject the proposed interpretation. The theoretical foundations of the study include cognitive models of agency, linguistic-pragmatic analysis of text modality, and the semiotic-dialogic approach (Bakhtin, Lotman). As a diagnostic indicator, we propose the Index of Interpretive Agency Consistency (IIAC), which links the user's subjective sense of autonomy to behavioral and linguistic markers. The paper contributes to the development of a human-centered model of AI interaction by foregrounding the user's perceptual strategies..