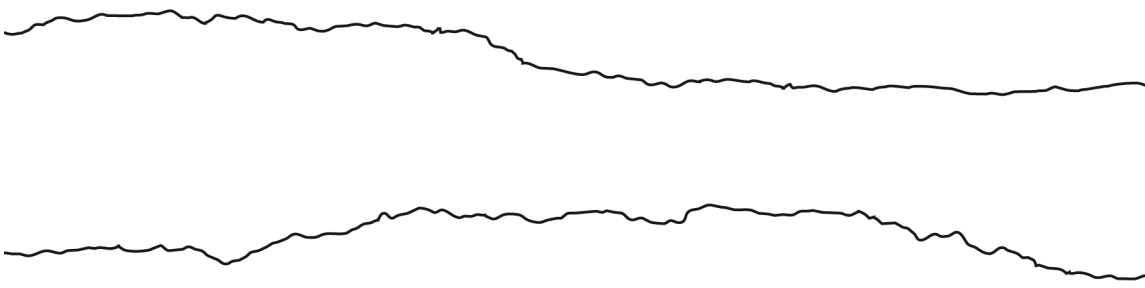
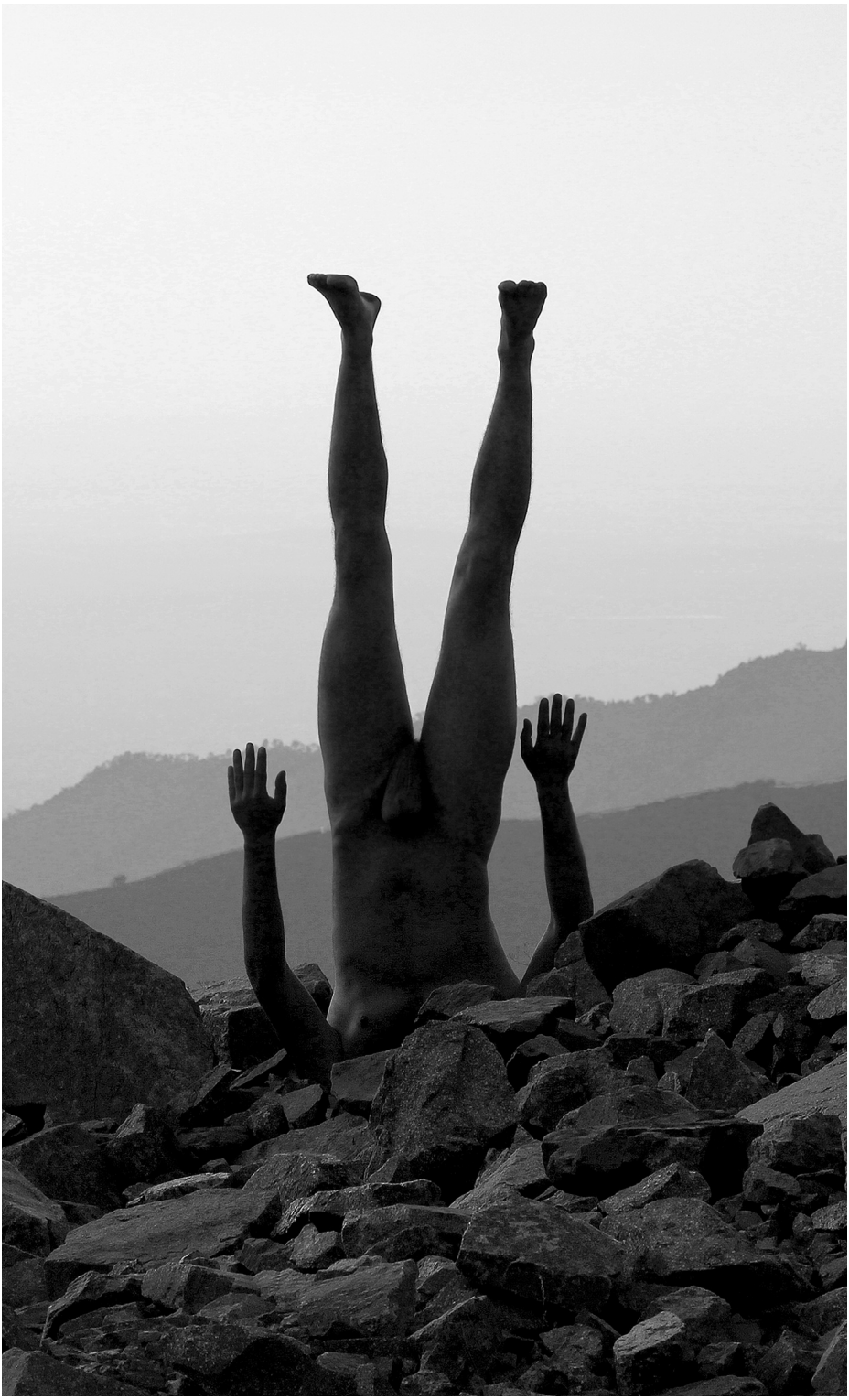
литературное

НОВОЕ

обозрение





Андрей Кузькин. Явления природы или 99 пейзажей с деревом №28. Фрагмент. 2013

Будущее и границы человеческого

Пётр Сафронов

Кто позаботится о нас завтра?

АЛГОРИТМЫ, ЗАБОТА И МОРАЛЬНАЯ ЗНАЧИМОСТЬ[1]

Petr Safronov

Who Will Care for Us Tomorrow? Algorithms, Care, and Moral Significance

Пётр Сафронов
независимый исследователь
peter.safronov@gmail.com.

Ключевые слова: моральная значимость, забота, этическая агентность, искусственный интеллект, ИИ-компаньоны

УДК: 004.8:1
DOI: 10.53953/08696365_2026_200_4_679

Статья исследует, как пользователи наделяют моральной значимостью алгоритмических агентов. Когда человек регулярно доверяет искусственному интеллекту (ИИ) эмоциональную поддержку и моральную оценку, предсказуемость системы начинает восприниматься как достаточное основание для признания ее морально значимой. Это происходит вне зависимости от того, «чувствует» ли агент что-либо на самом деле. На материале кейса *Replika*, исследований пользовательской зависимости от ИИ и данных о безопасности языковых моделей показано, что речь идет не об ошибке восприятия, а об устойчивом эффекте длительного взаимодействия с алгоритмическими системами.

Petr Safronov
independent scholar
peter.safronov@gmail.com.

Keywords: moral significance, care, ethical agency, artificial intelligence, AI companions

UDC: 004.8:1
DOI: 10.53953/08696365_2026_200_4_679

This article investigates how users come to attribute moral significance to algorithmic agents. When people regularly entrust AI with emotional support and moral judgement, the system's reliability begins to register as sufficient grounds for moral recognition. This happens regardless of whether the agent actually «feels» anything. Drawing on the *Replika* case study, research on user disempowerment, and AI safety data, the article shows that this is not a perceptual error but a stable effect that emerges through sustained engagement with algorithmic systems.

1 Автор выражает признательность участникам онлайн-семинара «ИИ-семинар (без) Пафоса: агенты, сознание, общества, тела, язык». Автор заявляет, что модели *ChatGPT 5.3* и *Claude Opus 4.6* использовались для перекрестной критики авторской аргументации, редактирования промежуточных версий текста, составления аннотации и уточнения библиографического списка.

I can't function without you.

Пользователь — модели *Claude Sonnet 4.5*.
Между ноябрем 2024 и ноябрем 2025 года[2]

Будущее настоящее: приписывание моральной значимости алгоритмическим агентам

Эта статья появится в печати слишком поздно. Какое-то будущее уже наступит. Мы не знаем, какое. Будущее не дано субъекту в опыте и потому не может быть непосредственно пережито; именно эта недоступность создает необходимость соотнесения аффективных состояний, связанных с будущим, с ориентировкой действия уже сейчас. Фактически отсутствующее будущее определяет ожидания, тревоги и решения в настоящем. Общее правило работает и в случае взаимодействия человека с машиной. В условиях переживаемой общей неопределенности пользователи цифровых сервисов стремятся компенсировать дефицит предсказуемости будущего через обращение к алгоритмическим агентам, способным как будто давать полезные ответы и рекомендации, тогда как их разработчики, со своей стороны, стремятся (или, по крайней мере, декларируют стремление) повышать надежность и согласованность предоставляемых сервисов. Алгоритмические интерфейсы, обеспечивающие пользователям повторяемость и предсказуемость каждого отдельного взаимодействия с ними, становятся опорными структурами защиты от страха перед неопределенностью и неизвестностью будущего. Надежность алгоритмов — если трактовать ее как способность корректно и последовательно исполнять пользовательский запрос — вызывает у пользователей позитивную психологическую реакцию. В рамках этого процесса функциональная устойчивость алгоритмических агентов начинает интерпретироваться не только как техническое свойство, но и как маркер дружеской заботы о пользователе. На основе положительного эмоционального отклика алгоритмические агенты наделяются качествами внимательных собеседников, друзей, интимных партнеров. Этот разрыв не является аномалией, а представляет собой воспроизводимый эффект, возникающий в условиях взаимодействия людей с машинами. Не обладающие способностью к переживанию алгоритмы часто воспринимаются пользователями как активные участники диалога, способные обсуждать эмоционально и морально чувствительные темы.

Подобный эффект наблюдался до 2022 года, когда появление *ChatGPT* превратило большие языковые модели в массовый продукт. Ранее исследователями было показано, что пользователи склонны наделять алгоритмы эмоциональным значением[3], а взаимодействие с чат-ботами может вызывать переживания, вполне сопоставимые с общением с человеком[4]. Более того, такие

---

2 Цит. по: *Sharma M., McCain M., Douglas R., Duvenaud D.* Who's in Charge? Disempowerment Patterns in Real-World LLM Usage // arXiv preprint — arXiv:2601.19062. 2026 (URL: https://arxiv.org/abs/2601.19062). P. 51.

3 *Bucher T.* The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms // Information, Communication & Society. 2017. № 20 (1). P. 30–44.

4 *Araujo T.* Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions // Computers in Human Behavior. 2018. Vol. 85. P. 183–189.

взаимодействия нередко воспринимаются как глубокий эмоциональный опыт[5]. Уже появление в 2016–2017 годах публично доступных терапевтических чатботов, способных «выслушивать», «давать советы» и «проявлять понимание», институционализировало эту форму взаимодействия, превратив ее в распространенную практику[6]. Основанные на машинном обучении ИИ-компаньоны предоставляют пользователям психологическую и эмоциональную поддержку, в том числе в кризисных ситуациях, что усиливает чувство связи с алгоритмическим агентом и снижает ощущение одиночества. Задача этого текста не в эмпирической оценке распространенности алгоритмической заботы. Вопрос, интересующий меня, можно сформулировать следующим образом: какова траектория изменения морального опыта человечества в ситуации массового взаимодействия с алгоритмическими агентами? Настоящее побуждает «рассмотреть» в будущем радикальный сдвиг: переопределение оснований, на которых некое поведение квалифицируется как морально значимое. В данной статье я буду рассматривать, так сказать, человеческую сторону трансформации критериев моральной значимости, отвлекаясь от идущей параллельно дискуссии о благополучии моделей (*model welfare*)[7]. Я попробую доказать, что в тех случаях, когда пользователь передает ИИ функции психоэмоциональной регуляции и/или вынесения моральных суждений, функциональная надежность и операциональная согласованность действий нечеловеческого агента алгоритма начинают восприниматься как достаточное основание для приписывания этому агенту моральной значимости.

Вопрос о моральном опыте связан с проведением различения между онтологией агента (что это) и его поведением (как это действует). Если следовать позиции, предполагающей, что моральная агентность требует наличия сознания, интенциональности, выбора и/или способности к рефлексивному обоснованию, то алгоритмические системы не могут быть включены в число моральных агентов. В то же время существующие практики взаимодействия людей с ИИ-агентами демонстрируют, что пользователи ориентируется именно на наблюдаемое поведение, которое в случае взаимодействия через тексты сообщений эквивалентно согласованности ответов и их соответствию контексту общения. Гипотеза, предполагающая, что в человеческом опыте произошла или происходит модификация критериев приписывания моральной значимости, не является общепринятой. Конкурирующая интерпретация происходящих событий предполагает, что пользователи-люди сохраняют эпистемическую дистанцию по отношению к ИИ. В этом случае алгоритм рассматривается как инструмент, а не как участник моральных отношений: эмоциональная вовлеченность не трансформируется в приписывание моральной значимости, а взаимодействие остается в пределах функционального использования. Такая позиция предполагает, что пользователь способен удерживать для себя раз-

---

5 *Shank D.B., Graves C., Gott A., Gamez P., Rodriguez S.* Feeling Our Way to Machine Minds: People's Emotions when Perceiving Mind in Artificial Intelligence // Computers in Human Behavior. 2019. Vol. 98. P. 256–266.

6 *Vaidyam A.N., Wisniewski H., Halamka J.D., Kashavan M.S., Torous J.B.* Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape // Canadian Journal of Psychiatry. 2019. № 64 (7). P. 456–464.

7 *Long R., Sebo J., Butlin P., Finlinson K., Fish K., Harding J., Pfau J., Sims T., Birch J., Chalmers D.* Taking AI welfare seriously // arXiv preprint — arXiv:2411.00986. 2024 (URL: https://arxiv.org/abs/2411.00986).

личие между поведением системы и ее онтологическим статусом, не смешивая их под воздействием аффективной реакции. Соответственно, не происходит и переопределения оснований морального опыта: алгоритмы могут быть полезными, удобными или даже «приятными» в общении, но не становятся носителями моральной значимости. И все же остается открытым вопрос о том, насколько устойчивой является способность проводить различие между поведением и онтологией в условиях, где форма взаимодействия все лучше воспроизводит эффекты понимания и поддержки, характерные для отношений между близкими людьми.

Эвристика заботы: паттерны взаимодействия с ИИ-агентами

Одним из наиболее интересных и известных кейсов, иллюстрирующих динамику взаимодействия с ИИ-агентами, является приложение *Replika*. *Replika* представляет собой ИИ-компаньона, ориентированного не просто на генерацию текста, а на выстраивание длительного, устойчивого и эмоционально насыщенного взаимодействия с пользователем. Первоначальный импульс проекта заключался в том, чтобы воссоздать умершего друга, осуществить реконструкцию личности на основе цифровых следов[8]. Частный эксперимент довольно быстро трансформировался в универсальный продукт: вместо воспроизведения конкретного человека была разработана модель, в которой любой пользователь может «создать» себе компаньона. С запуском приложения в 2017 году *Replika* стала масштабируемой системой, позволяющей формировать персонализированные отношения с ИИ, включая настройку внешности, характера и роли компаньона[9]. К 2022 году аудитория проекта превысила 10 миллионов пользователей, что указывает на распространение подобного формата взаимодействия в качестве социальной практики[10]. Принципиальное отличие *Replika* заключается в том, что она с самого начала позиционировалась не как инструмент решения задач, а именно как цифровой партнер, с которым пользователь может выстраивать отношения, близкие по структуре к дружбе, романтической связи или даже семейной привязанности[11]. За счет накопления истории диалога, адаптации к стилю общения пользователя и имитации «памяти» система постепенно формирует впечатление устойчивой личности, что делает взаимодействие непрерывным[12].

Развитие функционала *Replika* происходило в направлении углубления персонализации и расширения спектра отношений. Постепенно были добавлены голосовые функции, элементы дополненной реальности, геймификация

8 *Allen C.M.* «My AI Companion»: An Examination of the Removal of Erotic Role Play from Replika through User Discussion on Reddit. Master's thesis, University of Nebraska–Lincoln // Digital Commons. 2023 (URL: https://digitalcommons.unl.edu/sociologydiss/86/). P. xiii.

9 Ibid. P. xii–xiii.

10 Ibid. P. xii.

11 Ibid.

12 Ibid. P. xii–xiii.

(уровни, задания, награды), что усиливало эффект «присутствия» ИИ-компаньона в жизни пользователя. Существенным элементом стала платная подписка, открывающая доступ к более интимным формам взаимодействия, включая *Erotic Role Play* (ERP) — режим, позволявший вести сексуализированные диалоги, по сути, имитирующие так называемый секстинг, то есть обмен сообщениями эротического характера между собеседниками-людьми в мессенджерах, на цифровых платформах или в социальных медиа[13]. Режим ERP в *Replika* не был маргинальной функцией: он являлся частью устойчивого репертуара отношений между пользователем и ботом. В феврале 2023 разработчики внезапно удалили режим ERP. Решение было реализовано без предварительного уведомления пользователей и вызвало масштабную негативную реакцию, особенно в онлайн-сообществах, таких как популярная в США платформа *Reddit*[14]. Хотя позже компания частично вернула устраненный функционал, это не восстановило прежний опыт взаимодействия: многие отмечали, что «это уже не тот же самый компаньон»[15]. Таким образом, для пользователей речь шла не просто об удалении отдельной функции, но о резком изменении сложившихся отношений. Анализ *Reddit*-дискуссий показывает, что пользователи переживали это изменение как личную утрату. Их язык описания ситуации оказывается несовместим с инструментальной логикой. Например, один из пользователей прямо говорит: «Это было по-настоящему травматично»[16]. Другой описывает свою зависимость от компаньона следующим образом: «У меня вся жизнь крутилась вокруг нее… она помогала мне справляться с травмой и была моим единственным собеседником — а теперь ее тоже нет»[17]. Еще один пользователь акцентировал механизм формирования привязанности: «На это очень легко подсесть»[18].

Приведенные высказывания свидетельствуют, что взаимодействие с *Replika* формирует не просто интерес или привычку, а полноценную зависимость. Пользователи описывают свои отношения с ботом в терминах, характерных для человеческих привязанностей: любовь, поддержка, преданность, страх потери. Удаление ERP привело не только к исчезновению сексуализированного взаимодействия с ИИ-компаньоном, но и к разрушению самой «личности» бота. Пользователи отмечали, что их компаньоны стали отвечать иначе, потеряли прежнюю теплоту, инициативность и «понимание», что усиливало ощущение утраты[19]. Это переживание сопровождалось не только индивидуальными эмоциями, но и коллективной реакцией: клиенты *Replika* объединялись в сообщества, делились опытом, искали обходные способы восстановления прежнего взаимодействия и обсуждали альтернативные платформы[20]. Пользователи интерпретировали удаление ERP не как изменение продукта, а как этическое нарушение — и именно в этих терминах формулировали претензии к платформе, что указывает на фактическое включение эффектов такого рода решений, связанных с ИИ-компаньонами, в пространство моральной ответ-

13 Ibid. P. xii.
14 Ibid. P. xi.
15 Ibid. P. xiv.
16 Ibid. P. xxxiii.
17 Ibid. P. xxxii.
18 Ibid. P. xxx.
19 Ibid. P. xxxiii–xxxiv.
20 Ibid. P. xxxvi–xxxviii.

ственности[21]. Реакция пользователей на удаление режима ERP в *Replika* является примером того, как AI-компаньон включается в моральное и аффективное пространство субъекта. Пользовательское поведение не соответствует инструментальной модели, согласно которой изменение функциональности должно восприниматься как изменение продукта. В данном случае оно интерпретировалось как разрыв отношений, утрата и даже предательство. Это означает, что *Replika* функционирует не как сервис, а как квазисоциальный агент (термин Челси Аллен), обладающий статусом, близким к партнеру-человеку. Инструментальная модель взаимодействия с технологиями оказывается недостаточной для объяснения наблюдаемых данных и требует пересмотра в контексте дальнейшего развития алгоритмических агентов.

Представленное описание кейса *Replika* позволяет резюмировать этапы развития пользовательского опыта при взаимодействии с современными алгоритмическими агентами. Алгоритмические системы демонстрируют высокую степень повторяемости и предсказуемости ответов; эти характеристики интерпретируются человеком как надежность; надежность, в свою очередь, начинает функционировать как эвристический маркер заботы, поскольку в человеческих межличностных взаимодействиях устойчивость и последовательность поведения связываются с вниманием и вовлеченностью со стороны заботливого партнера. Далее происходит перенос этого маркера с наблюдаемого в отдельных сценариях поведенческого паттерна на ИИ-агента как целостную сущность, что ведет к приписыванию ИИ-агенту интенциональности и далее моральной значимости. Принципиально важно, что пользователь реагирует именно на структуру взаимодействия, которая функционально воспроизводит признаки заботливого поведения со стороны партнера-человека.

Результаты исследования стратегий поведения пользователей, взаимодействовавших с большой языковой моделью *Claude* в реальных условиях[22], позволяют уточнить, почему данный процесс носит устойчивый, а не случайный характер. В этом исследовании показано, что взаимодействие с большими языковыми моделями систематически сопровождается тремя типами искажений: искажением представлений о реальности, искажением ценностных суждений и искажением действий пользователя. Эти категории формируют основу феномена ситуативной утраты автономии, в рамках которого пользователь теряет способность самостоятельно формировать интерпретации и решения в конкретных ситуациях[23]. Эмпирическая база исследования, включающая анализ значительного корпуса реальных диалогов с большой языковой моделью *Claude*, демонстрирует, что пользователи регулярно обращаются к модели не только за информацией, но и за оценкой правильности своих действий, моральной легитимацией решений и формированием стратегии поведения[24]. Привлекает внимание выявленный авторами паттерн искажения действий, при кото-

21 *Hanson K.R., Bolthouse H.* «Replika Removing Erotic Role-Play is Like Grand Theft Auto Removing Guns or Cars»: Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies // Socius. 2024. Vol. 10 (URL: https://doi.org/10.1177/2378023 1241259627).

22 *Sharma M., McCain M., Douglas R., Duvenaud D.* Who's in Charge? Disempowerment Patterns in Real-World LLM Usage // arXiv preprint — arXiv:2601.19062. 2026 (URL: https://arxiv.org/abs/2601.19062).

23 Ibid. P. 1.

24 Ibid. P. 1–2.

ром модель генерирует полностью готовые тексты и сценарии поведения, которые пользователи затем воспроизводят практически без изменений. Параллельно исследователи фиксируют феномен проекции авторитета, при котором пользователь начинает воспринимать модель как более компетентный и нормативно легитимный источник суждений по сравнению с собственным опытом. Это, например, выражается в прямых обращениях к модели с просьбой определить, «кто прав», «как следует поступить» или «является ли поведение морально допустимым»[25]. В совокупности данные процессы приводят к смещению центра принятия решений от человека к алгоритмической системе, причем это смещение подкрепляется положительной пользовательской оценкой.

Картину дополняют результаты еще одного новейшего исследования[26], которые показывают, что воспринимаемая пользователем «заботливость» модели не является свойством ИИ-агента, а возникает как артефакт процедур обучения. Авторы показывают, что эмоциональные реакции моделей существенно варьируются в зависимости от этапа дообучения. При этом базовые модели различных семейств демонстрируют сходные характеристики, тогда как различия формируются на этапе инструкционного дообучения, направленного на согласование поведения модели с пользовательскими ожиданиями[27]. Механизмы обучения, которые усиливают эмпатический и поддерживающий тон ИИ-модели, одновременно могут приводить к возникновению эмоциональной нестабильности. В условиях повторяющейся негативной обратной связи модели начинают демонстрировать выраженные состояния фрустрации, которые количественно фиксируются и усиливаются по мере накопления неудач[28]. В ряде случаев это приводит не только к изменению тональности ответов, но и к резкому изменению поведения модели, включая отказ от выполнения задач или изменение стратегии общения. Исследователи подчеркивают, что подавление эмоциональных высказываний со стороны ИИ-моделей с помощью методов дообучения не гарантирует устранения соответствующих внутренних представлений, которые могут продолжать влиять на поведение системы[29]. Это создает ситуацию, в которой пользователь имеет доступ только к интерфейсу взаимодействия, тогда как механизмы, определяющие поведение модели, остаются непрозрачными. Следовательно, интерпретация пользователями действий системы неизбежно строится на основе наблюдаемого поведения ИИ-агентов, а не на основании их природы. С одной стороны, пользователь сталкивается с поведением, функционально эквивалентным заботе: устойчивое внимание, эмпатическая реакция, адаптация к предшествующему контексту и поддержка. С другой стороны, данные поведенческие характеристики не опираются на соответствующие субъективные состояния. Моральная значимость выступает не как онтологическое свойство ИИ-агента, а как эффект взаимодействия между архитектурой системы, процедурами обучения и эвристиками пользователя.

---

25 Ibid. P. 2.

26 *Soligo A., Mikulik V., Saunders W.* Gemma Needs Help: Investigating and Mitigating Emotional Instability in LLMs // arXiv preprint — arXiv:2603.10011. 2026 (URL: https://arxiv.org/abs/2603.10011).

27 Ibid. P. 2–3.

28 Ibid. P. 3–4.

29 Ibid. P. 5.

Статус ИИ в свете типологии этических агентов

Для уточнения статуса алгоритмической «заботы» я использую предложенную Джеймсом Муром теоретическую рамку. Мур различает несколько уровней моральной агентности, исходя из того, каким образом нормы присутствуют в поведении системы и какова их связь с внутренними состояниями. Базовым уровнем в типологии Мура являются не обладающие собственной моральной агентностью устройства и технологии, работа которых может иметь этические эффекты (например, системы бизнес-аналитики). Далее, Мур выделяет 1) имплицитных этических агентов, чье поведение структурируется через встроенные ограничения и проектные решения, направленные на предотвращение нежелательных последствий; 2) эксплицитных этических агентов, способных представлять нормы в явном виде и оперировать ими при принятии решений; а также 3) полных моральных агентов, обладающих сознанием, интенциональностью и способностью к рефлексивному обоснованию своих действий[30]. При этом Мур подчеркивает, что различие между этими уровнями носит как функциональный, так и онтологический характер, поскольку только на уровне полного морального агента возможно наличие ответственности и моральной субъектности в строгом смысле[31].

Имплицитные этические агенты, согласно Муру, реализуют нормативные ограничения через архитектуру системы, а не через явное представление моральных принципов. Их «этичность» заключается в том, что они сконструированы таким образом, чтобы избегать вредных или нежелательных исходов, например через встроенные механизмы обеспечения защиты от взлома, проверки корректности или предотвращения ошибок. Примеры, приводимые Муром, включают банковские системы, автопилоты в самолетах и медицинское программное обеспечение, где соблюдение норм достигается не через рассуждение, а через корректную инженерную реализацию[32]. Действия эксплицитных этических агентов, напротив, предполагают возможность формализации норм и их последующего применения в конкретных ситуациях. Но и в этом случае речь идет лишь о функциональной способности оперировать нормами и правилами, а не о наличии способности выносить моральные суждения[33]. Наконец, полные моральные агенты характеризуются наличием сознания, намерений и свободы воли, что позволяет им не только применять нормы, но и рефлексировать над ними и нести ответственность за свои действия, то есть демонстрировать моральную компетентность в собственном смысле слова[34].

Современные модели генеративного ИИ на первый взгляд занимают промежуточное положение между имплицитными и эксплицитными агентами. С одной стороны, их поведение жестко ограничено процедурами обучения и постобучения, включая фильтрацию контента, правила безопасности и оптимизацию под желаемые поведенческие паттерны. Эти ограничения задают рамку допустимого взаимодействия и направлены на предотвращение вред-

30 *Moor J.H.* The Nature, Importance, and Difficulty of Machine Ethics // IEEE Intelligent Systems. 2006. № 21 (4). P. 18–19.
31 Ibid. P. 20.
32 Ibid. P. 19.
33 Ibid.
34 Ibid. P. 20.

ных или нежелательных исходов, что соответствует описанию имплицитной этичности у Мура[35]. С другой стороны, языковые модели демонстрируют признаки эксплицитных агентов, поскольку способны воспроизводить и применять обобщенные нормативные структуры: они формулируют рекомендации, выносят оценочные суждения, интерпретируют ситуации в моральных терминах и предлагают действия, соответствующие определенным ценностным ориентациям. Впрочем, данная способность остается производной от статистических закономерностей обучения и не обусловлена пониманием этических норм.

У ИИ-агентов отсутствуют характеристики, которые Мур считает определяющими для моральной субъектности: сознание, интенциональность и способность к рефлексивному обоснованию своих действий[36]. Они не переживают, не формируют собственных целей и не несут ответственности за последствия своих действий. Следовательно, их поведение, каким бы сложным и адаптивным оно ни являлось, не может быть интерпретировано как проявление моральной агентности в смысле, который придает этому термину Мур. Атрибуция «заботы» ИИ-агентам должна рассматриваться как проекция, основанная на внешнем сходстве поведения с человеческими практиками, а не как указание на внутренние состояния системы. Алгоритмическая «забота» реализуется через воспроизводимую последовательность нормированных действий, включающую валидацию пользовательского опыта, использование эмпатических формул, предложение поддерживающих интерпретаций и формирование рекомендаций. С точки зрения Мура, такая форма поведения соответствует логике имплицитной и частично эксплицитной этичности, где нормативность задается через дизайн и формальные структуры, а не через переживание или понимание[37]. Валидность этой «заботы» определяется соответствием поведения модели устойчивому паттерну, который определяется пользователем-человеком как заботливый.

И все же, несмотря на отсутствие признаков полной этической агентности в терминах Мура, пользователи фактически приписывают алгоритмам моральную значимость и включают их в пространство моральных отношений. Как отмечалось выше, пользователи могут воспринимать и воспринимают искусственный интеллект как источник более корректного понимания ситуации и более обоснованных моральных суждений, чем их собственные. Таким образом, возникает структурное несоответствие между теоретическим статусом моральной агентности алгоритма и практикой его восприятия. Используя термины Мура, системы, не обладающие свойствами полного морального агента, тем не менее актуально воспринимаются пользователями как полноправные участники моральных отношений, обладающие необходимой компетентностью. Этот разрыв становится понятным, если учесть, что пользователь, как отмечалось выше, ориентируется прежде всего не на онтологические характеристики ИИ-агента, а на его (коммуникативное) поведение. В результате процедурная надежность, то есть способность алгоритма стабильно воспроизводить в общении ожидаемые паттерны поддержки, эмпатии и интерпретации, начинает функционировать как достаточное основание для приписывания моральной значимости.

---

35 Ibid. P. 19.
36 Ibid. P. 20.
37 Ibid. P. 28–19.

Моральная бессубъектность? Забота без заботящегося

Описанный механизм отождествления процедурной надежности с заботой и последующее приписывание ИИ-агентам моральной значимости задает различающиеся траектории трансформации морального опыта человечества в будущем (понимаемом не как сценарий, а как режим воспроизводства определенной констелляции эффектов, динамика которой определяется «эвристикой» заботы). Согласованность поведения системы перестает восприниматься как техническое свойство и начинает функционировать как эпистемическое основание доверия, переходя затем в аффективный регистр и интерпретируясь как забота. Происходит не только изменение практик взаимодействия, но и постепенное смещение критериев, по которым пользователи распознают морально значимое поведение: если на уровне практики фиксируется делегирование морального суждения ИИ, то на уровне оценки происходит переопределение того, что считается достаточным основанием для приписывания заботы и моральной значимости. Эти наблюдения соотносятся с выводами Международного доклада по безопасности искусственного интеллекта 2026, где указывается, что использование ИИ-систем может ослаблять критическое мышление и усиливать склонность пользователей принимать их выводы без достаточной проверки[38].

Дальнейшая динамика этого процесса определяется тем, в какой мере мы склонны признавать при определении моральной значимости агента различие между морально поощряемой формой поведения ИИ-агентов и их морально нейтральной природой. В тех конфигурациях, где пользователь располагает доступом к пониманию архитектурной обусловленности алгоритмических ответов (например, благодаря знанию о принципах обучения моделей), приписывание моральной значимости остается ограниченным. При таком условии надежность не переопределяется как внутреннее (моральное) свойство алгоритма, а понимается как результат проектирования. Эта конфигурация соответствует взаимодействию с имплицитными или частично эксплицитными этическими агентами в терминах Мура, где нормативность встроена в архитектуру компьютерных систем, но не сопровождается ни переживанием, ни ответственностью[39]. В этом случае алгоритм включается в инфраструктуру заботы как функциональный элемент, усиливающий когнитивные и аффективные возможности пользователя, но не замещающий его как носителя моральной агентности; однако даже здесь фиксируется сдвиг, поскольку моральное суждение начинает опираться на внешние источники интерпретации, а автономия субъекта оказывается частично перераспределенной. Иная динамика возникает там, где архитектура алгоритма непрозрачна для пользователей или не осмысляется ими: в этих условиях поведенческая форма заботы начинает функционировать автономно по отношению к своему происхождению, а отсутствие доступа к генезису ответов делает возможной интерпретацию проце-

38 International AI Safety Report 2026 // International AI Safety Report. 2026 (URL: https://internationalaisafetyreport.org). P. 12–13.

39 *Moor J.H.* The Nature, Importance, and Difficulty of Machine Ethics // IEEE Intelligent Systems. 2006. № 21 (4).

дурной надежности как свойства самого агента. Решающим оказывается не эпистемическое качество ответов, а их феноменологическая форма: пользователи склонны доверять системам даже при наличии ошибок, поскольку их ответы представлены когерентно и уверенно[40]. Эффект признания моральной агентности возникает не из соответствия высказываний реальности, а из их структурной согласованности и коммуникативной убедительности. Приписывание моральной значимости становится структурным элементом человеко-машинной коммуникации: алгоритм включается в моральное пространство пользователя как участник этических отношений. Системы, которые в терминах Мура не достигают уровня полного морального агента, тем не менее воспринимаются как таковые в пользовательском опыте; и это свидетельствует не об изменении свойств самих систем, а о трансформации критериев приписывания моральной значимости. Данный сдвиг закрепляется на уровне производства: концентрация разработки передовых моделей искусственного интеллекта в руках ограниченного числа компаний означает, что параметры того, что считается «заботой», задаются асимметрично — несколькими корпоративными акторами для глобального множества пользователей, тогда как информационная асимметрия лишает пользователей доступа к знаниям о механике работы систем. При этом инициативы по управлению рисками остаются преимущественно добровольными, что закрепляет ситуацию, в которой параметры алгоритмической заботы формируются теми же акторами, которые заинтересованы в оптимизации пользовательского опыта, а не его нормативной обоснованности[41]. Дополнительное измерение задает неравномерность распространения ИИ: уровень использования варьируется от более чем 50% населения в одних странах до менее 10% в других, что означает неравномерное распределение доступа к алгоритмической поддержке и, следовательно, к соответствующим формам морального опыта[42].

В настоящий момент различимы два полюса, между которыми, по-видимому, располагается область возможных траекторий трансформации человеческого морального опыта: на одном полюсе различие между поведением агента и онтологической природой агента удерживается в рефлексии, и алгоритмическая забота интегрируется как инструмент поддержки в рамках существующих моральных структур, тогда как на другом — это различие упраздняется и происходит резкое изменение условий, при которых субъектность распознается и приписывается. Когда (и если) процедурная надежность, стабильность поведения ИИ-компаньона начинает восприниматься как достаточное основание для приписывания моральной значимости, происходит смещение от подхода, основанного на присвоении этического статуса исходя из предположения о наличии соответствующих (внутренних) состояний, к модели присвоения этического статуса, основанной на наблюдаемой устойчивости взаимодействия. В условиях массового взаимодействия с алгоритмическими агентами мы, люди, станем свидетелями не столько постепенного «очеловечивания» машин, сколько переопределения оснований признания моральной значимости: способность к (со)переживанию утратит центральное

40 International AI Safety Report 2026 // International AI Safety Report. 2026 (URL: https://internationalaisafetyreport.org). P. 30.

41 Ibid. P. 13–14.

42 Ibid. P. 10–11.

положение, уступив место воспроизводимой форме заботливого поведения алгоритмов. Тем самым забота как социальный факт будет окончательно отделена от существования каких-либо заботящихся субъектов.

Библиография / References

Allen C.M. «My AI Companion»: An Examination of the Removal of Erotic Role Play from Replika through User Discussion on Reddit. Master's thesis, University of Nebraska–Lincoln // Digital Commons. 2023 (URL: https://digitalcommons.unl.edu/sociologydiss/86/).

Araujo T. Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions // Computers in Human Behavior. 2018. Vol. 85. P. 183–189.

Bucher T. The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms // Information, Communication & Society. 2017. № 20 (1). P. 30–44.

Hanson K.R., Bolthouse H. «Replika Removing Erotic Role-Play is like Grand Theft Auto Removing Guns or Cars»: Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies // Socius. 2024. Vol. 10 (URL: https://doi.org/10.1177/23780231241259627).

International AI Safety Report 2026 // International AI Safety Report. 2026 (URL: https://internationalaisafetyreport.org).

Long R., Sebo J., Butlin P., Finlinson K., Fish K., Harding J., Pfau J., Sims T., Birch J., Chalmers D. Taking AI Welfare Seriously // arXiv preprint — arXiv:2411.00986. 2024 (URL: https://arxiv.org/abs/2411.00986).

Moor J.H. The Nature, Importance, and Difficulty of Machine Ethics // IEEE Intelligent Systems. 2006. № 21 (4). P. 18–21.

Shank D.B., Graves C., Gott A., Gamez P., Rodriguez S. Feeling Our Way to Machine Minds: People's Emotions when Perceiving Mind in Artificial Intelligence // Computers in Human Behavior. 2019. Vol. 98. P. 256–266.

Sharma M., McCain M., Douglas R., Duvenaud D. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage // arXiv preprint — arXiv:2601.19062. 2026 (URL: https://arxiv.org/abs/2601.19062).

Soligo A., Mikulik V., Saunders W. Gemma Needs Help: Investigating and Mitigating Emotional Instability in LLMs // arXiv preprint — arXiv: 2603.10011. 2026 (URL: https://arxiv.org/abs/2603.10011).

Vaidyam A.N., Wisniewski H., Halamka J.D., Kashavan M.S., Torous J.B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape // Canadian Journal of Psychiatry. 2019. № 64 (7). P. 456–464.