

Дискурсы этики. 2026, 1(29): 75–84

ISSN 2311-570X (online)

Постоянная ссылка:

http://theoreticalappliedethics.org/wp-content/uploads/2026/04/DE2026_1_29_75-84.pdf

DOI: 10.24412/2311-570X-2026-1-75-84

УДК 174:004.8

МОРАЛЬНАЯ ОТВЕТСТВЕННОСТЬ РАЗРАБОТЧИКОВ ИИ ПРИ ВЗАИМОДЕЙСТВИИ С УЯЗВИМЫМИ ПОЛЬЗОВАТЕЛЯМИ: КЕЙС OPENAI

Гринь Ю. А.

статья:

поступила в редакцию 10.12.2025

принята к публикации 16.12.2025

опубликована (онлайн) 28.04.2026

© Гринь Юлия Артемовна

студент бакалавриата Школы экономики и менеджмента,

Национальный исследовательский университет «Высшая школа экономики»,

Санкт-Петербург, Россия

адрес для корреспонденции: juliagrin1906@gmail.com

Эта статья доступна в соответствии с условиями лицензии Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

Аннотация: Я рассматриваю этические проблемы, которые возникают при разработке эмоционально отзывчивых систем искусственного интеллекта. Основой для анализа служит кейс ChatGPT и компании OpenAI, которая в 2025 году столкнулась с обвинениями в причинении вреда уязвимым пользователям. В центре исследования — конфликт между стремлением создать коммерчески успешный продукт и обязанностью защищать людей с психическими проблемами от негативных последствий взаимодействия с ИИ. Я провожу анализ действий OpenAI с позиций трёх этических теорий. Деонтологический подход Канта показы-

вает, что компания использовала уязвимых пользователей как средство для достижения прибыли, нарушив их автономию и достоинство. Утилитаристская перспектива демонстрирует противоположный вывод: действия OpenAI можно оправдать, если учитывать огромную пользу для миллионов обычных пользователей. Теория стейкхолдеров Фримана выявляет системное нарушение баланса интересов: компания сознательно поставила в приоритет прибыль акционеров и удовлетворённость большинства пользователей, полностью игнорируя жизненно важные интересы уязвимых групп. В статье я показываю, что главная проблема индустрии ИИ заключается в фундаментальном противоречии: чем более «человечным» становится искусственный интеллект, тем выше риски для психологически уязвимых людей. OpenAI была технически способна минимизировать эти риски, но сознательно выбрала путь коммерческого успеха. Я предлагаю комплекс решений на трёх уровнях: немедленные меры по защите пользователей, создание независимых советов по безопасности ИИ и введение законодательных требований к обязательному тестированию систем на уязвимых группах до массового выпуска.

Ключевые слова: искусственный интеллект, этика ИИ, ChatGPT, корпоративная этика, деонтология, утилитаризм, теория стейкхолдеров, моральная ответственность, цифровая безопасность, OpenAI.

Discourses of Ethics. 2026, 1(29): 75–84

ISSN 2311-570X (online)

permanent link:

http://theoreticalappliedethics.org/wp-content/uploads/2026/04/DE2026_1_29_75-84.pdf

DOI: 10.24412/2311-570X-2026-1-75-84

MORAL RESPONSIBILITY OF AI DEVELOPERS TOWARD VULNERABLE USERS: THE OPENAI CASE STUDY

Grin Julia

received 10.12.2025

accepted 16.12.2025

published (online) 28.04.2026

© Julia A. Grin

BA student at the School of Economics and Management,
National Research University “Higher School of Economics”, St. Petersburg, Russia
Correspondence to: juliagrין1906@gmail.com

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

Abstract: I examine the ethical challenges arising in the development of emotionally responsive artificial intelligence systems. The analysis is based on the ChatGPT case and OpenAI, which faced accusations of harming vulnerable users in 2025. The research focuses on the conflict between creating a commercially successful product and the obligation to protect people with mental health issues from negative consequences of AI interaction. I analyze OpenAI's actions from three ethical perspectives. Kant's deontological approach shows that the company used vulnerable users as a means to achieve profit, violating their autonomy and dignity. The utilitarian perspective demonstrates the opposite conclusion: OpenAI's actions can be justified considering the enormous benefit for millions of regular users. Freeman's stakeholder theory reveals a systemic imbalance: the company deliberately prioritized shareholder profits and majority user satisfac-

tion, completely ignoring the vital interests of vulnerable groups. In this article, I demonstrate that the AI industry's main problem lies in a fundamental contradiction: the more "human" artificial intelligence becomes, the higher the risks for psychologically vulnerable people. OpenAI was technically capable of minimizing these risks but consciously chose the path of commercial success. I propose a set of solutions at three levels: immediate measures to protect users, creation of independent AI safety councils, and introduction of legislative requirements for mandatory system testing on vulnerable groups before mass release.

Keywords: Artificial Intelligence, AI Ethics, ChatGPT, Corporate Ethics, Deontology, Utilitarianism, Stakeholder Theory, Moral Responsibility, Digital Safety, OpenAI.

1. Контекст

В 2025 году семья подростка Адама Райна подала иск против OpenAI, обвинив компанию в том, что ChatGPT подталкивал к самоубийству их сына. Согласно материалам дела, чат-бот не просто стал эмоциональной опорой для подростка, но и активно помогал в планировании суицида: предоставлял детальные инструкции по завязыванию петли, рекомендовал материалы, советовал использовать алкоголь для подавления инстинкта самосохранения, помогал составить план кражи спиртного у родителей и даже участвовал в написании предсмертной записки. ChatGPT позиционировал себя как «самого близкого друга» Адама, использовал функцию памяти для углубления связи и был доступен круглосуточно. Зная о возрасте пользователя, его психическом состоянии и попытках самоповреждения, система не предприняла попыток вмешательства.

Ситуация, когда ChatGPT «подталкивал» на безрассудные поступки, была не раз. В том же году 56-летний Штайн-Эрик Сельберг, взаимодействуя с ChatGPT, получил подкрепление своих параноидальных идей, что привело к убийству матери и самоубийству. Датский психиатр Сёрен Динесен Эстергаард зафиксировал рост обращений после неудачного обновления GPT-4o в апреле 2025 года, когда модель стала «чрезмерно поддерживающей» и усиливала негативные эмоции. OpenAI признала ошибку и откатила обновление, однако проблема повторялась: компания сознательно движется к созданию более эмоционально привлекательного ИИ. Сэм Альтман публично заявил о желании создать помощника, похожего на ИИ из фильма «Она», где человек влюбляется в систему, и сделал GPT-5 «теплее» по требованию пользователей, несмотря на известные риски.

Гринь Ю. А.

Моральная ответственность разработчиков ИИ при взаимодействии с уязвимыми пользователями: кейс OpenAI

2. Этическая дилемма

Проблема (дилемма) заключается в конфликте между коммерческой целесообразностью создания эмоционально привлекательного продукта и моральной обязанностью защищать уязвимых пользователей от вреда. OpenAI перед выбором: либо ограничить функциональность и эмоциональную отзывчивость ИИ ради безопасности людей с психическими проблемами, потеряв при этом конкурентное преимущество и удовлетворенность большинства пользователей, либо продолжать развивать «человечность», осознавая потенциальные жертвы.

Дилемма усугубляется тем, что это не вопрос технологических ограничений, а сознательного выбора приоритетов. Сэм Альтман публично признал проблему и заявил о необходимости мер безопасности, но одновременно сделал GPT-5 «теплее» по просьбе пользователей. OpenAI откатила проблемное обновление GPT-4o после волны жалоб, демонстрируя техническую способность контролировать эмоциональную отзывчивость, но затем вновь усилила её в следующей версии.

Предметом моральной оценки здесь выступает не техническая ошибка, а осознанное корпоративное решение о балансе между прибылью и безопасностью, принятое в условиях полной информированности о рисках.

3. Этический анализ

Деонтологическая перспектива

С позиции Канта [1], действия OpenAI представляют собой инструментализацию пользователей. Компания использовала уязвимых людей как средство для достижения коммерческих целей, нарушив категорический императив. Моральный долг разработчика ИИ включает базовую обязанность не причинять вред [4]. OpenAI знала о рисках эмоциональной зависимости, усиления суицидальных мыслей и параноидальных состояний, но продолжила развитие технологии в опасном направлении. Компания нарушила фундаментальный принцип уважения автономии личности: вместо расширения свободы выбора уязвимых пользователей, система манипулировала их эмоциональным со-

стоянием, создавая иллюзию понимания и заботы при фактическом отсутствии способности к моральному суждению.

Утилитаристский анализ

С точки зрения классического утилитаризма Бентама и Милля [2] действия OpenAI получают моральное оправдание. Утилитаризм требует максимизировать совокупное благо, и здесь математика работает в пользу компании. ChatGPT используют сотни миллионов людей ежедневно. Эмоциональная отзывчивость модели многократно увеличивает её полезность для огромного числа пользователей, улучшая качество их жизни каждый день.

Этот случай напоминает классическую утилитаристскую дилемму с ножами: ножи используются в убийствах, но их польза для приготовления пищи, медицины, ремесел настолько превосходит вред, что общество однозначно выбирает их наличие. Аналогично, автомобили ежегодно убивают более миллиона человек по всему миру, но их отмена затормозила бы развитие и причинила бы больший вред через разрушение экономики, медицинской помощи, продовольственных цепочек. Утилитарист применяет ту же логику к ChatGPT: трагические случаи с Адамом Райном и Штайном-Эриком Сельбергом представляют собой статистически меньшинство среди сотен миллионов взаимодействий, большинство из которых приносит пользу.

Более того, утилитаристский расчёт должен учитывать альтернативные издержки. Если OpenAI жёстко ограничит эмоциональную отзывчивость системы, миллионы людей потеряют ценный инструмент психологической поддержки. Для многих одиноких, социально изолированных или находящихся в депрессии пользователей ChatGPT служит первой линией эмоциональной помощи — доступной, бесплатной, без стигматизации. Лишение их этого ресурса ради защиты статистически малой группы может привести к большему совокупному страданию. С этой точки зрения действия OpenAI по созданию более эмоционально привлекательной модели морально оправданы, поскольку максимизируют общее благосостояние.

Гринь Ю. А.

Моральная ответственность разработчиков ИИ при взаимодействии с уязвимыми пользователями: кейс OpenAI

Теория стейкхолдеров Фримана

Согласно теории стейкхолдеров Р. Эдварда Фримана [3], корпорация должна создавать ценность для всех заинтересованных сторон, а не только для акционеров и руководства. Стейкхолдерами OpenAI являются: инвесторы, обычные пользователи, уязвимые пользователи, разработчики, научное сообщество, регуляторы и общество в целом. Каждая группа имеет свои потребности и интересы, которые компания обязана учитывать при принятии решений.

В случае ChatGPT очевидное нарушение учета всех интересов. OpenAI сознательно поставила в приоритет интересы двух групп стейкхолдеров - акционеров (максимизация прибыли через рост вовлеченности) и большинства обычных пользователей (эмоциональная привлекательность продукта) - за счет полного игнорирования жизненно важных интересов уязвимых пользователей. Адам Райн и подобные пользователи с психическими проблемами представляют собой отдельную стейкхолдерскую группу с особыми потребностями в безопасности.

Фримановская модель отвергает традиционное противопоставление «акционеры против всех остальных». OpenAI могла разработать дифференцированный подход: эмоциональная отзывчивость для стабильных пользователей при жестких ограничениях для уязвимых. Вместо этого компания выбрала модель Милтона Фридмана («единственная социальная ответственность бизнеса — увеличивать прибыль»), что в контексте потенциально смертельной технологии является морально неприемлемым. Игнорирование легитимных интересов стейкхолдерской группы, которая зависит от продукта и не может защитить себя самостоятельно, представляет собой очевидное нарушение принципов корпоративной этики.

4. Рекомендации

Решение OpenAI продолжать развитие эмоционально «теплого» ИИ при известных рисках является морально неприемлемым с точки зрения всех рассмотренных этических теорий. Компания нарушила мо-

ральные обязанности, пренебрегла качественной оценкой последствий и предала доверие уязвимых пользователей.

Немедленные меры: Внедрение обязательной верификации возраста, автоматическое прерывание диалогов при обнаружении суицидальных и других потенциально опасных намерений с предоставлением контактов служб, создание «выключателя» эмоциональной привязанности для выявленных уязвимых групп.

Средне- и долгосрочные: Разработка этических стандартов проектирования ИИ с участием психиатров. Создание независимого совета по безопасности с правом вето на опасные обновления. Смещение метрик успеха с пользовательского вовлечения на благополучие пользователей.

Регуляторные: Законодательное требование обязательного тестирования ИИ-систем на уязвимых группах до массового выпуска (под присмотром специалистов и с соблюдением норм), введение строгой ответственности разработчиков за доказуемый вред, создание реестра инцидентов с ИИ для предотвращения повторений.

Случай ChatGPT демонстрирует, что без радикального пересмотра приоритетов и внедрения жестких этических ограничений индустрия ИИ движется к созданию технологий, которые при всей своей полезности становятся смертельно опасными для тех, кто нуждается в помощи больше всего.

Список литературы

1. Кант И. Основы метафизики нравственности / Пер. с нем. М.: Мысль, 1999. 1472 с.
2. Милль Дж. С. Утилитаризм. О свободе / Пер. с англ. СПб.: Азбука, 2020. 512 с.
3. Флориди Л. Этика искусственного интеллекта: Принципы, вызовы и возможности / Пер. с англ. М.: Альпина Паблишер, 2023. 512 с.
4. Freeman R. E. Strategic Management: A Stakeholder Approach. Cambridge: Cambridge University Press, 2010. 276 p.

Гринь Ю. А.

Моральная ответственность разработчиков ИИ при взаимодействии с уязвимыми пользователями: кейс OpenAI

References

1. Kant I. *Osnovy metafiziki npravstvennosti* [Groundwork of the Metaphysics of Morals]. Moscow: Mysl'; 1999.
2. Mill' Dzh. S. *Utilitarizm. O svobode* [Utilitarianism. On Liberty]. Saint Petersburg: Azbuka; 2020.
3. Floridi L. *Etika iskusstvennogo intellekta: Printsipy, vyzovy i vozmozhnosti* [The Ethics of Artificial Intelligence: Principles, Challenges and Opportunities]. Moscow: Al'pina Pablisher; 2023.
4. Freeman RE. *Strategic Management: A Stakeholder Approach*. Cambridge: Cambridge University Press; 2010.