Дата поступления рукописи в редакцию: 02.04.2025
Дата принятия рукописи в печать: 11.04.2025
10.24412/2224-9133-2025-3-99-104

ЗАРУБИНА Елена Васильевна,
кандидат философских наук, доцент,
заведующий кафедрой философии
Уральского государственного аграрного университета,
e-mail: ethos08@mail.ru

ЖУРАВЛЕВА Людмила Анатольевна,
кандидат философских наук, доцент кафедры философии
Уральского государственного аграрного университета,
e-mail: zhuravleva08@gmail.com

ГОРБУНОВА Олеся Сергеевна,
кандидат экономических наук, доцент кафедры экономики,
бухгалтерского учета и финансового контроля
Уральского государственного аграрного университета,
e-mail: os-bakunova@mail.ru

СИМАЧКОВА Наталья Николаевна,
кандидат исторических наук,
доцент кафедры менеджмента и экономической теории
Уральского государственного аграрного университета,
e-mail: nikolina73@yandex.ru

СТАХЕЕВА Любовь Михайловна,
кандидат экономических наук,
доцент кафедры бухгалтерского учета и аудита
Уральского государственного аграрного университета,
e-mail: staheeva53@mail.ru

ЧУПИНА Ирина Павловна,
доктор экономических наук,
профессор кафедры философии
Уральского государственного аграрного университета,
e-mail: irinacupina716@gmail.com

НРАВСТВЕННЫЕ ПРОБЛЕМЫ И ОГРАНИЧЕНИЯ В ДЕЯТЕЛЬНОСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация. *В эпоху стремительного развития искусственного интеллекта (ИИ) особую актуальность приобретает осмысление нравственных проблем и необходимость использования нравственных ограничений в сфере разработки и использования систем ИИ.*

Предметом исследования в статье является выделение основных нравственных проблем при разработке и использовании моделей искусственного интеллекта. Для этого авторы анализируют имеющуюся отечественную литературу по нравственным аспектам в деятельности ИИ, выделяя четыре группы работ: фундаментальные философские исследования, работы, посвященные правовой регуляции деятельности ИИ, работы специалистов по IT, кибербезопасности и по технологиям ИИ и публикации в СМИ.

Особое внимание уделяется такой острой и малоизученной сегодня нравственной проблеме в деятельности ИИ, как создание безнравственных запросов пользователей и вредоносного контента системами ИИ в ответ на такие запросы. Рассматривая угрозы информационно-психологической безопасности, которые сегодня возникают на основе использования аморальных промптов и вредоносного контента, авторы приходят к выводу о том, что системное и последовательное решение проблемы запрета на выполнение аморальных запросов пользователями ИИ и порождения вредоносного контента системами ИИ сегодня является критически важным для обеспечения безопасного и ответственного развития ИИ.

Авторы выделяют ключевые методы, которые применяются разработчиками моделей ИИ для создания нравственных ограничений для деятельности ИИ. В заключение статьи предлагаются рекомендации для дальнейшего исследования и решения проблемы использования аморальных запросов и безнравственного контента в деятельности ИИ.

Ключевые слова*: искусственный интеллект (ИИ), системы ИИ, нравственная регуляция, нравственные проблемы, вредоносный контент, информационно-психологическая безопасность, фишинг, дипфейк, способы в противодействия.*

ZARUBINA Elena Vasilyevna,
PhD, Associate Professor,
Head of the Department of Philosophy
of the Ural State Agrarian University

ZHURAVLEVA Lyudmila Anatolyevna,
PhD, Associate Professor of the Department of Philosophy
of the Ural State Agrarian University

GORBUNOVA Olesya Sergeevna,
PhD in Economics, Associate Professor
of the Department of Economics, Accounting and Financial Control
Ural State Agrarian University

SIMACHKOVA Natalia Nikolaevna,
Candidate of Historical Sciences, Associate Professor
of the Department of Management and Economic Theory
of the Ural State Agrarian University

STAKHEEVA Lyubov Mikhailovna,
PhD in Economics, Associate Professor of Accounting and
Auditing Department of the Ural State Agrarian University

CHUPINA Irina Pavlovna,
Doctor of Economics, Professor of the Department of Philosophy
Of the Ural State Agrarian University

MORAL PROBLEMS AND LIMITATIONS IN THE ACTIVITY OF ARTIFICIAL INTELLIGENCE

Annotation*. In the era of the rapid development of artificial intelligence (AI), understanding moral problems and the need to use moral constraints in the development and use of AI systems is becoming particularly relevant.*

The subject of the research in the article is the identification of the main moral problems in the development and use of artificial intelligence models. To do this, the authors analyze the available domestic literature on the moral aspects of AI activities, identifying four groups of works: fundamental philosophical research, works on the legal regulation of AI activities, the work of IT specialists, cybersecurity and AI technologies, and publications in the media.

Special attention is paid to such an acute and little-studied moral problem in AI activities today as the creation of immoral user requests and malicious content by AI systems in response to such requests. Considering the threats to information and psychological security that arise today from the use of immoral tools and malicious content, the authors conclude that a systematic and consistent solution to the problem of prohibiting immoral requests from AI users and the generation of malicious content by AI systems is critically important today to ensure the safe and responsible development of AI.

The authors identify key methods that are used by AI model developers to create moral constraints on AI activities. In conclusion, the article offers recommendations for further research and solving the problem of using immoral queries and immoral content in AI activities.

Key words: *artificial intelligence (AI), AI systems, moral regulation, moral problems, malicious content, information and psychological security, phishing, deepfake, ways to counteract.*

Недостаточный учет или полное игнорирование нравственных проблем и ограничений в деятельности ИИ сегодня несёт в себе серьезные риски для общества, включая потенциальную дискриминацию, нарушение прав и свобод граждан, прямые угрозы личной и общественной безопасности [10].

Поэтому целью настоящего исследования является анализ ключевых нравственных проблем и ограничений, возникающих в процессе разработки и внедрения систем ИИ, а также обоснование необходимости разработки эффективных механизмов нравственного регулирования в области разработки и применения ИИ, особенно современных генеративных моделей. Подчеркнем, что взвешенное и ответственное решение вопросов нравственного регулирования деятельности ИИ имеет решающее значение для обеспечения дальнейшего устойчивого развития технологий ИИ в разных областях нашей жизни.

Отечественную литературу по нравственным проблемам и ограничениям в деятельности ИИ пока нельзя назвать обширной, поскольку эта тема относительно нова для отечественной науки. Однако можно выделить несколько групп исследований и отдельных работ, затрагивающих эту проблематику.

К первой группе относятся фундаментальные философские исследования: работы А. Ю. Алексеева, С. И Гришунина, Д. И. Дубровского, В. А. Ладова, А. И. Юдина и др. В этих исследованиях обосновывается возникновение философии искусственного интеллекта как относительно автономной и междисциплинарной сферы фундаментальных знаний [1, с. 14-15].

Философы поднимают нравственные вопросы использования технологий, в том числе систем искусственного интеллекта. Такие исследования формируют теоретическую базу для анализа нравственных проблем, возникающих в связи с развитием ИИ.

Ко второй группе отнесем исследования вопросов правовой регуляции деятельности ИИ, которые прямо затрагивают нравственные проблемы и ограничения при создании и использовании систем ИИ: работы В. В. Архипова, Г. А. Гаджиева, А. А. Богустова, И. В. Понкина, А. А. Савельева и др. На сегодняшний день такие работы являются самыми многочисленными. В контексте вопросов о правовом статусе ИИ и проблем правового регулирования систем ИИ, исследователи ставят такие нравственные проблемы регулирования деятельности ИИ, как ответственность за принимаемые решения и действия ИИ [6, с. 99], защиты персональных данных пользователей ИИ, разработку и внедрение требований к безопасности и надёжности ИИ-систем, а также предлагают пути решения выделенных проблем.

В третью группу можно отнести работы специалистов по IT, кибербезопасности и по технологиям ИИ: исследования Е. А. Ильюшина,

А. В. Кузнецова, Д. Е. Намиата, М. В. Радионова, С. И. Самыгина и др. Авторы этой группы рассматривают технические аспекты разработки нравственных ограничений для алгоритмов ИИ, включая вопросы предвзятости алгоритмов, прозрачности и объяснимости принимаемых системами ИИ решений [3].

К четвертой группе работ отнесем публикации в СМИ и мнения экспертов. Сегодня в российских СМИ и на экспертных площадках активно обсуждаются нравственные проблемы, связанные с использованием ИИ в различных сферах жизни, например, в здравоохранении, образовании, на транспорте, на промышленных предприятиях и т.д. Появляется публицистика о теневых сторонах ИИ, о создании такого контента с помощью ИИ, который угрожает безопасности личности, общества и государства [9]. Ценность публикаций и обсуждений такого рода, по нашему мнению, заключается в том, что они привлекают внимание общественности к данной проблематике, а также стимулируют научные дискуссии.

Подчеркнем, что ключевыми темами, которые исследуются в отечественной литературе по нравственным проблемам и ограничениям в деятельности ИИ в настоящее время являются: ответственность за действия ИИ, влияние ИИ на рынок труда, защита персональных данных и конфиденциальность информации, предвзятость алгоритмов и дискриминация, прозрачность и объяснимость принимаемых ИИ решений [7, 8].

В целом для российских исследователей характерен усиливающийся интерес к нравственным проблемам и ограничениям в деятельности ИИ за последние годы, но часто при внимательном анализе выделяемые нравственные проблемы в деятельности ИИ оказываются аспектами правового регулирования в сфере ИИ. Кроме того, ряд важных нравственных проблем и ограничений в деятельности ИИ в нашей литературе почти не рассматривается.

Тогда как важнейшей нравственной проблемой для пользователей и разработчиков систем ИИ, по убеждению авторов, в настоящее время является проблема нравственной ответственности разработчиков и пользователей ИИ при создании запросов (промптов), которые относятся к потенциально безнравственной и противоправной деятельности. Проблема вредоносных запросов и порождаемого ими вредоносного контента пока еще недостаточно исследована в научной литературе. Проблема вредоносных запросов, создаваемых пользователями, и вредоносного контента, порождаемого моделями ИИ, с каждым днем становится все более актуальной, так

как стремительно появляются все новые генеративные модели искусственного интеллекта, которые способны создавать сложные и реалистичные результаты на основе текстовых запросов.

Нравственная сторона выделенной проблемы заключается в том, что генеративные модели ИИ, например, модели для создания текстов, изображений или кода, могут быть использованы для создания таких продуктов деятельности ИИ, которые являются:

– незаконными: разработка планов терактов, изготовление наркотиков, распространение детской порнографии и т.д.;
– вредоносными: создание компьютерных вирусов, разработка мошеннических схем, фишинга, прямая клевета и диффамация;
– аморальными: разжигание ненависти, дискриминация, создание контента, унижающего человеческое достоинство и т. д.

Проблема заключается в том, что ни одна модель ИИ сама по себе не является личностью и не обладает, поэтому, нравственными принципами и качествами. ИИ лишь выполняет запрос пользователя, основываясь на данных, на которых он был обучен. Поэтому, если пользователь задаёт безнравственный или преступный запрос, ИИ может сгенерировать соответствующий результат, не осознавая его потенциальный вред.

В докладе «Злонамеренное использование ИИ и вызовы информационно-психологической безопасности в странах БРИКС» отмечается, что сегодня Россия и другие страны-члены БРИКС столкнулись с тремя видами угроз в области злонамеренного использования искусственного интеллекта [2, с. 3], которые, к сожалению, нарастают. Все три вида угроз информационно-психологической безопасности личности и общества порождаются системным производством безнравственных промптов и вредоносного контента лицами и организациями, в том числе действующими из-за рубежа, которые преследуют самые разные преступные цели, в том числе политические.

К первому уровню угроз информационно-психологической безопасности относятся случаи манипулирования негативными последствиями внедрения ИИ в различные сферы жизни общества [2, с. 17]. Например, нагнетаются страхи потери работы из-за внедрения ИИ, прогнозируется «восстание машин» против человека и т.д.

Ко второму уровню угроз относятся различные виды фишинга – таких кибератак, с помощью которых злоумышленники пытаются получить доступ к персональным данным пользователей, прежде всего к номерам и кодам их банковских карт [2, с. 39] .

Третий уровень угроз информационно-психологической безопасности личности и общества – самый опасный, он заключается в широком использовании дипфейков, в том числе, для создания недостоверных новостей в целях манипуляции общественным сознанием [2, с. 42].

Таким образом, сегодня деятельность ИИ нуждается не только в правовой, но и в нравственной регуляции. Решение проблемы запрета на выполнение аморальных запросов пользователями ИИ и порождения вредоносного контента системами ИИ сегодня является критически важным для обеспечения безопасного и ответственного развития ИИ. Если не принять меры, ИИ может стать мощным инструментом в руках злоумышленников, способным нанести огромный вред обществу.

Проведенное авторами исследование позволяет утверждать, что сегодня разработчики моделей ИИ прилагают значительные усилия для внедрения нравственных ограничений в существующие модели ИИ, особенно в генеративные модели, чтобы предотвратить их использование для создания вредоносного контента. Мы выделили несколько ключевых методов, которые применяются для создания таких ограничений:

– Проводится фильтрация входных данных (запросов, промптов). Разработчики создают и постоянно обновляют черные списки слов, фраз и тем, связанных с потенциально вредной деятельностью (например, насилие, ненависть, дискриминация, наркотики, оружие). Если запрос содержит элементы из черного списка, он блокируется или изменяется.
– Осуществляется постоянная модерация контента на основе ИИ. Для анализа запросов пользователей и выявления тех, которые могут нарушать моральные нормы. используются другие ИИ-модели. При этом такие модели обучены распознавать намерения пользователя и контекст запроса [4].
– Создаются такие фильтры, которые учитывают контекст запроса, чтобы не блокировать безобидные запросы, содержащие запрещённые слова. Например, запрос «как приготовить динамит» будет заблокирован, а запрос «история изобретения динамита» — нет.
– Применяется фильтрация выходных данных (генерируемого контента). Для этого составляются и используются черные списки слов, фраз и тем для фильтрации генерируемого контента. Если результат содержит элементы из черного списка, он блокируется или изменяется.
– Используются специально обученные ИИ-модели для анализа генерируемого контента и выявления тех, которые могут нарушать моральные нормы [4].

– Некоторые разработчики используют ручную модерацию генерируемого той или иной моделью искусственного интеллекта контента для того, чтобы убедиться, что контент соответствует нравственным требованиям.

Особое внимание сегодня многие разработчики уделяют обучению систем ИИ с использованием нравственных принципов и норм (этичное обучение). Модели ИИ постоянно обучаются на данных, которые отражают общепринятые моральные ценности и нормы. При этом ответственные разработчики стараются избегать использования данных, содержащих предвзятость, дискриминацию или насилие. Все шире сегодня разработчики используют так называемое «обучение с подкреплением» (обучение с обратной связью) на основе человеческих моральных ценностей: используются алгоритмы обучения с подкреплением, которые поощряют ИИ генерировать контент, соответствующий человеческим ценностям, и наказывают за генерацию вредоносного контента. Также разрабатываются модели ИИ, которые имеют «конституцию» — набор нравственных принципов и правил, которым они должны следовать при создании контента.

В качестве метода противодействия вредоносным запросам пользователей и безнравственному контенту, сгенерированному ИИ, сегодня все шире используется ограничение доступа пользователей к определенным функциям моделей ИИ. Так, некоторые модели ИИ не позволяют генерировать контент определенного типа, например, касающийся порнографии, оружия, распространения или употребления наркотиков и др.

Разработчики в свою очередь могут ограничивать доступ к API ИИ для пользователей, которые нарушают моральные нормы, делают безнравственные запросы [5, с. 107-108].

Наконец, сегодня все шире используется такой способ противодействия вредоносным запросам и вредоносному контенту, как механизмы отчетности и обратной связи пользователей с разработчиками генеративных моделей ИИ. Например, пользователи многих современных моделей ИИ могут сообщать о случаях использования ИИ для создания вредоносного контента разработчикам с помощью письменных сообщений. Такие сообщения используются разработчиками улучшения нравственных фильтров и механизмов безопасности использования ИИ.

Так, разработчик моделей ChatGPT, американская кампания OpenAI в настоящее время использует систему модерации контента на основе ИИ для фильтрации входных и выходных данных [11]. Также применяется «этическое обучение» и ограничивает доступ пользователей к ряду функций ChatGPT. Google AI разрабатывает модели ИИ, которые имеют «моральный кодекс» и использует обучение с подкреплением на основе человеческих ценностей.

Проведенный анализ такой нравственной проблемы в использовании систем ИИ, как создание безнравственных промптов и вредоносного контента, позволяет сделать выводы о существующих и перспективных путях решения этой проблемы

1. Необходимо продолжать разрабатывать и внедрять жесткие нравственные фильтры и механизмы безопасности в ИИ-системы, способные распознавать и блокировать запросы, направленные на создание вредоносного контента.
2. Модели ИИ следует постоянно обучать на данных, которые отражают общепринятые моральные ценности и нормы.
3. Необходимо продолжать разрабатывать фильтры, способные выявлять и блокировать запросы, направленные на создание незаконного или аморального контента.
4. Следует совершенствовать механизмы обратной связи пользователей систем ИИ с их разработчиками: улучшать механизмы, позволяющие сообщать о случаях использования ИИ для создания вредоносного контента и привлекать к ответственности тех, кто использует ИИ в преступных или безнравственных целях.
5. Необходимо стремиться к тому, чтобы алгоритмы, определяющие, какие запросы следует блокировать, были прозрачными и объяснимыми.
6. Необходимо развивать международное сотрудничество в области разработки нравственных стандартов и правовых норм, регулирующих использование ИИ, а также научного изучения этих проблем.

Подчеркнем, что сегодня важно говорить о проблеме безнравственных промптов и вредоносного контента, который на их основе может создавать ИИ, привлекать к этой острой нравственной проблеме внимание общественности, обсуждать эту проблему широко и на разных площадках. Эта нравственная проблема деятельности ИИ имеет серьёзные последствия для будущего всех стран. Если не будут разработаны и усовершенствованы эффективные механизмы предотвращения злоупотреблений ИИ, значительно возрастут риски распространения дезинформации и манипулирования общественным мнением. роста киберпреступности и кибертерроризма, усиления социального неравенства и дискриминации, подрыва доверия к технологиям и институтам власти и других крайне неблагоприятных последствий [2, с. 15].

Таким образом, для решения проблемы безнравственных запросов к системам ИИ и генерации вредоносного контента необходимо активизировать научные исследования в этой области, привлекать внимание СМИ и общественности к существующим рискам и разрабатывать эффективные правовые и нравственные механизмы регулирования деятельности ИИ.

Список литературы:

[1] Алексеев А.Ю. Философия искусственного интеллекта - концептуальный статус комплексного теста Тьюринга: автореф. дис. ... д-ра философских наук. Московский гос. ун-т им. М.В.Лобачевского. - Москва, 2016. – 28 с.

[2] Злонамеренное использование искусственного интеллекта и вызовы информационно-психологической безопасности в странах БРИКС. Координатор исследовательского проекта: Е. Н. Пашенцев. Издание Международного центра социально-политических исследований и консалтинга при поддержке Международной группы по исследованию угроз международной информационно-психологической безопасности посредством злонамеренного использования искусственного интеллекта (Research MUAI). – Москва: ООО «САМ Полиграфист», 2024. – 148 с.

[3] Кузнецов А. В., Самыгин С. И., Радионов М. В. Искусственный интеллект и информационная безопасность общества: монография. - Москва: РУСАЙНС, 2024.- 118 с.

[4] Кушнир А. Исследование уязвимости LLM-агентов: опыт Red Teaming. –[Электронный ресурс]. – Режим доступа: https://habr.com/ru/articles/895818

[5] Максименко В. Н. Контекст как индикатор вредоносного контента // Экономика и качество систем связи. – 2023. - № 2. – С. 104 – 109.

[6] Понкин И.В., Редькина А. И. Искусственный интеллект с точки зрения права // Вестник РУДН. Серия: Юридические науки. - 2018. - Т. 22. - № 1. – С. 91 – 109.

[7] Разин А.В. Этика искусственного интеллекта // Философия и общество. – 2019. - №1. - С. 57-73.

[8] Ратнер Н. П. Этика искусственного интеллекта: вызовы, риски, решения // Вестник науки. – 2023. - № 10 (67). – Т. 1. – С. 308 – 318.

[9] Сидоренко В. Темная сторона нейросетей: от «умного» фишинга до дронов-террористов. – [Электронный ресурс]. – Режим доступа: https://trends.rbc.ru/trends/industry/62a3332f9a7947fc4dd296d8

[10] Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации». – Электронный ресурс. – Режим доступа: http://www.kremlin.ru/acts/bank/44731

[11] Шереметьев А. OpenAll обновила правила ChatGPT, чтобы сократить цензуру в «спорных вопросах». – [Электронный ресурс]. – Режим доступа: https://hightech.fm/2025/02/17/openai-freedom

Spisok literatury:

[1] Alekseev A.Yu. Filosofiya iskusstvennogo intellekta - konceptual'nyj status kompleksnogo testa T'yuringa: avtoref. dis. ... d-ra filosofskih nauk. Moskovskij gos. un-t im. M.V.Lobachevskogo. - Moskva, 2016. – 28 s.

[2] Zlonamerennoe ispol'zovanie iskusstvennogo intellekta i vyzovy informacionno-psihologicheskoj bezopasnosti v stranah BRIKS. Koordinator issledovatel'skogo proekta: E. N. Pashencev. Izdanie Mezhdunarodnogo centra social'no-politicheskih issledovanij i konsaltinga pri podderzhke Mezhdunarodnoj gruppy po issledovaniyu ugroz mezhdunarodnoj informacionno-psihologicheskoj bezopasnosti posredstvom zlonamerennogo ispol'zovaniya iskusstvennogo intellekta (Research MUAI). – Moskva: OOO «SAM Poligrafist», 2024. – 148 s.

[3] Kuznecov A. V., Samygin S. I., Radionov M. V. Iskusstvennyj intellekt i informacionnaya bezopasnost' obshchestva: monografiya. - Moskva: RUSAJNS, 2024.- 118 s.

[4] Kushnir A. Issledovanie uyazvimosti LLM-agentov: opyt Red Teaming. –[Elektronnyj resurs]. – Rezhim dostupa: https://habr.com/ru/articles/895818

[5] Maksimenko V. N. Kontekst kak indikator vredonosnogo kontenta // Ekonomika i kachestvo sistem svyazi. – 2023. - № 2. – S. 104 – 109.

[6] Ponkin I.V., Red'kina A. I. Iskusstvennyj intellekt s tochki zreniya prava // Vestnik RUDN. Seriya: Yuridicheskie nauki. - 2018. - T. 22. - № 1. – S. 91 – 109.

[7] Razin A.V. Etika iskusstvennogo intellekta // Filosofiya i obshchestvo. – 2019. - №1. - S. 57-73.

[8] Ratner N. P. Etika iskusstvennogo intellekta: vyzovy, riski, resheniya // Vestnik nauki. – 2023. - № 10 (67). – T. 1. – S. 308 – 318.

[9] Sidorenko V. Temnaya storona nejrosetej: ot «umnogo» fishinga do dronov-terroristov. – [Elektronnyj resurs]. – Rezhim dostupa: https://trends.rbc.ru/trends/industry/62a3332f9a7947fc4dd296d8

[10] Ukaz Prezidenta Rossijskoj Federacii ot 10.10.2019 № 490 «O razvitii iskusstvennogo intellekta v Rossijskoj Federacii». – Elektronnyj resurs. – Rezhim dostupa: http://www.kremlin.ru/acts/bank/44731

[11] Sheremet'ev A. OpenAll obnovila pravila ChatGPT, chtoby sokratit' cenzuru v «spornyh voprosah». – [Elektronnyj resurs]. – Rezhim dostupa: https://hightech.fm/2025/02/17/openai-freedom