ЭТИКА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПРОБЛЕМА МОРАЛЬНОЙ ОТВЕТСТВЕННОСТИ ПРИ ДЕЛЕГИРОВАНИИ АВТОНОМНЫХ РЕШЕНИЙ

Д.А. Бощенко, *студент*
Научный руководитель: *Т.П. Машихина, канд. пед. наук, доцент*
Волгоградский государственный университет
(Россия, г. Волгоград)

DOI:10.24412/2500-1000-2026-5-2-63-67

Аннотация. *В статье анализируются моральные дилеммы, возникающие при передаче решений всё более автономным системам искусственного интеллекта в таких областях, как беспилотный транспорт, медицинская диагностика и рекомендательные алгоритмы. Рассматривается концепция «разрыва ответственности» (responsibility gap), границы моральной агентности ИИ, проблема прозрачности алгоритмов и предвзятости данных. Сопоставляются позиции технологического оптимизма Л. Флориди и критических подходов Н. Бострома и Р. Спэрроу. В заключение обсуждаются возможные принципы распределения ответственности между разработчиком, пользователем и самой системой.*

Ключевые слова: *этика искусственного интеллекта; разрыв ответственности; моральная ответственность; автономные системы; моральная агентность; распределённая мораль; прозрачность алгоритмов.*

Стремительное внедрение систем искусственного интеллекта (ИИ) во все сферы человеческой деятельности ставит перед философской этикой фундаментальные вопросы, не имеющие аналогов в истории технологических революций. В отличие от традиционных орудий труда, которые являются лишь пассивным продолжением воли человека, современные ИИ-системы демонстрируют возрастающую степень автономии: они способны принимать решения без прямого вмешательства оператора, адаптироваться к изменяющимся условиям и даже вырабатывать новые стратегии поведения. Эта автономия одновременно порождает глубокую моральную проблему: кто несёт ответственность за последствия действий, совершённых автономной системой? Традиционные модели ответственности, основанные на чёткой причинно-следственной связи между действием агента и его результатом, оказываются неприменимы к ситуациям, где решение формируется внутри «чёрного ящика» алгоритма, недоступного для полного понимания и контроля со стороны человека [1]. Данная статья посвящена всестороннему анализу проблемы моральной ответственности при делегировании решений системам ИИ, получившей в современной литературе название «разрыв ответственности» (responsibility gap).

Термин «разрыв ответственности» был введён в философский дискурс Андреасом Маттиасом в 2004 году, который указал на парадоксальную ситуацию: по мере того как системы ИИ становятся всё более автономными и сложными, традиционные механизмы возложения моральной ответственности перестают работать [2]. Классическая модель ответственности предполагает выполнение двух условий – контрольного (control condition) и эпистемического (epistemic condition). Согласно первому, агент должен иметь возможность контролировать свои действия; согласно второму, он должен знать или иметь возможность знать, какие последствия эти действия повлекут. Разработчик нейросети не может предвидеть все возможные сценарии её поведения в реальной среде, а пользователь часто не понимает внутренней логики принятия решения из-за «проблемы чёрного ящика». Сам же алгоритм, не обладая сознанием, свободой воли и способностью к рефлексии, не может считаться моральным агентом в классическом понимании. В результате возникает ситуация, когда вред причинён, но формально некому вменить его в вину – ни один из участников взаимодействия не удовлетворяет в полной мере условиям моральной ответственности.

Попытки преодолеть этот разрыв породили несколько конкурирующих теоретических стратегий: отрицание существования разрыва, консервативные попытки его «закрыть» за счёт распределения ответственности между существующими человеческими агентами и радикальные предложения по расширению понятия моральной агентности на сами ИИ-системы. Одной из наиболее влиятельных попыток «закрыть» разрыв является концепция распределённой морали (distributed morality), предложенная Лучано Флориди. Флориди утверждает, что в условиях сложных социотехнических систем моральная ответственность не локализована в одном агенте, а распределена между множеством участников – разработчиками, пользователями, владельцами систем, регуляторами [3]. С этой точки зрения, разрыв ответственности оказывается иллюзией, порождённой устаревшим антропоцентрическим представлением о морали как об индивидуальном деянии. Однако подход Флориди сталкивается с серьёзной критикой: если ответственность распределена повсюду, то на практике она может оказаться нигде, поскольку размывание границ ведёт к её диффузии и, в конечном счёте, к безответственности. В этой связи перспективным представляется различение разных уровней ответственности – например, генеративной (кто создал условия для вреда) и апробационной (кто одобрил или использовал систему).

Противоположную позицию занимают те, кто, подобно Максимилиану Киенеру, утверждают, что разрыва ответственности как такового не существует, а есть, напротив, её изобилие (responsibility abundance) [4]. Киенер доказывает, что вокруг любой автономной системы существует множество агентов, которые в той или иной степени удовлетворяют условиям моральной ответственности, – от программистов, закладывающих архитектуру системы, до менеджеров, принимающих решение о её развёртывании, и конечных пользователей. Проблема заключается не в отсутствии ответственного, а в сложности справедливого распределения ответственности между множеством потенциальных кандидатов. Эта позиция, однако, вызывает возражения в тех случаях, где решения принимаются системой в режиме реального времени без участия человека, а цепочка причинности оказывается настолько длинной и опосредованной, что ни один из разработчиков не мог предвидеть конкретный исход. Как подчёркивают исследователи, ответственность не бесконечно делима – она имеет качественную, а не только количественную размерность [5].

Проблема моральной ответственности ИИ неразрывно связана с такими эпистемологическими вызовами, как непрозрачность (opacity) и предвзятость (bias) алгоритмов. Если решение, принятое ИИ, невозможно реконструировать на уровне, доступном человеческому пониманию (например, из-за огромного количества параметров глубокой нейронной сети), то становится невозможным и установление причинной связи между этим решением и действиями разработчиков. Современная литература различает три типа непрозрачности: непрозрачность как коммерческую тайну, непрозрачность как техническую неграмотность пользователя и фундаментальную эпистемическую непрозрачность, вытекающую из самой природы глубокого обучения. Именно последняя создаёт непреодолимые препятствия для классических моделей ответственности: если даже создатель алгоритма не может объяснить, почему было принято то или иное решение, то как можно вменить ему это решение в вину? Непрозрачность алгоритмов выступает не как техническая проблема, а как эпистемическая граница, меняющая саму структуру агентности и ответственности [7]. Возможно, вместо того чтобы требовать полного понимания причинно-следственных связей, мы должны разработать процедурные механизмы ответственности, основанные на контроле процесса разработки и внедрения.

Предвзятость алгоритмов, которая является следствием предвзятости обучающих данных или некорректно сформулированных целевых функций, создаёт дополнительное измерение разрыва ответственности. Если система отказывает в кредите соискателям определённой расы или пола, кто отвечает за эту дискриминацию? Разработчик, собравший предвзятые данные? Компания, утвердившая использование этих данных? Регулятор, не установивший стандарты? Или сама система, которая лишь математически оптимизирует заданный критерий? Опыт показывает, что в отсутствие чётких правил ответственности возникает то, что можно назвать «регрессом обвинения»:

каждый участник цепочки перекладывает вину на другого, и в конечном счёте никто не несёт реальной ответственности за причинённый вред. Этот феномен особенно ярко проявляется в контексте так называемой «проблемы многих рук» (problem of many hands), когда в создании и эксплуатации системы участвует такое большое количество людей и организаций, что установление индивидуальной причинной связи становится практически невозможным. Некоторые авторы, в частности Альбареда, доказывают, что разрыв ответственности в ИИ-этике сводим именно к проблеме многих рук и коллективной агентности, а не к какой-либо новой категории моральных феноменов [5].

Одним из наиболее ярких примеров обсуждаемой проблемы является сфера автономных боевых систем (летальных автономных роботов). Роберт Спэрроу в своей известной работе «Killer Robots» аргументирует, что использование таких систем принципиально неэтично именно из-за невозможности установить моральную ответственность за их действия. Если беспилотник совершает военное преступление, кто будет привлечён к ответственности? Командир, отдавший приказ о патрулировании? Программист, заложивший алгоритм распознавания целей? Производитель? Или сам беспилотник, которому невозможно инкриминировать виновное намерение? Спэрроу утверждает, что любая попытка возложить ответственность в этой цепочке наталкивается на непреодолимые препятствия, и поэтому единственным этически последовательным решением является запрет на разработку и применение полностью автономных летальных систем [6]. Этот аргумент, хотя и вызывает споры, демонстрирует, что проблема разрыва ответственности имеет не только академическое, но и судьбоносное практическое значение.

Рассмотрим проблему на более повседневном примере. Предположим, что рекомендательная система социальной сети, оптимизирующая время пребывания пользователя на платформе, начинает настойчиво предлагать пользователю депрессивный контент, что в конечном счёте приводит к ухудшению его психического здоровья. Кто несёт ответственность за этот вред? Разработчики алгоритма могут утверждать, что они лишь реализовали заказанную бизнесом оптимизационную функцию, и их код корректен. Менеджеры могут ссылаться на то, что они не обладали технической компетенцией для оценки долгосрочных психологических последствий. Сама платформа – юридическое лицо – может быть привлечена к гражданской ответственности, но это не затрагивает вопроса о моральной ответственности конкретных индивидов. Образуется классический разрыв: каждый выполнил свою работу, каждый следовал инструкциям, но в результате – реальный моральный вред, за который никто персонально не может быть привлечён к моральной ответственности [8].

Возможная стратегия выхода из этого тупика лежит в смещении акцента с ретроспективной ответственности (чьё конкретное действие привело к вреду) на проспективную ответственность (кто должен был предвидеть и предотвратить потенциальный вред). Вместо того чтобы спрашивать «кто виноват?» после того, как вред уже причинён, мы должны разработать нормативные стандарты для процесса разработки и внедрения систем ИИ, установив обязанности каждого участника жизненного цикла системы – от сбора данных до конечного применения. Этот подход, известный как «ответственность через процедуру» (procedural responsibility), уже применяется в регулировании финансового сектора. Ключевым элементом такой процедуры является требование о человеческом контроле (human oversight): все решения ИИ, имеющие значимые моральные последствия, должны быть в принципе проверяемы человеком и допускать возможность вмешательства. Если система устроена так, что такое вмешательство невозможно (например, в случае гиперскоростного трейдинга или автономного оружия), то она не должна применяться – это принципиальное ограничение на автономию ИИ [9].

Такой подход, однако, не решает всех проблем. Что делать с системами, которые обучаются в реальном времени и чьё поведение после развёртывания не может быть полностью предсказуемо? Как распределить ответственность между создателями базовой модели (например, большой языковой модели) и разработчиками, которые её дообучили на специфических данных и интегрировали в конкретный продукт? Вероятно, здесь требу-

ется многоуровневая модель ответственности: создатели базовой модели отвечают за общие архитектурные решения и безопасность «на уровне архитектуры»; дообучающие организации – за специфические риски, привнесённые на этапе тонкой настройки; интеграторы – за корректность внедрения и наличие механизмов контроля; пользователи – за соблюдение инструкций и границ применения. Такая многоуровневая модель является развитием идеи Флориди о распределённой морали, но с чёткими правилами разграничения уровней и видов ответственности [3].

Заключение

Проблема моральной ответственности при делегировании автономных решений системам ИИ не имеет простого и однозначного решения. Классические модели ответственности, выросшие из аристотелевской традиции и новоевропейского индивидуализма, оказываются неприменимы к ситуациям, где субъект действия отсутствует, причинно-следственные связи опосредованы непрозрачными алгоритмами, а масштаб возможного вреда несоизмерим с масштабом индивидуального контроля. Предложенные в литературе стратегии – от распределённой морали Флориди до радикального расширения понятия агентности – имеют как сильные, так и слабые стороны. Наиболее плодотворным представляется эволюционный путь, сочетающий аналитическое различение разных видов ответственности, разработку процедурных механизмов контроля и внедрение многоуровневых моделей распределения ответственности. Принципиальным остаётся тезис о том, что ответственность за действия ИИ, каким бы автономным он ни был, в конечном счёте лежит на человеке – разработчике, пользователе или владельце системы. Дальнейшие исследования должны быть направлены на разработку практических рекомендаций для законодателей, инженеров и пользователей, а также на философское осмысление возможных трансформаций самого понятия моральной ответственности в условиях цифровой эпохи.

Библиографический список

1. Müller V.C. Ethics of Artificial Intelligence and Robotics // The Stanford Encyclopedia of Philosophy / Ed. by E.N. Zalta. – 2020. – URL: https://plato.stanford.edu/archives/sum2020/entries/ethics-ai/.

2. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata // Ethics and Information Technology. – 2004. – Vol. 6, № 3. – P. 175-183.

3. Floridi L., Sanders J.W. On the Morality of Artificial Agents // Minds and Machines. – 2004. – Vol. 14, № 3. – P. 349-379.

4. Kiener M. AI and Responsibility: No Gap, but Abundance // Journal of Applied Philosophy. – 2025. – Vol. 42, № 1. – P. 357-374.

5. Llorca Albareda J. Uncovering the gap: challenging the agential nature of AI responsibility problems // AI and Ethics. – 2025. – P. 1-14.

6. Sparrow R. Killer Robots // Journal of Applied Philosophy. – 2007. – Vol. 24, № 1. – P. 62-77.

7. Goetze T.S. Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement // Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). – 2022. – URL: https://philarchive.org/rec/GOEMTG.

8. Плотников В.В., Косников М.С. Философия искусственного интеллекта и границы ответственности человека и алгоритма // Философия науки. – 2026. – № 1. – С. 115-131.

9. Nyholm S. Responsibility and Autonomous Technologies // The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence. – Cambridge University Press, 2025. – P. 63-78.

ETHICS OF ARTIFICIAL INTELLIGENCE: THE PROBLEM OF MORAL RESPONSIBILITY IN DELEGATING AUTONOMOUS DECISIONS

D.A. Boshchenko, *Student*
Supervisor: *T.P. Mashikhina, Candidate of Pedagogical Sciences, Associate Professor*
Volgograd State University
(Russia, Volgograd)

Abstract. *This article analyzes the moral dilemmas that arise when decisions are delegated to increasingly autonomous artificial intelligence systems in areas such as autonomous vehicles, medical diagnostics, and recommendation algorithms. It examines the concept of the "responsibility gap", the limits of AI moral agency, the problem of algorithmic opacity and data bias. The positions of technological optimism (Floridi) and critical approaches (Sparrow) are compared. The article concludes by discussing possible principles for distributing responsibility among developers, users, and the system itself.*

Keywords: *artificial intelligence ethics; responsibility gap; moral responsibility; autonomous systems; moral agency; distributed morality; algorithmic opacity.*