Человек, общество и государство в философской перспективе / Individual, Society, and State in Philosophical Perspective

Научная статья
DOI: 10.12731/3033-5981-2026-18-3-613
УДК 17.023.32

Original article
EDN: EWAPWF

Проблема моральной ответственности в контексте развития сильного искусственного интеллекта

С.М. Мальцева[1,2], А.Д. Грачева[1], М.Б. Ротанова[3]
[1]Мининский университет, Нижний Новгород, Российская Федерация
[2]НИПС - филиал ПривГУПС, Нижний Новгород, Российская Федерация
[3]Нижегородский государственный лингвистический университет им. Н.А. Добролюбова, Нижний Новгород, Российская Федерация

Аннотация

Обоснование. Стремительное развитие технологий машинного обучения и генеративных нейросетей обострило дискуссию о границах моральной субъектности и ответственности в ситуациях, где автономные системы принимают решения, затрагивающие жизнь и благополучие людей. Сильный интеллект – это предполагаемая техническая система, которая способны выполнять интеллектуальные задачи на уровне человека, то есть с пониманием и возможностью самообучения. В своем развитии он проходит несколько этапов: от контролируемого через уменьшение надзора до минимального к автономной и самодостаточной стадии. Вместе с этим меняется и мораль: от попытки встроить ее в алгоритмы к пересмотру сущности и оснований перед лицом агента, отличного или даже превосходящего человека. Классические этические теории, выстроенные вокруг свободной воли и рационального выбора, оказываются недостаточными для квалификации действий агентов, лишённых сознания в философском смысле.

Цель исследования – выявить концептуальные ограничения применения категории моральной ответственности к системам сильного искусственного интеллекта и предложить трехуровневую модель распределённой ответственности.

Гипотеза: моральная ответственность за действия автономных интеллектуальных систем не может быть локализована в одном субъекте; она распределяется между разработчиками, операторами и институциональной средой.

Материалы и методы. Методологической основой исследования выступает концептуальный анализ, позволяющий реконструировать содержание философ-

ских категорий (моральная ответственность, субъектность, интенциональность) и оценить границы их применения к новым объектам. Параллельно используется сравнительный метод, посредством которого сопоставляются позиции представителей различных этических традиций.

Новизна исследования. Предложена интерпретация проблемы ответственности через призму концепции «вакуума ответственности» А. Маттиаса, дополненной положениями этики заботы и институциональной этики, что позволяет преодолеть индивидуалистическую редукцию.

Результаты. Показано, что ни модель добровольного действия Аристотеля, ни критерий автономии воли Канта не применимы к искусственным агентам напрямую. Аргумент Сёрла о невозможности подлинной интенциональности у вычислительных систем остаётся состоятельным, однако не снимает вопроса об ответственности за последствия. Обосновано, что адекватной рамкой выступает распределённая ответственность, включающая три уровня: техническую подотчётность разработчика, операционную ответственность пользователя и регуляторную ответственность институтов.

Ключевые слова: моральная ответственность; сильный искусственный интеллект; моральная субъектность; вакуум ответственности; этика технологий; автономные системы; распределённая ответственность; философия техники

Для цитирования. Мальцева, С. М., Грачева, А. Д., & Ротанова, М. Б. (2026). Проблема моральной ответственности в контексте развития сильного искусственного интеллекта. *Russian Social and Humanitarian Studies / Российские социогуманитарные исследования*, *18*(3), 84-100. https://doi.org/10.12731/3033-5981-2026-18-3-613

The problem of moral responsibility in the context of strong artificial intelligence development

S.M. Maltseva[1,2], A.D. Gracheva[1], M.B. Rotanova[3]
[1]Minin University, Nizhny Novgorod, Russian Federation
[2]Nizhny Novgorod Institute of Railway Engineering - branch of the Volga State University of Railway Engineering, Nizhny Novgorod, Russian Federation
[3]Linguistic University of Nizhny Novgorod, Nizhny Novgorod, Russian Federation

Abstract

Background. The rapid development of machine learning technologies and generative neural networks has intensified the debate about the boundaries of moral subjectivity and responsibility in situations where autonomous systems make decisions affecting people's lives and well-being. Strong intelligence is an assumed technical system that is capable of performing intellectual tasks at the human level, that is, with

understanding and the possibility of self–learning. It goes through several stages in its development: from controlled through reduced supervision to minimal to an autonomous and self-sufficient stage. At the same time, morality is also changing: from an attempt to embed it in algorithms to a revision of the essence and grounds in the face of an agent who is different or even superior to a person. Classical ethical theories, built around free will and rational choice, are insufficient to qualify the actions of agents devoid of consciousness in the philosophical sense.

Purpose – to identify the conceptual limitations of applying the category of moral responsibility to systems of strong artificial intelligence and to propose a three-level model of distributed responsibility.

Hypothesis: moral responsibility for the actions of autonomous intelligent systems cannot be localized in one entity; it is distributed among developers, operators and the institutional environment.

Materials and methods. The methodological basis of the research is a conceptual analysis that allows reconstructing the content of philosophical categories (moral responsibility, subjectivity, intentionality) and assessing the boundaries of their application to new objects. In parallel, a comparative method is used, through which the positions of representatives of various ethical traditions and the **novelty of the research** are compared. The interpretation of the problem of responsibility through the prism of the concept of the «vacuum of responsibility» is proposed. Mattias, supplemented by the provisions of ethics of care and institutional ethics, which makes it possible to overcome individualistic reduction.

Results. It is shown that neither Aristotle's model of voluntary action nor Kant's criterion of will autonomy are directly applicable to artificial agents. Searle's argument about the impossibility of genuine intentionality in computing systems remains valid, but it does not remove the question of responsibility for the consequences. It is proved that an adequate framework is distributed responsibility, which includes three levels: technical accountability of the developer, operational responsibility of the user and regulatory responsibility of institutions.

Keywords: moral responsibility; strong artificial intelligence; moral agency; responsibility gap; ethics of technology; autonomous systems; distributed responsibility; philosophy of technology

For citation. Maltseva, S. M., Gracheva, A. D., & Rotanova, M. B. (2026). The problem of moral responsibility in the context of strong artificial intelligence development. *Russian Social and Humanitarian Studies*, *18*(3), 84-100. https://doi.org/10.12731/3033-5981-2026-18-3-613

Введение

Глобальные изменения последних лет, связанные с всеобщей информатизацией, а теперь уже и повсеместным внедрением искус-

ственного интеллекта, обострили социальные и этические противоречия. Сегодня мы уже можем не только предполагать, а анализировать накопившиеся последствия этих процессов: отчужденность, усталость, одиночество, социальный разрыв, несправедливость, проблемы коммуникации [2; 9; 10].

Вопрос о моральной ответственности принадлежит к числу старейших в западной философской традиции. От Аристотеля до И. Канта, от утилитаризма Дж. С. Милля до деонтологии К. Корсгаард, мыслители разных эпох искали критерии, позволяющие вменять субъекту ответственность за совершённое действие [1; 5; 15; 17]. Эти критерии, при всех различиях, сходились в одном: ответственность предполагает агента, наделённого сознанием, свободой воли и способностью к рефлексии. Появление автономных интеллектуальных систем, способных принимать решения без непосредственного участия человека, ставит под сомнение универсальность этой предпосылки.

Актуальность проблемы определяется не абстрактной перспективой создания «сильного» искусственного интеллекта, обладающего сознанием и самоосознанием, а вполне практическими ситуациями. Беспилотные автомобили уже сегодня вынуждены «выбирать» между столкновениями с разными объектами; системы медицинской диагностики формируют рекомендации, определяющие тактику лечения; алгоритмы кредитного скоринга влияют на доступ граждан к финансовым ресурсам. Когда решение принято машиной, возникает закономерный вопрос: кто несёт ответственность за его последствия? Классические модели атрибуции ответственности не дают однозначного ответа, поскольку строились для субъектов, обладающих интенциональностью.

Связь между сильным искусственным интеллектом, искусственными агентами и моральной ответственностью – это фундаментальный вызов человечеству, ставящий под сомнение всю его этическую систему. Традиционная модель ответственности была рассчитана только на человека, единственно обладающего свободой воли и сознанием. Сильный искусственный интеллект уже превзошел че-

ловека по когнитивным способностям и грозит превзойти в автономии. Его действия все более сложны и непредсказуемы. Даже если мы не признаем наличие у него сознания, это не исключает наличие у него адаптивности, автономности и интерактивности, делая его со-субъектом, способным производить в мире значимые эффекты.

Анализ литературы показывает, что проблема моральной ответственности в контексте развития искусственного интеллекта разрабатывается преимущественно в зарубежной традиции. Л. Флориди и Дж. Сандерс предложили концепцию «моральной подотчётности» артефактов, не наделённых субъектностью [13; 14]. А. Маттиас ввёл вызвавшее среди ученых резонанс понятие «вакуума ответственности», описывающего ситуацию, когда ни разработчик, ни пользователь не могут в полной мере предвидеть поведение обучающейся системы. [12; 16; 20; 23]. Н. Бостром проанализировал риски, связанные с появлением сверхинтеллекта, и указал на проблему согласования целей [3]. В отечественной философии к данной теме обращались Е.А. Никитина, В.А. Лекторский, Б.Г. Юдин, исследующие взаимосвязь технологического прогресса и трансформации субъектности [6; 7; 11].

Вместе с тем ряд аспектов остаётся дискуссионным. Не прояснён вопрос о том, применимы ли к автономным системам классические критерии моральной субъектности (свобода воли, осознанность, способность к суждению). Не разработана удовлетворительная модель распределения ответственности между множественными участниками процесса создания и эксплуатации интеллектуальных систем.

Цель исследования – выявить концептуальные ограничения применения категории моральной ответственности к системам сильного искусственного интеллекта и предложить трехуровневую модель распределённой ответственности.

Для достижения цели поставлены следующие *задачи:* проанализировать классические критерии моральной субъектности применительно к искусственным агентам; реконструировать аргумент Дж. Сёрла [8] о невозможности подлинной интенциональности у вычислительных систем и оценить его следствия для этической

теории; предложить трёхуровневую модель распределённой ответственности.

Научная новизна состоит в том, что концепция «вакуума ответственности» дополнена положениями институциональной этики и этики заботы, что позволяет преодолеть индивидуалистическую редукцию, свойственную большинству существующих подходов.

Материалы и методы

Методологической основой исследования выступает концептуальный анализ, позволяющий реконструировать содержание философских категорий (моральная ответственность, субъектность, интенциональность) и оценить границы их применения к новым объектам. Параллельно используется сравнительный метод, посредством которого сопоставляются позиции представителей различных этических традиций: деонтологии (И. Кант [5]), этики ответственности (Г. Йонас [4]), аналитической философии сознания (Дж. Сёрл [8]), информационной этики (Флориди [13; 14]).

Теоретическую базу составляют несколько ключевых понятий. Под сильным искусственным интеллектом, вслед за Дж. Сёрлом, понимается гипотетическая система, обладающая подлинным пониманием и сознанием, а не только способностью имитировать разумное поведение [8]. Моральная ответственность рассматривается как нормативное отношение между агентом и его действием, предполагающее возможность вменения похвалы или порицания. «Вакуум ответственности», согласно А. Маттиасу, возникает в ситуации, когда сложность обучающейся системы делает невозможной полную атрибуцию её поведения какому-либо конкретному лицу [16].

Ход анализа выстроен не как простое сопоставление авторских позиций, а как проверка условий вменения ответственности. Сначала из классических теорий выделяются признаки агента: способность понимать обстоятельства поступка; возможность соотнести действие с нормой; контроль над последствиями хотя бы в пределах разумного предвидения. Затем эти признаки переносятся на искусственные системы и проверяются по каждому пункту. Такая после-

довательность нужна потому, что модель ответственности нельзя выводить из самого факта автономности алгоритма: автономность в техническом смысле ещё не даёт моральной самостоятельности.

Дальнейший этап связан с разрывом между субъектностью и практическим вредом. Если система не обладает сознанием и интенциональностью, её нельзя описывать как виновного субъекта в строгом смысле. При этом последствия её решений остаются морально значимыми для людей, которые попадают в зону действия алгоритма. Поэтому анализ переходит от вопроса о вине машины к вопросу о распределении обязанностей между теми, кто проектирует систему, допускает её к применению и контролирует последствия её работы.

Результаты и обсуждение

Аристотель в «Никомаховой этике» связывает ответственность с двумя условиями: добровольностью действия и осведомлённостью агента о существенных обстоятельствах [1]. Действие считается добровольным, если его источник находится в самом деятеле и деятель осознаёт частные обстоятельства поступка. Применение этих критериев к автономным системам наталкивается на принципиальную трудность. Нейросетевая модель, обученная на массиве данных, действительно порождает выходные данные «изнутри себя», то есть формально источник действия локализован в системе. Тем не менее характер этого «изнутри» радикально отличается от аристотелевского понимания у модели отсутствует способность к осознанному практическому рассуждению, которая для Аристотеля неотделима от добродетели.

Кантовская этика ставит ещё более жёсткое условие. Моральный субъект, по И.Канту, определяется автономией воли, то есть способностью действовать в соответствии с законом, который субъект сам себе полагает [5]. Категорический императив предполагает, что агент способен к рефлексии над максимой своего поступка и к универсализации этой максимы. Вычислительная система, какой бы сложной она ни была, оперирует функциями оптимизации, а не максимами. Разница здесь не количественная, а качественная: оптимизация целевой функции принципиально отличается от мо-

рального самоопределения. И. Кант, по-видимому, отказал бы даже идеальному искусственному агенту в статусе морального субъекта, поскольку последний не обладает разумом как способностью к автономному законодательству.

Аргумент Дж. Сёрла о «китайской комнате» был сформулирован ещё в 1980 г., однако сохраняет значимость для современных дискуссий. Дж. Сёрл проводит различие между синтаксической обработкой символов и семантическим пониманием. Компьютерная программа, каким бы совершенным ни был её алгоритм, манипулирует символами по формальным правилам; она не понимает смысла обрабатываемой информации. Из этого следует, что приписывание интенциональности вычислительным системам лишено оснований. Без интенциональности нет подлинного действия, есть лишь выполнение программы; без подлинного действия невозможна моральная субъектность [8].

Против аргумента Дж. Сёрла выдвигались многочисленные возражения. Одно из наиболее известных, ответ «системы в целом», утверждает, что понимание возникает не у отдельного компонента, а у системы как целого. Дж. Сёрл парировал это возражение, заметив, что даже если оператор интернализует все компоненты системы, понимания не возникнет [8]. Для целей настоящего исследования существенен не столько вопрос об интенциональности сам по себе, сколько его этическое следствие: если у системы нет интенциональности, ей нельзя вменить ответственность в кантовском или аристотелевском смысле.

Признание невозможности прямого приписывания ответственности машине не снимает проблемы; оно лишь переформулирует её. А. Маттиас показал, что в случае обучающихся систем возникает «вакуум ответственности»: ни программист, ни пользователь, ни владелец не могут предвидеть все решения системы после обучения [16; 24]. Программист задаёт архитектуру и начальные условия, но не контролирует поведение модели после дообучения на новых данных. Пользователь применяет систему, не располагая техническим знанием о её внутреннем устройстве. Владелец несёт юридическую, но не эпистемическую ответственность. Автор опи-

сывал системы, обучающиеся лишь в определенной ограниченной области, и это уже было сложно для понимания. Сильный интеллект, построенный на нейросетях, имеет универсальное назначение и обладает фундаментальной непрозрачностью; приписывание ответственности становится невозможным, ведь мы уже не понимаем причин принятых им решений. Так теряется контроль уже над всем процессом взаимодействия, а не только над отдельным актом.

Л. Флориди предлагает различать моральную ответственность и моральную подотчётность. Подотчётность не требует интенциональности; она опирается на функциональный критерий: агент подотчётен, если его действие может быть оценено как морально значимое по его последствиям, независимо от наличия у агента сознания [13]. Такой подход позволяет говорить об ответственности за поведение искусственного агента, перенеся фокус с вопроса «кто виноват?» на вопрос «кто должен отвечать?». Этот сдвиг, впрочем, сам по себе не указывает на конкретного носителя ответственности.

Г. Йонас в «Принципе ответственности» аргументировал, что технологическая мощь человечества порождает новый тип этического обязательства: ответственность перед будущими поколениями и перед сохранением условий для подлинно человеческой жизни [4]. Если экстраполировать позицию Г. Йонаса на сферу искусственного интеллекта, то разработчик автономной системы несёт перспективную ответственность, обращённую в будущее: он обязан предвидеть потенциальные последствия своего изобретения и минимизировать риски. Эта ответственность не индивидуальна; она коллективна, поскольку ни один разработчик не создаёт систему в одиночку. Принятые в последние годы международные постановления и регламенты отражают эту позицию [18; 19; 22].

Таким образом, вначале из аристотелевского и кантовского подходов выделен предел прямого переноса: алгоритм не выполняет действие как осознанный поступок и не полагает себе моральный закон. Поэтому машина не становится носителем вины даже тогда, когда её решение причиняет вред. Этот вывод не закрывает проблему. Он лишь показывает, что искать ответственного нужно не вну-

три алгоритма, а в цепочке людей и организаций, которые задали архитектуру системы, условия её обучения и область применения.

Затем был учтён аргумент Дж. Сёрла о разрыве между синтаксической обработкой и пониманием. Если принять этот аргумент, искусственный агент не может отвечать за смысл собственного решения, поскольку смысл не возникает из одной формальной операции. Но вред от решения всё равно входит в социальную практику: врач получает рекомендацию, банк отказывает клиенту в кредите, водитель взаимодействует с беспилотной системой. Именно поэтому ответственность перемещается к тем участникам, которые способны оценивать риск до внедрения, во время эксплуатации и после возникновения ущерба.

На этом этапе позиция А. Маттиаса вызывает центральное затруднение: обучающаяся система может изменить поведение после обновления данных, поэтому один разработчик не предвидит все траектории её дальнейшей работы. Пользователь тоже не контролирует архитектуру модели. Институция, допускающая систему в оборот, часто работает с уже готовым продуктом и не видит каждого технического решения. Разрыв не снимается поиском одного виновного. Он требует разделения ответственности по стадиям жизненного цикла системы [16].

Разграничение Л. Флориди между ответственностью и подотчётностью помогает перейти от персональной вины к обязанностям сопровождения системы. Подотчётность здесь не заменяет моральную ответственность полностью; она фиксирует минимум, без которого автономная система превращается в закрытый источник решений. На этом уровне появляется требование объяснимости: не всякое решение модели можно сделать прозрачным до конца, но отсутствие хотя бы проверяемой процедуры делает распределение ответственности пустой формулой [13].

Так получаются три узла ответственности. Разработчик связан с архитектурой системы, качеством данных, тестированием и объяснимостью модели. Оператор связан с режимом применения: он решает, где система допустима, когда требуется человеческая про-

верка и какие решения нельзя оставлять без контроля. Институты задают внешнюю рамку: сертификацию, стандарты безопасности, правила аудита и порядок возмещения вреда. Модель, таким образом, выводится из различия между тремя формами контроля: техническим, эксплуатационным и нормативным.

Синтезируя рассмотренные позиции, можно предложить трёхуровневую модель распределённой ответственности. Техническая подотчётность разработчика предполагает обязанность обеспечить прозрачность и объяснимость алгоритмов, провести тестирование на предмет систематических ошибок и предусмотреть механизмы аварийного отключения. Операционная ответственность пользователя включает обязанность применять систему в пределах специфицированных условий, контролировать её выходные данные и не делегировать критические решения машине без верификации. Регуляторная ответственность институтов охватывает сертификацию, стандартизацию и правовое регулирование автономных систем [18; 19; 21; 22].

Подобная модель не лишена ограничений. Границы между уровнями размыты: разработчик может одновременно выступать оператором, а институт, регулирующий отрасль, может быть заинтересованной стороной. Проблема усугубляется тем, что обучающиеся системы эволюционируют: ответственность, зафиксированная на момент создания, может оказаться неадекватной спустя несколько циклов переобучения. Здесь уместно вернуться к идее Г. Йонаса о перспективной ответственности: она должна быть не статичной, а динамической, обновляемой вместе с развитием системы.

Выводы

Проделанный анализ позволяет сформулировать несколько выводов. Ни аристотелевская модель добровольного действия, ни кантовский критерий автономии воли не могут быть перенесены на искусственные агенты без утраты своего содержания. Аргумент Дж. Сёрла о невозможности подлинной интенциональности у вычислительных систем остаётся состоятельным; его опровержения, при всей их изобретательности, не достигают цели. Из этого следу-

ет, что искусственный интеллект, каким бы «сильным» он ни был в функциональном смысле, не может быть признан моральным субъектом в классическом понимании.

Между тем отказ от приписывания ответственности машине не означает, что проблема ответственности исчезает. Напротив, она становится острее, поскольку традиционная дихотомия «субъект ответственности, объект действия» перестаёт работать. «Вакуум ответственности», описанный А. Маттиасом, реален; он уже проявляется в судебных спорах, связанных с авариями беспилотных транспортных средств и с ошибками медицинских диагностических систем.

Предложенная трёхуровневая модель распределённой ответственности не претендует на окончательное решение. Она намечает направление, в котором классическая этика может быть адаптирована к условиям технологического прогресса. Распределение ответственности между разработчиком, оператором и институтом позволяет избежать двух крайностей: полного снятия ответственности с человека (что аморально) и возложения всей ответственности на одного участника (что нереалистично) [1].

Описанная модель опирается не на механическое суммирование философских позиций, а на последовательное выявление трёх дефицитов: дефицита моральной субъектности у машины; дефицита предвидения у отдельного человека; дефицита устойчивого контроля без институтов. Каждому дефициту соответствует отдельный уровень ответственности. Поэтому разработчик, оператор и регулятор включены в модель не как условный перечень лиц, а как разные точки контроля над жизненным циклом автономной системы.

Выводы статьи соотносятся с позицией Л. Флориди, настаивающего на разграничении ответственности и подотчётности, и дополняют её институциональным измерением. Позиция Н. Бострома о необходимости согласования целей подтверждается, однако согласование целей недостаточно без механизмов распределённого контроля. Результаты исследования коррелируют также с отечественными разработками Е.А. Никитиной и Б.Г. Юдина, подчёркивающими трансформацию субъектности в условиях цифровизации [7; 11].

Ограничения исследования связаны с тем, что данный анализ сосредоточен преимущественно на западной философской традиции; буддийская и конфуцианская этика, предлагающие иные модели моральной субъектности, остались за пределами рассмотрения. Перспективным направлением дальнейшего исследования может стать кросс-культурный сравнительный анализ подходов к ответственности за действия искусственных агентов, а также разработка формализованных критериев, позволяющих фиксировать динамическую ответственность на каждом этапе жизненного цикла автономной системы.

Категория моральной ответственности не утрачивает значимости; она требует структурного усложнения, перехода от индивидуальной атрибуции к распределённой модели, учитывающей множественность агентов и динамический характер технологической среды.

Список литературы

1. Аристотель. (2020). *Никомахова этика*. Москва; Берлин: Директ-Медиа, 222 с.
2. Баксанский, О. Е., & Сорокина, С. Г. (2025). Этика искусственного интеллекта: сбалансированный подход к развитию и применению. *Общество: философия, история, культура, 1*, 23–33. https://doi.org/10.24158/fik.2025.1.2, EDN: https://elibrary.ru/AVFHGB
3. Бостром, Н. (2016). *Искусственный интеллект. Этапы. Угрозы. Стратегии*. Москва: Манн, Иванов и Фербер, 496 с.
4. Йонас, Х. (2004). *Принцип ответственности. Опыт этики для технологической цивилизации*. Москва: Айрис-пресс, 480 с.
5. Кант, И. (2005). *Критика практического разума*. Санкт-Петербург: Наука, 528 с.
6. Лекторский, В. А. (2024). Человек и системы искусственного интеллекта. В *Искусственный интеллект: материалы заседания 14 октября 2023* (с. 20–50, 76–77, 164–167). Москва: Новые печатные технологии.
7. Никитина, Е. А. (2014). Искусственный интеллект: философия, методология, инновации. *Философские проблемы информационных технологий и киберпространства, 2*(8), 108–122. EDN: https://elibrary.ru/TENWVP
8. Сёрл, Дж. (1990). Разум мозга – компьютерная программа? *В мире науки, 3*, 7–13.
9. Сорокина, С. Г. (2023). Искусственный интеллект в контексте междисциплинарных исследований языка. *Вестник Кемеровского государственного университета. Сер.: Гуманитарные и общественные науки, 7*(3), 267–280. https://doi.org/10.21603/2542-1840-2023-7-3-267-280, EDN: https://elibrary.ru/FEIQKS

10. Филиппович, Ю. С., & Филиппович, М. С. (2025). Опыт доверия в контексте усталости и одиночества человека в современном обществе. *Вестник Мининского университета, 13*(3). https://doi.org/10.26795/2307-1281-2025-13-3-14, EDN: https://elibrary.ru/SRYPUQ
11. Юдин, Б. Г. (2018). *Человек: выход за пределы*. Москва: Прогресс-Традиция, 472 с.
12. Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press, 237 p. https://doi.org/10.7551/mitpress/12549.001.0001
13. Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d, EDN: https://elibrary.ru/ECMQOX
14. Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001
15. Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press, 290 p. https://doi.org/10.1017/CBO9780511554476
16. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1, EDN: https://elibrary.ru/WQSUMU
17. Mill, J. S. (1863). *Utilitarianism*. London: Parker, Son and Bourn, 95 p.
18. OECD. (2019; обновлено в 2024). *Recommendation of the Council on Artificial Intelligence*. Paris: OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (дата обращения: 24.05.2026).
19. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, 12 июля. https://eur-lex.europa.eu/eli/reg/2024/1689/oj (дата обращения: 24.05.2026).
20. Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x, EDN: https://elibrary.ru/FNKIAH
21. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Gaithersburg: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
22. UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380455 (дата обращения: 24.05.2026).
23. Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Minds and Machines, 34*(3). Article 20. https://doi.org/10.1007/s11023-024-09674-0, EDN: https://elibrary.ru/STVLOD

24. Veluwenkamp, H., & Hindriks, F. (2024). Artificial agents: Responsibility & control gaps. *Inquiry*. Advance online publication. https://doi.org/10.1080/0020174X.2024.2410995

References

1. Aristotle. (2020). *Nicomachean Ethics*. Moscow; Berlin: Direct-Media, 222 p.
2. Baksansky, O. E., & Sorokina, S. G. (2025). Ethics of artificial intelligence: A balanced approach to the development and application. *Society: Philosophy, History, Culture, 1*, 23–33. https://doi.org/10.24158/fik.2025.1.2, EDN: https://elibrary.ru/AVFHGB
3. Bostrom, N. (2016). *Artificial Intelligence: Stages. Threats. Strategies*. Moscow: Mann, Ivanov and Ferber, 496 p.
4. Jonas, H. (2004). *The Principle of Responsibility: An Attempt at an Ethics for the Technological Civilization*. Moscow: Iris-Press, 480 p.
5. Kant, I. (2005). *Critique of Practical Reason*. St. Petersburg: Nauka, 528 p.
6. Lektorsky, V. A. (2024). Man and artificial intelligence systems. In *Artificial Intelligence: Proceedings of the Meeting of October 14, 2023* (pp. 20–50, 76–77, 164–167). Moscow: New Printing Technologies.
7. Nikitina, E. A. (2014). Artificial intelligence: Philosophy, methodology, innovations. *Philosophical Problems of Information Technologies and Cyberspace, 2*(8), 108–122. EDN: https://elibrary.ru/TENWVP
8. Searle, J. (1990). Is the mind of the brain a computer program? *In the World of Science (Russian edition of Scientific American)*, 3, 7–13.
9. Sorokina, S. G. (2023). Artificial intelligence in the context of interdisciplinary studies of language. *Bulletin of Kemerovo State University. Series: Humanities and Social Sciences, 7*(3), 267–280. https://doi.org/10.21603/2542-1840-2023-7-3-267-280, EDN: https://elibrary.ru/FEIQKS
10. Filippovich, Yu. S., & Filippovich, M. S. (2025). The experience of trust in the context of fatigue and loneliness of a person in modern society. *Vestnik of Minin University, 13*(3). https://doi.org/10.26795/2307-1281-2025-13-3-14, EDN: https://elibrary.ru/SRYPUQ
11. Yudin, B. G. (2018). *Man: Going Beyond the Limits*. Moscow: Progress-Tradition, 472 p.
12. Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press, 237 p. https://doi.org/10.7551/mitpress/12549.001.0001
13. Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d, EDN: https://elibrary.ru/ECMQOX
14. Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001

15. Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press, 290 p. https://doi.org/10.1017/CBO9780511554476
16. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1, EDN: https://elibrary.ru/WQSUMU
17. Mill, J. S. (1863). *Utilitarianism*. London: Parker, Son and Bourn, 95 p.
18. OECD. (2019; updated 2024). *Recommendation of the Council on Artificial Intelligence*. Paris: OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed: 24.05.2026).
19. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, July 12. https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed: 24.05.2026).
20. Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x, EDN: https://elibrary.ru/FNKIAH
21. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Gaithersburg: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
22. UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380455 (accessed: 24.05.2026).
23. Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Minds and Machines, 34*(3). Article 20. https://doi.org/10.1007/s11023-024-09674-0, EDN: https://elibrary.ru/STVLOD
24. Veluwenkamp, H., & Hindriks, F. (2024). Artificial agents: Responsibility & control gaps. *Inquiry*. Advance online publication. https://doi.org/10.1080/0020174X.2024.2410995

ДАННЫЕ ОБ АВТОРАХ

Мальцева Светлана Михайловна, канд. филос. наук, доцент, доцент кафедры философии и теологии; доцент кафедры общеобразовательных и профессиональных дисциплин
Мининский университет; НИПС - филиал ПривГУПС
ул. Ульянова, 1, г. Нижний Новгород, 603950, Российская Федерация; пл. Комсомольская, 3, г. Нижний Новгород, 603011, Российская Федерация
maltsewasvetlana@yandex.ru

Грачева Арина Дмитриевна, студент кафедры профессионального образования и управления образовательными системами
Мининский университет

ул. Ульянова, 1, г. Нижний Новгород, 603950, Российская Федерация
Gracheva.Arina28@mail.ru

Ротанова Мира Борисовна, канд. филос. наук, доцент кафедры философии, общественных коммуникаций и туризма
Нижегородский государственный лингвистический университет им. Н.А. Добролюбова
ул. Минина, 31а, г. Нижний Новгород, 603155, Российская Федерация
mrotanova@lunn.ru

DATA ABOUT THE AUTHORS

Svetlana M. Maltseva, PhD, Associate Professor, Associate Professor of the Department of Philosophy and Theology; Associate Professor of the Department of General Education and Professional Disciplines
Minin University; Nizhny Novgorod Institute of Railway Engineering - branch of the Volga State University of Railway Engineering
1, Ulyanov Str., Nizhny Novgorod, 603950, Russian Federation; 3, Komsomolskaya Sq., Nizhny Novgorod, 603011, Russian Federation
maltsewasvetlana@yandex.ru
ORCID: https://orcid.org/0000-0002-7273-1852
Scopus Author ID: 57204862811

Arina D. Gracheva, Student of the Department of Professional Education and Management of Educational Systems
Minin University
1, Ulyanov Str., Nizhny Novgorod, 603950, Russian Federation
Gracheva.Arina28@mail.ru

Mira B. Rotanova, PhD, Associate Professor of Philosophy, Public Communications and Tourism
Linguistic University of Nizhny Novgorod
31a, Minin Str., Nizhny Novgorod, 603155, Russian Federation
mrotanova@lunn.ru

Поступила 05.06.2026
После рецензирования 02.07.2026
Принята 09.07.2026

Received 05.06.2026
Revised 02.07.2026
Accepted 09.07.2026