УДК 1:004.8 EDN VDSNDF

Искусственный интеллект и кризис ответственности в призме кантовского императива

© Н.Г. Багдасарьян[1], М.С. Бодрина[2]

[1]МГТУ им. Н.Э. Баумана, Москва, 105005, Россия
[2]МГУ имени М.В. Ломоносова, Москва, 119991, Россия

Представлен анализ проблематики кризиса ответственности в процессах применения искусственного интеллекта. Показано, что обострение давних проблем ответственности порождает то, что можно обозначить как разрыв вменения, при котором значимые действия и их последствия перестают быть однозначно приписываемыми сознательной моральной воле. Сложность, неоднозначность и новизна проблематики ответственности в контексте искусственного интеллекта требуют обращения к надежному методологическому фундаменту, каковым является категорический императив Иммануила Канта. Отмечено, что моральная философия Иммануила Канта — с ее акцентом на автономию, самозаконоустановление и категорический императив — предоставляет строгую теоретическую рамку для объяснения того, почему искусственный интеллект как гетерономная система не может быть признан моральным агентом. Проанализировано, каким образом Национальная стратегия развития искусственного интеллекта в Российской Федерации отвечает на эту проблему, формируя правовую и техническую архитектуру, призванную закрыть данный «кантианский разрыв». Стратегия рассмотрена как политико-правовая рамка, переводящая философские подходы в плоскость регуляторной практики и предлагающая пути к сохранению человеческого суверенитета и восстановлению подотчетности в условиях, формируемых неморальными, алгоритмическими системами. Исследование основано на критическом анализе современных философско-правовых, политико-философских и нормативных подходов к искусственному интеллекту.

Ключевые слова: *Кант, разрыв вменения, этика, ответственность, управление искусственным интеллектом, доверенный ИИ, стратегия ИИ России*

Стремительное развитие искусственного интеллекта (ИИ), его внедрение в социальную и политическую жизнь представляет собой не только технологическое изменение, но и вызов устоявшимся моральным и правовым предпосылкам, подрыв общепризнанных структур ответственности, что можно обозначить как разрыв вменения[1].

[1] Разрыв вменения — это правовая и теоретическая проблема, возникающая из-за сложности и способности к принятию решений систем искусственного интеллекта, что затрудняет применение традиционного уголовного и деликтного права, усложняя определение ответственности за халатные действия или причиненный вред.

По мере того как системы ИИ приобретают все более высокую степень функциональной автономии, существующие модели ответственности оказываются под напряжением. Действия, имеющие существенные последствия, перестают быть однозначно приписываемыми сознательному, рассуждающему субъекту, возникает разрыв подотчетности.

Проблема этики и ответственности по мере распространения применения ИИ практически во всех сферах социальной жизни — в управлении персоналом, образовании, науке, экономической деятельности, шоу-бизнесе и др. — приобретает статус первостепенной. Ее интенсивное осмысление ведется в разнообразных дискурсах, в первую очередь в правовом и философском, что отражено, например, в выпуске журнала «Философия» (2025, т. 9, № 4), включающем работы Д. Быльевой и А. Нордманна «Онтолитический эффект искусственного интеллекта: расширяя понимание ответственных исследований и инноваций» [1], Е. Труфановой «Надежность и ответственность как ключевые вопросы применения ИИ-систем: человек в петле ответственности» [2] и П. Чэна, Чж. Чжана «Механизм формирования ответственности и логика этического управления в воплощенном искусственном интеллекте» [3]. Осознание силы и опасности ИИ привело к этическому повороту и стремлению справиться с этой проблемой на самых разных уровнях — от частных корпораций до государственных структур. Ответом стала подготовка руководств, рекомендаций, кодексов, правил разработки и применения ИИ. Однако даже сам факт лавинообразного потока документов-регламентаций — свидетельство того, что проблема этики и ответственности в сфере ИИ далека от своего разрешения. Показательно появление работы Л. Манна с коротким и однозначным названием «Бесполезность этики ИИ» («The uselessness of AI ethics») [4].

Учитывая многоаспектность проблематики ответственности ИИ, рассмотрим лишь один ее срез, связанный с природой современного разрыва вменения, опираясь на моральную философию Иммануила Канта, которая определяет моральную агентность через автономию, самозаконоустановление и следование категорическому императиву. Тем самым формируются концептуальные границы осмысления данной проблемы. Рассмотрение ИИ в призме кантовского подхода показывает, что системы ИИ лишены подлинного намерения и морального самоуправления и потому не могут рассматриваться в качестве моральных агентов. Это позволяет прояснить природу современного разрыва вменения, а не замаскировать его.

В определенной мере принципы проектирования Национальной стратегии развития ИИ в Российской Федерации (далее — Стратегия)

дают ответ на данный разрыв. Акцент на недопустимости делегирования моральных решений, механизмах сертификации «доверенных технологий искусственного интеллекта» [5], а также требованиях прозрачности и человеческого контроля направлен на сохранение человеческой ответственности в процессах, опосредованных ИИ. Вместо попытки приписать машинам моральную агентность, такой подход рассматривает ИИ как высокоразвитый инструмент, функционирование которого остается подчиненным человеческому суждению. В этом отношении стратегия очерчивает модель управления ИИ, которая является одновременно практико-ориентированной и философски последовательной, согласуя технологическое регулирование с сохранением человеческой моральной ответственности, лежащей в самом основании кантовской этики.

Исследование основано на критическом анализе современных философско-правовых, политико-философских и нормативных подходов к ИИ. Методологическая основа сочетает нормативно-ценностный анализ с философским и эпистемологическим исследованием, что позволяет целенаправленно рассмотреть проблемы ответственности, агентности и суверенитета в технологически опосредованных системах. Системный подход используется для интерпретации ИИ, правового регулирования и государственного управления как взаимосвязанных элементов более широкого социотехнического порядка. Междисциплинарная перспектива, опирающаяся на юриспруденцию, философию (этику и философскую антропологию), а также исследования науки и техники, обеспечивает комплексный анализ возникающих регуляторных вызовов.

Искусственный интеллект и традиционно понимаемая ответственность. Попытки приблизить ИИ к человеческому мышлению порождают не только технические, но и фундаментальные философские вопросы. Философия Иммануила Канта предоставляет масштабный инструментарий для их анализа. Кант утверждал, что познание не возникает из пассивного восприятия информации, а формируется благодаря активной роли сознания в структурировании опыта. Его так называемая коперниканская революция состоит в утверждении, что предметы должны сообразовываться с познанием, т. е. условия человеческого понимания определяют саму возможность явления мира для людей. Здесь прослеживается фундаментальное противоречие в исследовании ИИ: в то время как человеческое познание, по Канту, опирается на внутренние условия опыта, системы ИИ функционируют посредством внешне сконструированных вычислительных механизмов.

Для Канта познание представляет собой активный процесс синтеза. Многообразие чувственности — исходный материал опыта —

упорядочивается посредством априорных форм и понятий тремя центральными способностями:

1) *чувственность* (Sinnlichkeit) есть способность получать созерцания, которые неизбежно даны в формах пространства и времени. Эти формы не извлекаются из опыта, а, напротив, делают сам опыт возможным;

2) *рассудок* (Verstand) структурирует созерцания посредством категорий, таких как причинность и единство. Как пишет Кант, «мысли без содержания пусты, созерцания без понятий слепы» [6]. Категории определяют способы, в которых предметы и события могут быть помыслены;

3) *воображение* (Einbildungskraft) опосредует чувственность и рассудок, синтезируя многообразие созерцания в связный опыт. Кант характеризует его как необходимую, но частично непрозрачную умственную деятельность, позднее описываемую комментаторами через такие понятия, как «синтезы рассудка, сила воображения, схематизм, эпигенезис» [7, ст. 1035].

Как соотносятся эти элементы познавательной деятельности с ИИ? Если некоторые измерения познания допускают сопоставление с операциями по правилам, то другие принципиально сопротивляются подобной формализации:

- рассудок как вычисление. Деятельность рассудка, подчиненная правилам, в определенной степени может быть сопоставлена с вычислительными процессами. Именно в этой области современные системы ИИ демонстрируют наибольшую эффективность — прежде всего в задачах формальной обработки, распознавания закономерностей и вывода на основе данных;
- воображение как невычислимый синтез. Напротив, синтезирующая деятельность воображения с трудом переводится в алгоритмические термины. Хотя системы ИИ способны следовать правилам или порождать результаты, внешне напоминающие творчество, они не осуществляют доконцептуального объединения опыта, которое Кант считает фундаментальным для познания. Этот синтез не является следствием применения правил, но представляет собой условие самой возможности познания, регулируемого правилами.

Пределы ИИ становятся еще более очевидными при обращении к кантовскому учению о способности суждения (Urteilskraft). Кант различает два вида суждений:

1) определяющее суждение, подводящее частное под уже заданное правило;

2) рефлексивное суждение, которое ищет или порождает соответствующее правило для новой ситуации.

Рефлексивное суждение зависит от динамического взаимодействия воображения и рассудка — процесса, который Кант описывает как их «свободную игру» («in einem freien Spiele») [8, ст. 55]. Эта способность лежит в основе творчества, понятийного новаторства и эстетического суждения. Однако системы ИИ функционируют в рамках фиксированных архитектур и критериев оптимизации даже тогда, когда их результаты кажутся оригинальными. Как подчеркивается в современной кантовской литературе, «рефлексия в составе познавательной способности» является необходимым условием подлинного познания [7, ст. 1041].

Эти соображения имеют прямое отношение к дискуссиям о сверхинтеллекте. С кантовской точки зрения сама эта идея проблематична. Любой интеллект — человеческий или иной — действует в определенных условиях опыта. Кант определяет фундаментальное условие как «трансцендентальное единство апперцепции… то единство, посредством которого все многообразие, данное в созерцании, объединяется в понятие предмета» [9, ст. 120]. Именно это синтетическое, объективное единство впервые конституирует связный мир для познающего субъекта. Даже если система ИИ способна изменять собственный код, она не порождает ту фундаментальную рамку, в пределах которой опыт вообще становится возможным. Она обрабатывает многообразие данных, но не является источником трансцендентального единства, делающего предмет опыта мыслимым. В кантовском смысле она не полагает собственного бытия.

Следовательно, проблема заключается не в том, что ИИ станет автономным субъектом, превосходящим человеческое познание. Напротив, он может все в большей степени выглядеть агентоподобным, не обладая при этом подлинной агентностью. В результате появляются системы, которые действуют мощно и порой непредсказуемо, но лишены способности к рефлексивному суждению и моральной ответственности.

Такие системы неподотчетны не потому, что превосходят человеческое познание, а потому, что лишены того единства сознания, которое у Канта служит основанием самой возможности ответственности. Для Канта моральная ответственность предполагает автономную волю, которая, по его словам, «не просто подчинена закону, но подчинена ему таким образом, что она должна рассматриваться и как устанавливающая закон для самой себя и именно потому как впервые подчиненная закону, автором которого она может считать себя» [10, ст. 34]. ИИ, напротив, целиком функционирует в мире, сконституированном человеческим познанием; он сам этот мир не конституирует и не способен к самозаконоустановляющей автономии. В этом смысле ИИ представляет собой новый тип деятеля — эффективный

и непрозрачный, но принципиально лишенный намерения и самотворящей воли, необходимой для подлинной ответственности.

Почему искусственный интеллект не может быть моральным агентом? Для Канта моральная агентность укоренена в автономии, понимаемой не как свобода выбора между альтернативами, а как способность к самозаконоустановлению. Он определяет автономию как «свойство воли быть законом самой себе» [11, ст. 47]. Существо является морально автономным лишь постольку, поскольку оно определяет свои максимы посредством чистого практического разума, а не под воздействием эмпирических побуждений или внешнего руководства. Автономия, таким образом, лежит в основании морального достоинства; как отмечает Кант, она есть «основание достоинства человеческой и всякой разумной природы» [11, ст. 43]. Концепция получает свое нормативное выражение в категорическом императиве, требующем действовать лишь по тем максимам, которые могут быть желаемы в качестве всеобщего закона [11].

Искусственные системы лишены способности к такому самозаконоустановлению. Их поведение целиком определяется внешне заданными целями, режимами обучения и процедурами оптимизации. В кантовских терминах они действуют в условиях гетерономии, а не автономии. Поскольку лишь самозаконоустанавливающее существо может рассматриваться как цель сама по себе, ИИ не может быть признан моральным субъектом. Он остается тем, что Кант называет *Sache* — «вещью» [11, ст. 37], а не лицом. Лицо, поясняет Кант, есть субъект, «действия которого могут быть вменены» [12, ст. 83], обладающий моральной личностью как «свободой разумного существа под моральными законами» [12, ст. 83] и потому связанный только теми законами, которые он дает себе сам. Системы ИИ этим условиям не отвечают.

Следовательно, системы ИИ могут приводить к последствиям, но не несут ответственности за них. Они действуют без воли. Хотя их процессы напоминают рассуждение, они не устанавливают для себя принципы действия. Данная особенность отчетливо раскрывается в рамках проблемы идентификации максимы. Кантовская моральная оценка сосредоточена на максиме, понимаемой как субъективный принцип воления, лежащий в основе поступка. Напротив, «решение» нейронной сети возникает из распределенных статистических процессов, а не из сознательно принятого принципа. Никакая связная максима не может быть приписана подобным механизмам. Хотя человеческие наблюдатели могут интерпретировать результаты работы ИИ как выражение намерения, Кант подчеркивает, что моральная оценка касается «принципа воли, независимо от целей» [11, ст. 13].

Поскольку ИИ действует исключительно в плоскости эмпирической причинности, он не способен обосновывать свое поведение априорным моральным законом.

Дополнительная сложность обусловлена коллективным и распределенным характером разработки ИИ. Современные системы формируются усилиями проектировщиков, институтов, наборов данных и инфраструктур, что подрывает требуемое для морального вменения единство. Кант понимает вменение как возложение ответственности на агента в качестве автора действия [11, ст. 37–40]. Когда системы ИИ порождают неожиданные или эмерджентные результаты, становится неясно, чья именно воля — если таковая вообще имела место — определяла действующий принцип. Кант прямо указывает, что неведение или отсутствие контроля могут устранять вменяемость. В контексте ИИ все чаще наблюдаются формы причинности без ясного авторства.

Эти концептуальные проблемы имеют прямые правовые последствия. Значительная часть западной правовой теории, прежде всего уголовное право, опирается на кантовскую модель агентности. Требование *mens rea* предполагает субъекта, способного понимать правовые нормы и свободно решать их нарушить. Кант полагает, что наказание оправданно лишь в том случае, если оно обращено к разумному агенту, свободно пожелавшему неправомерного поступка. Системы ИИ, лишенные как воли, так и морального самозаконоустановления, этим критериям не соответствуют. Схожие затруднения возникают в деликтном праве[2], опирающемся на такие понятия, как предсказуемость и разумная агентность. Системы ИИ способны причинять вред, но они не могут совершать правонарушение в кантовском смысле.

На более глубоком уровне данное ограничение обусловлено отсутствием того, что Кант называет фактом разума — непосредственного осознания морального закона в разумных существах. Кантовская моральная философия предполагает несводимое моральное сознание, которое не может быть выведено из эмпирической причинности. ИИ инвертирует эту структуру: он демонстрирует функциональную рациональность без морального сознания — техническую изощренность без нормативного основания. Его агентность потому является лишь кажущейся: сложной по функционированию, но пустой в моральном отношении.

[2] Деликт (лат. delictum) — правонарушение частноправового характера, выражающееся в нарушении права или запрета и причиняющее вред частному лицу, его семье либо имуществу, в результате чего возникают новые права и правовые обязанности — деликтные обязательства (obligationes ex delicto), реализуемые посредством исков из правонарушений (actiones ex delicto) [13].

Делегирование значимых решений таким системам чревато размыванием различия между подлинными моральными агентами и неавтономными инструментами. Для Канта разумное существо должно быть способно рассматривать себя как цель саму по себе и как законодательного члена «царства целей» («kingdom of ends») [11, ст. 42], подчиняющегося лишь тем законам, которые оно дает себе само. ИИ не может удовлетворить этому требованию, поскольку он полностью детерминирован эмпирическими причинами.

В этом отношении ИИ вводит тип деятеля, для которого кантовская теория не предлагает однозначной категории. Данные системы не являются ни лицами, ни просто инертными объектами. Они лишены автономии, но при этом порождают действия с серьезными моральными и социальными последствиями. Результатом становится сущность, которая действует без воления, производит эффекты без намерения и имитирует практическое рассуждение, оставаясь при этом вне морального сообщества агентов.

От метафизических вопросов к регуляторным решениям. Рассмотрение ИИ как кантовского морального агента представляет собой фундаментальную категориальную ошибку. Кантовские понятия автономии, чистого практического разума и самозаконоустановления предполагают форму агентности, неприменимую к системам, действующим целиком на основе гетерономных правил. Вместо попыток вписать ИИ в моральную рамку, разработанную для разумных существ, более продуктивно рассматривать его как особый класс мощных артефактов, управление которыми должно основываться на их функциональных возможностях, а не на метафизических утверждениях об их статусе.

Концептуальные трудности, возникающие в современных дискуссиях, во многом обусловлены распространенностью интенциональной установки (по Д. Деннету). Как поясняет Деннет, «интенциональная установка — это стратегия интерпретации поведения некоторой сущности… при которой ее рассматривают так, как если бы она была разумным агентом, определяющим свой “выбор” действия с учетом своих убеждений и желаний» [14, ст. 27]. Хотя такая стратегия может быть практически полезной, она поощряет приписывание убеждений, желаний или намерений системам, которые лишь демонстрируют функциональную эффективность. С кантовской точки зрения это ведет к смещенным дискуссиям о машинном сознании или моральном понимании применительно к неразумным системам. Для регуляторных целей релевантен не метафизический, а функциональный вопрос: действует ли система таким образом, что в целях управления ее следует рассматривать, *как если бы* она была агентом?

Кантовское различие между лицами и вещами предоставляет четкую нормативную рамку. Лица — это субъекты, «действия которых

могут быть им вменены», тогда как вещи — это сущности, которым ничего не может быть вменено — Sache ist ein Ding («Вещь есть вещь») [11, ст. 37]. Кант далее отмечает, что сущности могут подчиняться законам, не обладая автономией: «Все в природе действует согласно законам», однако лишь разумные существа «действуют согласно представлению о законе» [11, ст. 24]. Данное различие позволяет понимать системы ИИ как подчиненные законам, не смешивая их с моральными агентами. На этом основании может быть сформулировано функциональное определение агентности (Functionalist Definition of Agency, FDA) для регуляторных контекстов. Из него следуют два функциональных критерия.

1. Операционная автономия (функциональная, а не моральная). Система ИИ квалифицируется как агент в регуляторном смысле, если она способна воспринимать среду, обрабатывать информацию и действовать без постоянного человеческого вмешательства. Это отражает организованное, целенаправленное поведение, но не тождественно кантовской моральной автономии. Кант прямо противопоставляет автономию гетерономной мотивации, указывая, что когда повиновение закону обусловлено внешними стимулами или санкциями, «императив, которому мы следуем, подчиняясь этому закону, является гипотетическим императивом», например: «если ты не хочешь попасть в тюрьму или хочешь попасть в рай… то должен соблюдать этот закон» [11, ст. XXIV]. Системы ИИ, функционирующие исключительно на основе внешне заданных правил и целей, являются в этом смысле парадигматически гетерономными.

2. Способность к нормативно значимым эффектам. Система должна быть способна производить результаты, которые существенно затрагивают права, благополучие или безопасность лиц. Любая сущность с подобными эффектами требует механизмов управления, направленных на защиту морального сообщества, независимо от того, может ли она быть признана моральным агентом.

В совокупности эти критерии формируют регуляторную категорию «подотчетных артефактов» (accountable artifact[s]) [15, ст. 21]. Данные системы не могут быть отнесены ни к категории моральных субъектов, ни к разряду простых инструментов: они представляют собой гетерономные сущности, чья сложность и масштаб социального воздействия обусловливают необходимость структурированного надзора. Предложенная концептуальная модель согласуется с кантовским тезисом о том, что достоинство присуще исключительно автономным разумным существам, и одновременно учитывает, что неразумные артефакты требуют регулирования, соразмерного порождаемым ими рискам.

Такой подход позволяет избежать упрощенной оппозиции «ИИ как лицо» — «ИИ как инструмент». Вместо этого он признает степени

функциональной агентности при сохранении четких моральных границ. Кант настаивает, что лишь разумные существа могут «действовать согласно представлению о законах» [11, ст. 24], — способность, которой ИИ не обладает. В то же время регулирование мощных неразумных систем в целях защиты человеческого достоинства полностью согласуется с кантовским рассуждением.

Наконец, функционалистский подход смещает регуляторный фокус от ретроспективного обвинения к проспективному управлению. Поскольку системы ИИ действуют без воли и без максим, традиционное моральное вменение к ним неприменимо. Регулирование должно, следовательно, сосредоточиваться на институциональном и техническом проектировании, обеспечивающем согласование поведения ИИ с человеческими ценностями и интересами, защиту условий человеческой автономии при одновременном управлении системами, функционирующими вне сферы моральной агентности. Как подчеркивает Кант, «автономия есть, следовательно, основание достоинства человеческой природы и всякой разумной природы» [11, ст. 43].

Вектор дальнейшего развития. Разрыв вменения. Одной из ключевых проблем управления искусственным интеллектом является то, что Кант охарактеризовал бы как нарушение условий моральной ответственности — размывание четкой цепочки вменения. В кантовской концепции справедливость предполагает возможность отнесения поступков к рациональному субъекту, способному к свободному выбору (*causa libera*). Он проводит принципиальное различие между лицами — субъектами, действия которых подлежат вменению, и вещами, которым ничто не может быть вменено. Системы ИИ осложняют эту структуру, производя последствия без автора в указанном смысле. Их результаты возникают в силу технических процессов, а не морального агентства, что порождает структурный разрыв ответственности и повышает риск ее диффузии.

Национальная стратегия развития ИИ в России отвечает на эту проблему, последовательно сохраняя жесткое различие между лицами и вещами. Искусственному интеллекту прямо отказывается в моральной или правовой субъектности; он определяется как инструментальный объект: «Комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека» [16]. Ответственность за все последствия однозначно сохраняется за человеческими акторами, что зафиксировано в запрете на делегирование моральной ответственности: «Не допускается делегирование системам искусственного интеллекта ответственного нравственного выбора…

Ответственность за все последствия работы систем искусственного интеллекта всегда несет физическое или юридическое лицо…» [5].

Такое положение жестко закрепляет вменение в человеческой сфере. Независимо от уровня автономности или технической сложности системы ИИ ответственность не может быть на нее перенесена. Вещь не способна нести моральную ответственность, что позволяет сохранить кантовскую структуру подотчетности.

В то же время Стратегия учитывает распределенный характер разработки и внедрения ИИ. Системы ИИ создаются и используются множеством участников — разработчиками, поставщиками данных, интеграторами, операторами инфраструктуры и конечными пользователями, что порождает известную проблему множества рук. Вместо того чтобы допускать рассеивание ответственности, Стратегия предлагает ее дифференциацию по ролям и степени причиненного вреда: «Разграничение ответственности организаций — разработчиков и пользователей технологий искусственного интеллекта исходя из характера и степени причиненного вреда» [5].

Таким образом, вменение поддерживается не через приписывание действий единому агенту, а посредством структурированного распределения ответственности. В поддержку этого подхода Стратегия вводит категорию «доверенные технологии искусственного интеллекта» [5]. Такие системы подлежат обязательной сертификации («системы сертификации решений» [5]), оценке соответствия («системы оценки соответствия» [5]) и включению в национальный реестр («реестр апробированных доверенных технологий искусственного интеллекта» [5]). Правила государственных закупок дополнительно требуют использования федеральными органами власти исключительно сертифицированных систем. Сертификация выступает юридическим механизмом восстановления прослеживаемой цепочки ответственности даже в условиях множественности участников.

Стратегия также формирует эпистемические условия, необходимые для вменения. Принципы «прозрачность» [5] и «достоверность исходных данных» [5] направлены на обеспечение возможности анализа, оспаривания и отнесения решений, принимаемых с использованием ИИ, к ответственным субъектам. Надзор выстраивается на основе «риск-ориентированного подхода» [5], предполагающего соразмерность регулирования уровню потенциального вреда.

Несмотря на эти меры, ряд кантовских затруднений сохраняется. Хотя ответственность четко закрепляется за человеческими акторами, Стратегия не в полной мере конкретизирует механизмы ее реализации, в частности, остается неясным, должна ли ответственность носить строгий, виновный или смешанный характер. Кроме того, использование перспективных методов ИИ, нередко разрабатываемых

и поддерживаемых за пределами национальной юрисдикции, дополнительно осложняет вменение, поскольку авторство может быть юридически удаленным, фрагментированным или непрозрачным.

Защита человеческой автономии от симулированного агентства. Для Канта человеческое достоинство основано не на интеллекте или техническом мастерстве, а на автономии воли — способности самому себе давать моральный закон. Именно эта способность к самозаконодательству определяет личность и проводит границу между субъектами и объектами. Поэтому системы ИИ, способные имитировать адаптивную автономность, включая формы «сильного» [5] ИИ, которые могут адаптироваться к меняющимся условиям без прямого человеческого вмешательства, порождают специфическую моральную проблему. Риск заключается не в потенциальном обретении такими системами личностных свойств, а в ослаблении, вытеснении или утрате отчетливости человеческого морального авторства.

Национальная стратегия развития ИИ отвечает на этот вызов, подтверждая неделегируемость морального выбора. Запрет на передачу ответственного нравственного решения системам ИИ прямо применяется в данном контексте: человек остается единственным носителем моральной ответственности. Системы ИИ могут содействовать, предоставляя рекомендации, прогнозы или оптимизационные решения, но не способны принимать решения в морально ответственном смысле. В кантовских терминах это означает сохранение суверенитета воли перед лицом подмены гетерономными механизмами.

Этот принцип подкрепляется институциональным устройством. В чувствительных сферах допускается использование только сертифицированных «доверенных технологий искусственного интеллекта» [5] и лишь после экспертной оценки и государственного одобрения посредством включения в национальный реестр. Таким образом, человеческое суждение вновь вводится на этапе внедрения, гарантируя, что системы ИИ допускаются к практике лишь после нормативной оценки.

Стратегия также стремится ограничить эпистемическую зависимость от ИИ. Через образовательные и просветительские меры она поощряет критическое отношение к технологиям ИИ, а не безрефлексивное доверие к ним. Этот акцент перекликается с кантовским призывом *Sapere aude* — «Имей мужество знать» («dare to know») [17], поскольку автономия предполагает не только правовые гарантии, но и развитие способности к рациональному суждению у граждан и специалистов.

На нормативном уровне этот подход имеет выраженно антропоцентрический характер. «Гуманистический подход» [5] утверждает права и свободы человека как высшие ценности, а принцип «уважение

автономии и свободы воли человека» [5] запрещает использование систем ИИ таким образом, который подрывал бы человеческий выбор или интеллектуальную независимость. Рациональная свобода человека рассматривается как фундаментальный элемент общественного и политического порядка.

Этический надзор дополнительно поддерживает эту ориентацию посредством «Кодекса этики в сфере искусственного интеллекта» [5], назначения специальных ответственных лиц («уполномоченные по этике» [5]) и постоянного национального и международного диалога. Таким образом, автономия защищается не только правовыми нормами, но и институционализированной рефлексией.

Вместе с тем остаются нерешенные напряжения. Стратегия не проводит четкого различия между моделями *human-in-the-loop* и *human-on-the-loop*. Кроме того, опосредованные механизмы, такие как алгоритмическое подталкивание, рейтинговые системы или поведенческая настройка, способны влиять на человеческие решения, формально не нарушая правовых требований. В результате воля человека остается юридически защищенной, но на практике оказывается уязвимой перед новыми формами воздействия.

Прозрачность, критический контроль и овладение непредсказуемостью. Категорический императив Канта требует, чтобы максима, направляющая действие, была способна к универсализации [11]. В применении к управлению ИИ это порождает практическую трудность: каким образом возможно этическое оценивание систем, чье функционирование основано на непрозрачных вероятностных моделях, обученных на масштабных массивах данных? Национальная стратегия развития ИИ отвечает на данный вызов посредством многоуровневой регуляторной модели, призванной повысить постижимость, обеспечить этический контроль и управлять неопределенностью.

Первым элементом этой модели является объяснимость. Принцип «прозрачность» [5] требует, чтобы процессы и результаты работы систем ИИ поддавались объяснению. Объяснимость выполняет как эпистемическую, так и нормативную функции: без минимального уровня постижимости невозможны ни доверие, ни подотчетность.

Второй элемент связан с целостностью данных. Требование «достоверность исходных данных» [5] устанавливает стандарты происхождения данных, их качества, разметки и совместимости. Эти меры снижают риск скрытой предвзятости и обеспечивают открытость эмпирической основы принятия решений с использованием ИИ для проверки и анализа.

Третий элемент вводит этическую фильтрацию через процедуры сертификации. «Доверенные технологии искусственного интеллекта» [5] оцениваются не только по техническим параметрам, но и по этическим критериям, таким как объективность, недискриминационность

и предотвращение вреда. Сертификация и оценка соответствия выступают процедурными аналогами кантовской универсализации: системы проверяются на общую допустимость до их внедрения.

Четвертый элемент направлен на работу с эффектами эмерджентности и непредсказуемости. Пилотные зоны («создание пилотных зон для апробации» [5]) и механизмы государственного мониторинга позволяют наблюдать за функционированием систем ИИ в контролируемых условиях. Реестр доверенных технологий не является статичным: системы могут повторно оцениваться, приостанавливаться или исключаться на основании данных, полученных после внедрения. Надзор выстраивается на основе риск-ориентированного подхода, при котором интенсивность регулирования соотносится с потенциальным ущербом, в том числе применительно к крупным генеративным моделям.

Этическое управление дополнительно поддерживается Кодексом этики, институтом уполномоченных по этике, а также постоянными общественными и экспертными консультациями. Этическая рефлексия рассматривается не как разовая процедура соблюдения требований, а как непрерывный процесс.

Несмотря на перечисленные меры, сохраняются существенные ограничения. В случае «больших фундаментальных моделей» [5] полная объяснимость может быть технически недостижимой, что ограничивает подлинную постижимость их функционирования. Кантовское требование прозрачности и универсализируемости максим может быть воспроизведено лишь процедурно, но не реализовано в полном объеме. Это указывает на более общий вывод: управление способно снижать и контролировать моральные риски, связанные со сложными нерациональными системами, но не может устранить их полностью.

В заключение следует отметить, что эффективное управление ИИ должно быть ориентировано на сохранение человеческой подотчетности, а не на приписывание машинам способности к этическому суждению. Национальная стратегия развития ИИ в России иллюстрирует такой подход, юридически закрепляя кантовское различие между моральными лицами и инструментальными вещами. Она отвечает на разрыв вменения посредством институционально обеспечиваемых мер: запрета на делегирование нравственного выбора, требования прозрачности и сертификации «доверенных технологий искусственного интеллекта» [5], а также четкого закрепления правовой ответственности за человеческими акторами. В рамках этой модели ИИ рассматривается строго как инструмент, чья функциональная автономность подчинена человеческой власти через риск-ориентированный надзор, этическую оценку и механизмы, обеспечивающие объяснимость и возможность оспаривания.

Таким образом, Стратегия предлагает последовательную модель управления гетерономными системами. Ее цель состоит в защите человеческого морального сообщества и той автономии, достоинства и ответственности, на которых оно основано, посредством обеспечения того, чтобы искусственный интеллект оставался направляемым человеческим суждением и подотчетным ему.

ЛИТЕРАТУРА

[1] Быльева С.Б., Нордманн А. Онтолитический эффект искусственного интеллекта. *Philosophy Journal of the Higher School of Economics*, 2025, № 4, с. 84–104.

[2] Труфанова Е.О. Надежность и ответственность как ключевые вопросы применения ИИ-систем. *Philosophy Journal of the Higher School of Economics*, 2025, № 9, с. 105–122.

[3] Чэн П., Чжан Ч. Механизм формирования ответственности и логика этического управления в воплощенном искусственном интеллекте. *Philosophy Journal of the Higher School of Economics*, 2025, № 4, с. 123–151.

[4] Munn L. The uselessness of AI ethics. *AI Ethics*, 2023, no. 3, pp. 869–877.

[5] Указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации». *Президент России.* URL: http://www.kremlin.ru/acts/bank/44731 (дата обращения 16.12.2025).

[6] Mahlmann M. Theoretische Philosophie. *Rechtstheorie*. URL: https://www.rwi.uzh.ch/elt-lst-mahlmann/rechtstheorie/kant/de/html/u2_lo2_1.html (дата обращения 16.05.2026).

[7] Шиян А.А., Петров В.Б. Трансцендентальный поворот в современной философии — VIII: метафизика, эпистемология, трансцендентальная когнитивистика и искусственный интеллект. *Вестник Российского университета дружбы народов. Серия: Философия*, 2023, № 4, с. 1033–1041.

[8] Kant I. *Kritik der Urteilskraft*. Leipzig, Verlag von Felix Meiner, 1922, 394 p.

[9] Kant I. *Kritik der reinen Vernunft*. Leipzig, Leopold Voss, 1868, 619 p.

[10] Sensen O. *Kant on moral autonomy*. New York, Cambridge University Press, 2013, 311 p.

[11] Kant I. *Groundwork of the Metaphysics of Morals*. Cambridge, Cambridge University Press, 1998, 77 p.

[12] Abakare C.O. Kantian Ethics and the Hesc Research: A Philosophical Exploration. *Predestinasi*, 2021, no. 2, pp. 79–92.

[13] Козлов С.С., Пирогов П.П. *Римское право*. Мурманск, Мурманский арктический университет, 2024, 238 с.

[14] Dennett D.C. *Kinds Of Minds: Toward An Understanding Of Consciousness*. New York, Basic Books, 2015, 184 p.

[15] Eslambolchilar P., Bødker M., Chamberlain A. Ways of Walking: Understanding Walking's Implications for the Design of Handheld Technology via a Humanistic Ethnographic Approach. *Human Technology*, 2016, no. 1, pp. 5–30.

[16] Статья 2. Основные понятия, используемые в настоящем Федеральном законе. *КонсультантПлюс.* URL: https://www.consultant.ru/document/cons_doc_LAW_351127/c5051782233acca771e9adb35b47d3fb82c9ff1c/ (дата обращения 21.12.2025).

[17] Kant I. What is Enlightenment? *Columbia.edu.* URL: https://www.columbia.edu/acis/ets/CCREAD/etscc/kant.html (дата обращения 16.09.2025).

Статья поступила в редакцию 13.07.2026

Ссылку на эту статью просим оформлять следующим образом:

Багдасарьян Н.Г., Бодрина М.С. Искусственный интеллект и кризис ответственности в призме кантовского императива. *Гуманитарный вестник*, 2026, вып. 4. EDN VDSNDF

Багдасарьян Надежда Гегамовна — д-р филос. наук, профессор, заслуженный работник высшей школы Российской Федерации, профессор кафедры «Социология и культурология» МГТУ им. Н.Э. Баумана. e-mail: ngbagda@mail.ru

Бодрина Маргарита Сергеевна — магистрант кафедры «Информационное обеспечение внешней политики» факультета мировой политики МГУ имени М.В. Ломоносова. e-mail: bodrina1010@mail.ru

Artificial Intelligence and the Crisis of Responsibility through the Lens of Kant's Imperative

© N.G. Bagdasaryan[1], M.S. Bodrina[2]

[1] Bauman Moscow State Technical University, Moscow, 105005, Russia
[2] Lomonosov Moscow State University, Moscow, 119991, Russia

An analysis of the crisis of responsibility in the processes of artificial intelligence application is presented. It is shown that the exacerbation of long-standing problems of responsibility gives rise to what can be designated as a "responsibility gap", in which significant actions and their consequences cease to be unequivocally attributable to conscious moral will. The complexity, ambiguity, and novelty of the problem of responsibility in the context of artificial intelligence require recourse to a reliable methodological foundation, which is provided by Immanuel Kant's categorical imperative. It is noted that Kant's moral philosophy — with its emphasis on autonomy, self-legislation, and the categorical imperative — provides a rigorous theoretical framework for explaining why artificial intelligence, as a heteronomous system, cannot be recognized as a moral agent. It is analyzed how the National Strategy for the Development of Artificial Intelligence in the Russian Federation responds to this problem by forming a legal and technical architecture designed to close this "Kantian gap". The Strategy is considered as a politico-legal framework that translates philosophical approaches into regulatory practice and offers ways to preserve human sovereignty and restore accountability in conditions shaped by non-moral, algorithmic systems. The study is based on a critical analysis of contemporary philosophical-legal, politico-philosophical, and regulatory approaches to artificial intelligence.

Keywords: *Kant, responsibility gap, ethics, responsibility, AI governance, trustworthy AI, Russia's AI strategy*

REFERENCES

[1] Bylieva S.B., Nordmann A. Ontoliticheskiy effekt iskusstvennogo intellekta [Ontolytic Effects of AI]. Filosofiya. *Zhurnal vysshey shkoly ekonomiki — Philosophy Journal of the Higher School of Economics*, 2025, no. 4, pp. 84–104.

[2] Trufanova E.O. Nadezhnost i otvetstvennost kak klyuchevye voprosy primeneniya II-sistem [Trustworthiness and Responsibility as the Key Issues of the AI Application]. *Filosofiya. Zhurnal vysshey shkoly ekonomiki — Philosophy Journal of the Higher School of Economics*, 2025, no. 9, pp. 105–122.

[3] Cheng P., Zhang C. Mekhanizm formirovaniya otvetstvennosti i logika eticheskogo upravleniya v voploshchennom iskusstvennom intellekte [The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence*]. Philosophy Journal of the Higher School of Economics*, 2025, no. 4, pp. 123–151.

[4] Munn L. The uselessness of AI ethics. *AI and Ethics*, 2023, no. 3, pp. 869–877.

[5] Ukaz Prezidenta Rossiyskoy Federatsii ot 10.10.2019 g. № 490 "O razvitii iskusstvennogo intellekta v Rossiyskoy Federatsii" [Decree of the President of the Russian Federation dated October 10, 2019 No. 490 "On the Development of Artificial Intelligence in the Russian Federation"]. *President of Russia*. Available at: http://www.kremlin.ru/acts/bank/44731 (accessed December 16, 2025).

[6] Mahlmann M. Theoretische Philosophie. *Rechtstheorie*. Available at: https://www.rwi.uzh.ch/elt-lst-mahlmann/rechtstheorie/kant/de/html/u2_lo2_1.html (accessed May 16, 2026).
[7] Shiyan A.A., Petrov V.B. Transtsendentalnyy povorot v sovremennoy filosofii — VIII: metafizika, epistemologiya, transtsendental'naya kognitivistika i iskusstvennyy intellekt [Transcendental Turn in Contemporary Philosophy — VIII: Metaphysics, Epistemology, Transcendental Cognitive Science and Artificial Intelligence]. *Vestnik Rossiyskogo universiteta druzhby narodov. Seriya: Filosofiya — RUDN Journal of Philosophy*, 2023, no. 4, pp. 1033–1041.
[8] Kant I. *Kritik der Urteilskraft* [Critique of Judgment]. Leipzig, Verlag von Felix Meiner, 1922, 394 p.
[9] Kant I. *Kritik der reinen Vernunft* [Critique of Pure Reason]. Leipzig, Leopold Voss, 1868, 619 p.
[10] Sensen O. *Kant on moral autonomy*. New York, Cambridge University Press, 2013, 311 p.
[11] Kant I. *Groundwork of the Metaphysics of Morals*. Cambridge, Cambridge University Press, 1998, 77 p.
[12] Abakare C.O. Kantian Ethics And The Hesc Research: A Philosophical Exploration. *Predestinasi*, 2021, no. 2, pp. 79–92.
[13] Kozlov S.S., Pirogov P.P. *Rimskoe pravo* [Roman law]. Murmansk, Murmansk Arctic University Press, 2024, 238 p.
[14] Dennett D.C. *Kinds of Minds: Toward an Understanding of Consciousness*. New York, Basic Books, 2015, 184 p.
[15] Eslambolchilar P., Bødker M., Chamberlain A. Ways of Walking: Understanding Walking's Implications for the Design of Handheld Technology Via a Humanistic Ethnographic Approach. *Human Technology*, 2016, no. 1, pp. 5–30.
[16] Statiya 2. Osnovnye ponyatiya, ispol'zuemye v nastoyaschem Federalnom zakone [Article 2. Basic concepts used in this Federal Law]. *KonsultantPlyus*. Available at: https://www.consultant.ru/document/cons_doc_LAW_351127/c5051782233acca771e9adb35b47d3fb82c9ff1c/ (accessed December 21, 2025).
[17] Kant I. What is Enlightenment? *Columbia.edu*. Available at: https://www.columbia.edu/acis/ets/CCREAD/etscc/kant.html (accessed September 16, 2025).

Bagdasaryan N.G., Dr. Sc. (Philosophy), Professor, Honored Worker of Higher Educati on of the Russian Federation, Professor, Department of Sociology and Cultural Studies, Bauman Moscow State Technical University. e-mail: ngbagda@mail.ru

Bodrina M.S., Master's Student, Department of Information Support for Foreign Policy, Faculty of World Politics, Lomonosov Moscow State University.
e-mail: bodrina1010@mail.ru