# ПРИГЛАШЕНИЕ К ДИСКУССИИ

DOI: 10.19181/socjour.2025.31.1.5
EDN: NRRIYS

***А.В. АГАНОВА*[1,2], *И.В. КАТЕРНЫЙ*[3,4]**

[1] Национальный исследовательский университет «Высшая школа экономики».
101000, Москва, ул. Мясницкая, д. 11.
[2] Фонд «Общественное мнение» (ФОМ).
123376, Москва, ул. Рочдельская, д. 15, стр. 16А.
[3] МГИМО МИД России.
119454, Москва, проспект Вернадского, д. 76.
[4] Институт социологии ФНИСЦ РАН.
109544, Москва, ул. Большая Андроньевская, д. 5. стр.1.

# МОРАЛЬНЫЙ СТАТУС ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: В ПОИСКАХ СОЦИОЛОГИЧЕСКОЙ ПЕРСПЕКТИВЫ

*Аннотация.* В статье рассматривается проблематика моральной агентности и моральной пациентности искусственного интеллекта (ИИ). В первой части статьи дается обзор основных теоретических аргументов, а также некоторых экспериментальных данных относительно (не)возможности наделения машин моральным статусом. Демонстрируется, что антропоморфизм и анимизм в восприятии ИИ, будучи культурно обусловленными, в значительной мере влияют на моральное признание искусственного интеллекта. Во второй части статьи выделяются основные исследовательские векторы, представляющиеся перспективными для рассмотрения сквозь социологическую призму, в частности: (а) фокусировка на конкретных контекстах взаимодействия с ИИ; (б) поиск социально-культурных различий в приписывании моральной агентности и пациентности, а также в моральных ожиданиях от ИИ, в восприятии проблем, связанных с его использованием; (в) социологическая экспертиза проблем разработки этического дизайна; (г) проблема делегирования морально-релевантных решений ИИ. В завершение статьи проблема морального статуса ИИ моделируется через стратификацию интегрированной моральной ответственности за социальные последствия в человеко-машинных взаимодействиях, в рамках которой выделяются несколько уровней, демонстрирующих усиление морального непризнания ИИ с ростом его автономности и разумности.

*Ключевые слова:* искусственный интеллект; моральная агентность; моральная пациентность; мораль; социология морали.

**Для цитирования:** *Аганова А.В., Катерный И.В.* Моральный статус искусственного интеллекта: в поисках социологической перспективы // Социологический журнал. 2025. Том 31. № 1. С. 92–109. DOI: 10.19181/socjour.2025.31.1.5 EDN: NRRIYS

## Введение

Как известно, согласно деонтологической этике И. Канта мораль дана человеку не по природе, а благодаря уникальному сочетанию в нем разума и воли. Это делает человека способным не только самоподчиняться долгу, но и относиться к себе, «равно как и всякому разумному существу», как к цели и никогда — как к средству [7, с. 530]. Идеи автономности и рациональности легли в основу всех последующих прокантианских представлений о предельном моральном статусе человека как источнике всякой нормативности [39]. Однако современные технологические разработки в области искусственного интеллекта создают фронтирную зону коммуникаций, чувствительных к моральным обязательствам, где автономность и рациональность более не являются сугубо человеческой привилегией. Характерно, что сам Кант видел в человеке лишь эмпирическое проявление «разумного естества», выступающего как априорное начало императивных нравственных принципов независимо от человеческой природы как таковой [8, с. 182].

Искусственный интеллект (далее — ИИ) за последние полвека очень быстро прошел путь развития от простого калькулятора до сложных нейросетей и робототехнических систем, приближающихся по своим возможностям к индивидуальной человеческой воле и человеческому мышлению (и в определенных сферах превосходящих их). Способность к (само)обучению и адаптации, автономное принятие решений, понимание контекста, обработка естественного языка и другие когнитивные функции приближают ИИ к статусу кантовского «разумного существа» как носителя моральных качеств и производителя моральных последствий. Уже сегодня искусственные агенты делают научные открытия на уровне Нобелевской премии [52], все чаще используются специалистами в качестве источников для вынесения вердиктов о жизни и здоровье людей [33], к ним обращаются за житейским советом и эмоциональной поддержкой [36]. Появляется возможность делегирования не только рутинных практик, но и действий, связанных со сложными управленческими процессами, что существенно расширяет пределы включения искусственных агентов в область институциональных коммуникаций [26]. При этом моральные дилеммы, возникающие перед нечеловеческими акторами, практически никогда не могут быть решены с помощью простых алгоритмических правил [17]. В ситуациях, когда ИИ функционирует с относительной автономией, а сама моральная дилемма не может быть решена однозначным образом, оказывается неясным, кто в конечном счете будет виновен, если произойдет несчастный случай или нарушение прав человека [47].

Включение ИИ в область социальных взаимодействий, влекущих морально-релевантные последствия, актуализирует вопрос о восприятии искусственных акторов в качестве субъектов и объектов морально-релевантных действий (то есть их моральной агентности и моральной пациентности соответственно [12]), другими словами — вопрос восприятия их морального статуса. Между тем проблематика

моральной агентности и пациентности напрямую связана с понятием морального сознания. Это одна из центральных категорий социальной теории, однако внимание к ней снизилось в связи с общим кризисом социологии морали во второй половине XX в. В последние десятилетия предпринимаются попытки ревитализации социологического подхода к изучению моральных явлений [4; 5], но в рамках т. н. новой социологии морали проблематика человеко-машинной коммуникации практически не затрагивается. Однако, на наш взгляд, социологический подход к изучению коммуникации «человек – машина» способен концептуально насытить имеющиеся эмпирические данные в этой области, а кроме того, расширить и саму область изучения морального сознания за счет рассмотрения качественно новых социальных процессов [12].

Таким образом, актуальным становится вопрос морального стратифицирования ИИ[1] в континууме от неодушевленной вещи до самосознающего субъекта как носителя полного морального статуса и тех социальных последствий, к которым ведет это признание. Цель статьи — систематизация основных аргументов в дискуссии о моральном статусе ИИ (разворачивающейся прежде всего в философской традиции), а также обзор некоторых экспериментальных данных (источником которых являются главным образом когнитивные науки) относительно восприятия моральной агентности и пациентности ИИ. Это позволит обозначить потенциально перспективные векторы для социологического изучения проблематики.

## Моральный статус ИИ: обзор междисциплинарной теоретической дискуссии и экспериментальных данных

### *Моральная агентность ИИ*

Хотя большинством специалистов все еще признается, что даже самые продвинутые формы ИИ не обладают моральным статусом, сравнимым с человеческим, теоретические подходы к вопросу наделения ИИ моральной агентностью пестрят разнообразием. Согласно самой «мягкой» позиции, для признания актора моральным агентом ему достаточно быть способным совершать морально-релевантные действия (которые, в свою очередь, заключаются в том, чтобы как минимум избегать нанесения вреда) [26]. Авторы, придерживающиеся этой линии, полагают: развитие ИИ требует смены парадигмы в том, что касается критериев наделения акторов моральным статусом. Здесь нужно основываться не на вопросе, какие онтологические свойства присущи машине, а на ее перформативных способностях и на том, какие типы отношений «человек – машина» возможны [41]. Иными словами, вопрос морального статуса ИИ предлагается рассматривать сквозь призму бихевиористского подхода.

В связи с этим одним из наиболее дискуссионных оснований наделения ИИ моральной агентностью выступает интенциональность. Исследователи, апеллирующие к наиболее радикальной точке зрения — невозможности отнесения

[1] В этой статье ИИ будет определяться как область всех человеко-машинных коммуникаций, основанных на имитации человеческого поведения и мышления.

нечеловеческих акторов к категории моральных агентов, — отмечают, что условием интенциональности является наличие сознания [25]. По мысли же других авторов, которые придерживаются более «мягкой» позиции, не стоит рассматривать видимые последствия действий машин исключительно как функцию от интенциональности пользователей, разработчиков и т. д., поскольку таким образом игнорируются эмерджентные способности ИИ. Его самообучаемость, согласно этой точке зрения, требует новой концептуальной основы, признающей определенную свободу действий и «прединтенциональность» [45].

Не менее дискуссионно и другое основание, напрямую связанное с интенциональностью, — ответственность [36]. В качестве одного из аргументов против наделения машины ответственностью (помимо собственно отсутствия интенциональности) приводят невозможность обладания моральным авторитетом [50]. Однако этот тезис оказывается спорным: именно благодаря непрозрачности, функционированию по типу «черного ящика», алгоритмы воспринимаются как более объективные и безошибочные (и тем самым порой незаметно воспроизводят этически спорные решения) [41].

Особое внимание в этой связи получила проблема так называемого «пробела ответственности» (*responsibility gap*), который можно определить как сложности в приписывании и распределении ответственности в ситуациях, когда действия автономных машин повлекли морально-релевантные последствия [32]. В таких случаях остается неясным, где сосредоточивается ответственность при нанесении вреда: лежит ли она на самой системе ИИ (или же на другой технической части, к ИИ не относящейся), на пользователе (который продолжает эксплуатировать систему, несмотря на непонимание принципов ее функционирования), на менеджерах или на разработчиках; эта ответственность индивидуальна, институциональна или распределена между всеми акторами (и если да, то в каких долях) [28; 35]. Разумеется, сложности с распределением ответственности между разными акторами характерны не только для контекстов, где одним из акторов является машина. Однако особенность «пробела ответственности» в случае коммуникации «человек – машина» заключается в том, что вследствие (а) антропоморфизации ИИ и одновременно (б) восприятия алгоритмов как объективных и безошибочных, (в) сложности с отслеживанием цепочки действий, (г) потенциальной эмерджентности поведения машин возникает возможность снятия ответственности (намеренного или ненамеренного) с человеческих акторов [33; 53]. Кроме того, определение ответственных за причинение вреда сторон в значительной мере зависит от уровня абстрактности интерфейса, делающего одни стороны жизненного цикла ИИ видимыми, а другие — невидимыми [51].

Между тем в контексте разработки искусственных агентов, которым может быть делегирована ответственность в морально-релевантных ситуациях, приобретает значение то, в каких дискурсивных рамках разворачивается обоснование необходимости таких разработок. Как правило, наиболее представленными оказываются консеквенциалистские аргументы: например, в дискуссии о беспилотных автомобилях чаще всего звучит тезис о возможности спасать больше человеческих жизней в дорожных происшествиях. Однако подобные обыденные оценки моральной приемлемости, игнорирующие асимметрию (не)благоприятных последствий,

могут быть достаточно упрощенными, становясь тем самым фактором социального неравенства (беспилотные автомобили, не запрограммированные на приоритетность жизни пассажира, скорее всего, окажутся просто невостребованными) [40].

Наконец, если говорить о моральной агентности машин, то выделяются и ряд других оснований, согласно которым ИИ потенциально можно отнести к моральным агентам, — интерактивность, адаптивность и автономность [35]. Причем критерию автономности уделяется довольно пристальное внимание: утверждается, например, что в случае ИИ трактовка автономности как свободы воли неадекватна, поэтому предлагается интерпретировать ее бихевиористски — как способность реагировать на стимулы окружающей среды. Другие интерпретации включают также способность к (относительному) самоуправлению и самоконтролю. В конечном счете автономность машин предлагается определять как способность достигать поставленных целей и воздействовать на окружающую среду в течение некоторого времени без внешних интервенций, а также изменять свои собственные действия (хотя только в предзаданных рамках) [23]. Впрочем, критерий автономности оказывается проблематичным в практической плоскости: экспериментальный опыт демонстрирует, что человек в момент взаимодействия с роботом не способен определить, действует ли тот автономно или управляется оператором [1].

### *Моральная пациентность ИИ*

В отличие от вопроса моральной агентности, проблеме наделения машин моральной пациентностью посвящено не очень много исследований. В частности, предпринимаются попытки разработки измерительных шкал для экспериментальных планов (напр.: [20]). Теоретическую дискуссию по данной проблеме едва ли можно назвать насыщенной. При этом внимание уделяется не столько реконцептуализации понятия моральной пациентности применительно к человеко-машинному взаимодействию (например, обсуждается, должны ли нечеловеческие акторы отвечать антропоцентричным критериям [19]), сколько вопросу о принципиальной возможности отнесения машин к категории моральных пациентов.

В отношении последней проблемы просматриваются два лейтмотива. Во-первых, поднимается вопрос: как при отсутствии способности к субъективному переживанию объект может быть конечным адресатом блага или вреда? По мнению некоторых авторов, сознание не является условием моральной пациентности, а некую сущность можно наделять моральной пациентностью постольку, поскольку она является телеологической системой [42]. Во-вторых, отмечается, что «конечным» моральным пациентом в человеко-машинной коммуникации оказывается сам человек. Последнее аргументируется конструированием социальных роботов преимущественно антропоморфными, что наделяет машину символическим значением [34]. Эксперименты показывают, что насилие по отношению к антропоморфному роботу оценивается так же, как и физический буллинг в адрес человека [21]. Впрочем, эта связь не столь очевидна: например, ролики с нанесением физического вреда антропоморфному роботу вызывают куда больше негативных эмоций, чем ролики с убийством человеческого персонажа компьютерной игры [40] (вероятно, в этом случае на восприятие влияет физическая воплощенность).

Важно также учитывать, что антропоморфизм как фактор морального признания имеет насыщенное культурное происхождение и не является единственно возможным паттерном продвинутой коммуникации с ИИ. В 2023 г. появились результаты кросс-культурного исследования морального отношения к роботам в Японии и США, где подчеркивается особое эмоционально насыщенное восприятие искусственных агентов японцами за счет культурных паттернов анимизма [38]. На то существуют как минимум две базовые причины: культурная и социальная. В контексте культуры повсеместное увлечение робототехникой поддерживается традиционными идеями синто, согласно которым дух Будды живет во всех вещах, включая механические. Если на Западе пугают злобным и бездушным Терминатором, то в Японии робот по своей «духовной» сути ничем не отличается от живого человека. Эмоциональная привязанность к домашнему роботу может соперничать с дружбой с реальным человеком, а похоронить по всем правилам вдруг сломавшуюся любимую роботизированную собачку Айбо не будет считаться отклонением от нормы. Что касается социальной причины, то высокая продолжительность жизни вкупе с низкой рождаемостью и почти нулевой иммиграцией диктует необходимость развития отрасли, способной массово производить эмоциональных роботов-сиделок для пожилых японцев. Все это создает особый коммуникативный паттерн как моральной агентности роботов, так и их моральной пациентности в форме человеческой заботы, эмоциональной привязанности и даже духовной связи с человеком.

Насыщенный моральный статус ИИ как коммуникативного партнера все больше просматривается в активном внедрении в повседневную жизнь встроенных голосовых помощников (в навигаторах, телефонах, умных колонках, умных домах) и виртуальных операторов. Общение с подобными собеседниками создает двойственную ситуацию сопряжения человеческой и нечеловеческой онтологий, что способно вызывать морально значимые эффекты. Когда компания Google представила в 2018 г. голосовую технологию Duplex, способную самостоятельно звонить абоненту и разговаривать неотличимым от человека образом (меняя интонации, делая паузы, запинаясь и даже делая оговорки), эксперты посчитали подобную антропоморфизацию неэтичной и даже жуткой [43], что заставило компанию-разработчика ввести правило для робота: обязательно представляться в начале любого разговора. Впоследствии доступ к этой технологии и вовсе ограничили. В целом же антропоморфность влияет на восприятие роботов нелинейным образом: чувство аффекта и отвращения, вызываемое чрезмерным антропоморфным реализмом, получило название «зловещей долины» — uncanny valley (существуют и другие вариации этого феномена — «зловещий обрыв» (uncanny cliff) или «зловещая стена» (uncanny wall) [9].

Отечественные исследования также показывают, что ставшие сегодня обыденными разговоры с роботами-операторами по телефону эффективны, только когда абонент понимает, что разговаривает с ИИ, и адаптивно смешивает паттерны общения «как с человеком» и «как с машиной», тем самым рефлексивно помогая преодолеть онтологический разрыв и достичь целей коммуникации [11]. Активное социальное вовлечение голосовых роботов в повседневную жизнь наиболее интенсивно ведет к росту их воспринимаемой моральной агентности и пациентности. Все большая рутинизация коммуникативных ситуаций с участием машин неиз-

бежно очеловечивает ИИ, наделяя его моральными качествами и ожиданиями, которые могут восприниматься как положительно, так и отрицательно. Например, способность умных колонок всегда быть включенными и слышать разговоры домочадцев включает паттерны поведения, характерные для присутствия чужих людей [10]. Пользователи могут избегать приватных разговоров в присутствии колонки или стараться не разглашать чувствительную информацию. В то же время необходимость произнесения имени для активации умной колонки (так называемое *wake-word*) укрепляет восприятие голосового помощника как обладающего собственной личностью. Некоторые, говоря о своих помощниках, часто прибегают к гендерным местоимениям («она»). При этом те пользователи, которые называют голосовых помощников личными местоимениями, с большей вероятностью используют их для ведения бесед, в то время как другие, обозначающие помощников как объект («оно»), больше склонны воспринимать их как технологический девайс, обеспечивающий определенный функционал. Кроме того, голосовые ассистенты используются, как правило, в приватном пространстве дома, что также побуждает относиться к ним как к партнерам по взаимодействию. В представлении пользователя голосовые помощники также могут играть разнообразные роли — не только партнера по взаимодействию, но и друга, личного помощника, слуги или эксперта [27]. Разговорные сюжеты заигрываний, сексуальные оскорбления и даже домогательства также возможны в ситуации феминизации коммуникативного интерфейса ИИ [13]. Таким образом, способ категоризации голосового помощника, степень его антропоморфизации имеют некоторую соотнесенность со стратегиями его использования [27].

## ИИ в морально-релевантных контекстах: перспективы социологического подхода

По словам С. Хитлин и С. Вейзи, область интереса новой социологии морали охватывает не только нормы и ценности, но и нарративы, идентичности, институты, символические границы и когнитивные схемы, социальные и исторические различия в восприятии морального, а также социальные процессы, создающие и поддерживающие определенные концепции морали [37].

Между тем изучение приписывания моральных метакачеств ИИ в настоящее время является прежде всего стезей когнитивных наук (впрочем, как и исследования морали в целом). Это, в свою очередь, приводит к фокусировке на экспериментальных планах, оставляя без должного внимания конкретные контексты взаимодействия. С одной стороны, представляется, что в условиях бурного развития ИИ (а значит, и меняющихся практик) такая фокусировка неизбежно будет приводить к неполноте знаний о приписывании моральных метакачеств ИИ. С другой — в условиях продолжающейся институционализации новой социологии морали невнимание к актуальным проблемам несет для нее риск оказаться за бортом научной дискуссии [4].

В связи с этим можно выделить несколько векторов, потенциально перспективных для рассмотрения проблемы морального статуса ИИ через социологическую призму.

Прежде всего, восприятие моральных метакачеств ИИ может варьироваться в зависимости от контекста взаимодействия (например, в случае согласия/несогласия с решениями машины, при относительно беспроблемном протекании взаимодействия или в случае существенных «поломок», в приватной или публичной обстановке и т. д. [18]). Без внимания в настоящее время остается и поиск социальных различий — не только в приписывании агентности и пациентности, но и в восприятии моральной способности ИИ и моральных проблем, связанных с его использованием (или, напротив, в восприятии преимуществ делегирования морально-релевантных решений). В частности, интерес может представлять различие между обыденным и профессиональным знаниями о работе алгоритмов (например, представления об их объективности, стратегии объяснения принятых алгоритмом морально-релевантных решений и восприятие источника этих решений). Проблема агентности и пациентности ИИ трактуется профессионалами как вопрос экспертной разработки ценностно-сенситивного дизайна и нормативного регулирования инструментов машинного обучения и больших данных для эффективной имплементации соответствующих этических кодексов («Кодекс этики в сфере ИИ», «Сбер: Принципы этики ИИ», “Ethics guidelines for trustworthy AI”, “Montréal Declaration for Responsible Development of Artificial Intelligence”, “Principles for Accountable Algorithms and a Social Impact Statement for Algorithms” и др.). Кодификация моральной повестки также сталкивается здесь с проблемами ответственности в контуре «общество – разработчики – собственники – искусственный интеллект». Требуют социологической экспертизы несколько взаимосвязанных проблем разработки «этического дизайна»: (а) универсальные гуманитарные проблемы (человеческая автономность и независимость, защита данных); (б) экспертный надзор (подконтрольность обществу); (в) ценностный детерминизм (борьба с имплементацией предрассудков, предубеждений, нелегального контента в ИИ); (г) корпоративизм как локус этического контроля (проблема ИИ как частной собственности больших корпораций); (д) социально ответственные проектирование и эксплуатация ИИ (соблюдение всех международных конвенций, прав человека в разработке и работе ИИ); (е) прозрачность технологии (информированность населения о рисках ИИ; например, допустимо ли не быть информированным о том, что медицинский или юридический совет был дан человеку чат-ботом).

Заслуживает внимания и дальнейший поиск социально-культурных различий в моральных ожиданиях от ИИ. Подтверждением этому служит известное глобальное онлайн-исследование Массачусетского технологического института 2016–2017 гг. «Моральная машина», которое собрало мнения почти 40 миллионов людей со всех уголков мира об их моральных предпочтениях в ситуации столкновения с различными вариантами т. н. проблемы вагонетки[2] [17]. Респондентам предоставлялась возможность морально взвесить жизни людей и животных, пас-

[2] «Проблема вагонетки» (англ. — *Trolley problem*) — мысленный этический эксперимент, согласно классической версии которого на пути неуправляемой вагонетки привязаны пять человек. Они погибнут, если вагонетка продолжит двигаться. Можно перевести стрелки, и в таком случае вагонетка поедет по второму пути, к которому привязан один человек. Предполагается, что ИИ должен научиться видеть моральные дилеммы подобного рода и принимать максимально взвешенное решение, например, в случае дорожной аварии беспилотного транспорта.

сажиров и пешеходов, женщин и мужчин, молодых и старых, стройных и не очень, представителей высшего и низшего классов, большее и меньшее количество жертв в ситуации жизни и смерти. В итоге были выявлены значительные гендерные, региональные, культурные и институциональные различия морального выбора. Но три вывода более или менее объединяют всех: предпочтение людей животным, молодых — людям старшего возраста и спасения большего количества жизней — меньшему. Хватит ли этих принципов для разработки экспериментальной этики ИИ, все еще неясно. По мнению исследователей, будущим производителям необходимо учитывать особенности местных моральных кодексов в тех регионах, где традиции накладывают сильный отпечаток на коллективные представления.

При рассмотрении разных моделей восприятия ИИ в морально-релевантных контекстах следует решать и задачу теоретического насыщения. Важным шагом в этом является переход от понятий, по Г. Абенду [16], «тонких» (например, приемлемость) к «насыщенным», имеющим культурные и институциональные предпосылки (например, справедливость). Логично предположить, что чувствительность к различным «насыщенным» моральным концептам может варьироваться в зависимости не только от контекста, но и от других переменных, включая такие ценности, как технооптимизм, технопатернализм и технопессимизм.

Что касается делегирования морально-релевантных решений ИИ, то экспериментальные данные демонстрируют: принятие машинами морально-релевантных решений вызывает интуитивное отторжение. Однако представление об алгоритмах как о более непредвзятых и экспертных по сравнению с людьми, по всей видимости, помогает подобное отторжение преодолевать [24]. Исходя из этого интерес может представлять то, какие факторы влияют на приемлемость и желательность делегирования морально-релевантных решений ИИ. Такими факторами могут выступать воспринимаемые символические выгоды, связанные с личной идентичностью «первопроходца» новых технологий [27] или же, например, восприятие интенциональности алгоритма и его способности к субъективному переживанию. Изучение этих факторов вновь должно учитывать контекстуальность: во-первых, можно предположить, что для тех ИИ, которые воплощают роль личного ассистента, подчеркнутое интеллектуальное превосходство машины может оказываться нежелательным. Во-вторых, при делегировании морально-релевантных решений приобретает значение то, где проходят границы допустимого делегирования (например, в зависимости от сенситивности контекста). В-третьих, делегирование морально-релевантных решений предполагает и изучение того, в каких дискурсивных рамках разворачивается обоснование необходимости подобных разработок (или, напротив, их опасности). И, в-четвертых, это рефлексивное восприятие человеком собственной агентности (в том числе в тех организационных контекстах, где алгоритмизация оказывается вынужденной). Здесь явственно проступает напряженность между повышением эффективности (а также символическими выгодами) и непрозрачностью алгоритмических архитектур выбора.

С учетом сказанного социологическое обоснование получает разработка модели моральной классификации ИИ на основе идеи иерархии морального статуса как реляционного феномена [30; 31; 46]. Эволюция моральной феноменологии в сфере зоозащиты, а также биоэтики показывает, что нечеловеческие субъекты и объекты могут располагаться на разных уровнях социального признания — от пренебрежимого

статуса (например, в качестве еды) до юридического приравнивания прав некоторых видов к правам человека. В этом смысле проблема моральной пациентности ИИ, в частности, может быть решена в рамках теории когнитивной эквивалентности, представленной Г. Шевлиным, где он сопоставляет когнитивный статус машин и животных с последующими моральными выводами. Это означает, что ИИ-систему следует считать «психологическим моральным пациентом» в той мере, в которой она обладает когнитивными механизмами, общими с другими субчеловеческими существами, каковых мы также считаем обладающими подобным статусом [48]. Однако непрозрачность подобных сравнений даже на экспертном уровне и нерешенность проблемы агентности заставляют искать более редуцированные подходы. Как показали эксперименты с буллингом, инкультурированная этика добродетели также требует, чтобы наше отношение к нечеловеческим объектам отвечало нормам «человечности» или «цивилизованности». Таким образом, проблема морального статуса ИИ может быть представлена как стратификация интегрированной моральной ответственности за социальные последствия в человеко-машинных взаимодействиях, в рамках которой можно выделить несколько уровней пациентности и агентности, демонстрирующих кумулятивный эффект моральной декогеренции ИИ, то есть усиление морального непризнания с ростом его автономности и разумности (см. рис.).


**Пирамида морального статуса (МС) искусственного интеллекта**

(1) Пациентный моральный статус ИИ получает как источник утилитарного блага и объект норм добродетели, что подразумевает непричинение ему вреда, необходимый уход и заботу со стороны человека как проявление морального патернализма по отношению к тому, что призвано служить человеку и повышать качество его жизни. Многочисленные свидетельства добровольной помощи прохожих застрявшим роботам-курьерам на улицах Москвы являются примером спонтанного поведения подобного рода [14]. Помимо сугубо деривативной ценности нечто, имеющее внутреннюю целостность, телеологию и динамические свойства («метаболизм»), претендует на ценность само по себе вне зависимости от того, живой это организм или искусственное создание, а значит, дополнительно отвечает требованиям минимального морально-пациентного статуса [22]. В то же время, как было показано, антропоморфизм или анимизм в восприятии воплощенного

ИИ сильно влияет на степень эмоционального отклика людей. Последующие уровни представленной пирамиды подразумевают не исключение, а надстраивание моральных статусов ИИ как превосходящих друг друга по степени агентности и одновременно усиливающих моральную декогеренцию.

(2) Пренебрежимый агентный моральный статус имеют все существующие модели ИИ, которые полностью основаны на программном коде и не способны к эмерджентному поведению. Операциональная моральная и юридическая ответственность здесь всегда возлагается на человеческих агентов — собственника, разработчика и оператора ИИ (в случае нарушения прав человека) либо пользователя (в случае эксплуатации ИИ ненадлежащим образом). В эту категорию входят почти все существующие модели ИИ, включая нейросети уровня System 1 и беспилотный транспорт вплоть до четвертого уровня автономности, однако проблема «пробела ответственности», неизбежно возникающая в данном случае, сама может иметь морально сомнительные решения. Характерно, что после пяти лет мучительного судебного разбирательства по делу о гибели пешехода от тестируемого автопилота компании Uber в 2018 г., все закончилось добровольным признанием вины тестировщика, сидевшего за рулем, а сама компания избежала наказания, хотя недоработки программного уровня были доказаны. То же самое произошло в случае ДТП с человеческими жертвами в 2019 г. с участием автопилота Tesla: ответственным был признан водитель, а не компания [49].

(3) Разделенный (с человеком) морально-агентный статус могут иметь те технические разработки, которым вменяется способность к автономному принятию решений, самообучению и функциональной интенциональности в контингентных (заранее непредусмотренных) условиях выбора. Передовые нейронные сети, различные экспертные системы в медицине, судебной системе, финансовом анализе, беспилотный транспорт пятого уровня автономности, а также действующие модели автономного беспилотного оружия подпадают под эту категорию, но имеют функционально ограниченную моральную чувствительность, поскольку их инструментальная эффективность превалирует над этичностью руководящих принципов. Многочисленные случаи дискриминации, несправедливых рекомендаций, нарушений частной жизни фиксируются в практике внедрения подобных систем по всему миру (подробнее см.: [15]). Уже на этом уровне проблема ИИ как «максимизатора скрепок» (Н. Бостром) не может быть решена эффективно сугубо средствами имплементации нормативных систем, а значит, ответственность и моральная подотчетность в таких формах человеко-машинного взаимодействия оказываются функционально эфемерными. Неспроста еще в 2015 г. С. Хокинг и более сотни других ученых подписали открытое письмо против исследований в области автономного летального оружия [2], а в 2023 г. уже более 30 тыс. человек подписали новое письмо с призывом заморозить обучение более продвинутых LLM-систем, нежели ChatGPT-4, из-за рисков моральной непрозрачности [44]. Тем не менее дальнейшая разработка профессиональных этических кодексов, например в логике «морали сотрудничества», допускает возможность создания более робастных и транспарентных протоколов функционирования ИИ с обязательным участием экспертной панели [6].

(4) На делегированный морально-агентный статус уже претендуют нейросети поколения System 2 или супер-ИИ (AlphaGo, Strawberry). В отличие от ИИ первого поколения, который лишь эмулировал когнитивные процессы на основе перебора

тысяч и миллионов готовых решений и информационного ассортимента, System 2 не симулирует автономность и рациональность, а создает принципиально новый опыт, обладая своеобразным копирайтом на производство решений. Супер-ИИ демонстрирует качества морально обязанного субъекта: память о собственном прошлом, самооценку, способность рассуждать и принимать решения, не основанные на регенерации шаблонов, а творческие, исходящие из ситуации здесь и сейчас. Подобные системы приобретают не только когнитивную свободу, но и морально значимые конечные цели (интерес к выживанию), как живые существа. Однако уровень их моральной компетентности остается «серой зоной», поэтому широкое применение нечеловеческого разума сталкивается с моральным парадоксом: чем больше машина похожа на человека, тем меньше у нее шансов на социальное признание. Различные конечные цели у двух равных моральных систем (человеческой и искусственной) могут вступать в противоречие друг с другом (моральная конкуренция), а общая инструментальная сходимость (преследование таких целей, как самозащита, самосовершенствование, поддержание полезности, расширение доступа к ресурсам и др.) делает такой конфликт почти неизбежным. Именно поэтому такие системы имеют (пока) строго ограниченное применение (игра в го, например).

(5) Полный моральный статус. Эту высшую страту моральной классификации всегда занимал человек, так как степени его автономности и рациональности, как предполагал Кант, уникальны и нормативны. Появление ИИ как морально обязанной и обязывающей личности, равной человеку, потенциально вступает в противоречие с самой идеей его моральной пациентности и имеет настолько непредсказуемые последствия, что возникает превентивная ответственность за недопущение появления подобных созданий. Дж. Баррат, автор книги с красноречивым названием «Последнее изобретение человечества», предлагает как можно скорее заключить специальное международное соглашение, предусматривающее программирование апоптоза (самоуничтожения) посттьюринговых компьютерных систем при достижении ими определенного уровня развития [3, с. 266–269]. Однако с социологической точки зрения развитие ИИ следует рассматривать как институциональный, а не сугубо технический вопрос. Поэтому запуск «сильного ИИ» как социотехнической структуры так или иначе будет сдерживаться механизмами социального морфостазиса, которые призваны отслеживать институционально значимые угрозы уже на уровне национальной безопасности, — так же, как это происходило в прошлом с ядерным оружием. Культурные, экономические, юридические ограничения тоже будут накладывать отпечаток на перспективы разработок в этой сфере, ведя к созданию скорее более слабых, но индигенизированных, то есть локальных, ценностно насыщенных, и при этом конкурирующих систем ИИ в различных регионах мира.

Важно также подчеркнуть, что «народное» описание когнитивного статуса продвинутого ИИ гораздо более склонно присваивать ему качества агентности, нежели экспертные оценки. Последние эксперименты показывают, что большинство людей считают, например, ChatGPT «имеющим опыт сознания» [29]. И чем интенсивнее «общение» с этой нейросетью у испытуемого, тем больше у него уверенности в этом. Случаи, когда ИИ в восприятии самих людей толкал их на попытки убийства и самоубийства, также имеются. Явное расхождение экспертных и «народных» оценок возможностей ИИ способно существенно повлиять на развитие дискурса о его моральном статусе в ближайшем будущем.

## Заключение

Таким образом, можно говорить о том, что проблематика приписывания ИИ моральных метакачеств, несмотря на свою обширность, имеет все же достаточно много лакун, которые довольно сложно закрыть инструментами поведенческих и когнитивных наук, в рамках которых она сейчас наиболее активно изучается. Представляется, что внимание к этой проблемной области со стороны социологии, включая социологию морали, способно серьезным образом ее расширить и обогатить. Несмотря на развитие преимущественно в смежных научных областях, проблематика морального статуса ИИ является безусловно социологической ввиду лиминальности социального положения искусственных акторов и вариабельности социально-культурных и институциональных контекстов, стоящих за восприятием их в качестве моральных агентов и пациентов. Это позволяет относить понятие морального статуса к «насыщенным» концептам (в то время как в когнитивных науках в фокусе рассмотрения находятся «тонкие» представления [16]). В условиях моральной декогеренции ИИ, то есть усиления его морального непризнания с ростом автономности и разумности, основой социологически значимого понимания проблемы морального статуса ИИ оказываются такие факторы, как «народные» представления об ИИ, доминирующие ценностно-нормативные комплексы, а также риски институциональной безопасности. Пирамида морального статуса ИИ тем не менее демонстрирует, что перспектива наделения искусственных агентов социально и юридически значимыми правами как «лиц нечеловеческой природы» (по примеру защиты прав животных) вполне возможна в недалеком будущем на основе сохранения нормативности антропоцентрической этики. Наконец, социологический взгляд на проблематику морального статуса ИИ, по нашему мнению, не только углубляет обоснование этической экспертизы разработки систем ИИ, но и позволяет внести вклад в понимание человеческой моральной способности в целом.

### *СВЕДЕНИЯ ОБ АВТОРАХ*

**Аганова Анастасия Вячеславовна** — аспирант, Аспирантская школа по социологическим наукам, Национальный исследовательский университет «Высшая школа экономики»; старший аналитик, Фонд «Общественное мнение» (ФОМ). **Телефон**: +7 (495) 772-95-90, доб. 12454. **Электронная почта**: avaganova@hse.ru

**Катерный Илья Владимирович** — доктор социологических наук, профессор, кафедра социологии, МГИМО МИД России; ведущий научный сотрудник, Институт социологии ФНИСЦ РАН. **Телефон**: +7 (495) 225-40-89. **Электронная почта**: yarkus@mail.ru

**Sotsiologicheskiy Zhurnal = Sociological Journal. 2025. Vol. 31. No. 1. P. 92–109. DOI:** 10.19181/socjour.2025.31.1.5

Research Article

***ANASTASIA V. AGANOVA[1,2], ILYA V. KATERNYI[3,4]***

[1] HSE University.
11, Myasnitskaya st., 101000, Moscow, Russian Federation.

[2] Foundation “Public Opinion” (FOM).
15, bl. 16A, Rochdelskaya st., 123376, Moscow, Russian Federation.
[3] MGIMO-University.
76, Vernadskogo avenue, 119454, Moscow, Russian Federation.
[4] Institute of Sociology of FCTAS RAS.
5, bl. 1, Bolshaya Andronievskaya st., 109544, Moscow, Russian Federation.

## The Moral Status of Artificial Intelligence: in Search of a Sociological Perspective

*Abstract.* The article examines AI from the perspective of moral agency and moral patiency. In the forepart, an overview of the main theoretical arguments as well as some experimental data regarding the (im)possibility of grounding the moral status of intelligent machines is provided. Evidence shows that perceived anthropomorphism and animism, this being culturally dependent, significantly influence the moral recognition of AI. The second part of the article highlights some research perspectives that seem promising for sociology of morality to be applied to the AI field: (a) discrimination of moral situations specific to interaction with AI, (b) discovering the cultural background of moral agency and patiency of AI, (c) expertise in AI code of ethics development, (d) inquiry into AI value alignment when delegating morally relevant decisions. The remainder of the article interprets moral status of AI as a stratification of integrated moral responsibility for social consequences in human-machine interactions and represents a pyramid within which several layers are distinguished, demonstrating the tendency to link the moral ‘non-recognition’ of AI with enhancing its autonomy and ‘consciousness’.

*Keywords:* AI; moral agency; moral patiency, morality; sociology of morality.

**For citation:** Aganova, A.V., Katernyi, I.V. The Moral Status of Artificial Intelligence: in Search of a Sociological Perspective. *Sotsiologicheskiy Zhurnal = Sociological Journal*. 2025. Vol. 31. No. 1. P. 92–109. DOI: 10.19181/socjour.2025.31.1.5

*INFORMATION ABOUT THE AUTHORS*

**Anastasia V. Aganova** — Postgraduate Student, Doctoral School of Sociology, HSE University; Senior Analyst, Foundation “Public Opinion” (FOM). **Phone**: +7 (495) 772-95-90, ext. 12454. **E-mail**: avaganova@hse.ru
**Ilya V. Katernyi** — Doctor of Sociological Sciences, Professor, Department of Sociology, MGIMO-University; Leading Researcher, Institute of Sociology of FCTAS RAS. **Phone**: +7 (495) 225-40-89. **E-mail**: yarkus@mail.ru

## ЛИТЕРАТУРА / REFERENCES

1. *Абрамов Р.Н., Катечкина В.М.* Социальные аспекты взаимодействия человека и робота: опыт экспериментального исследования // Журнал социологии и социальной антропологии. 2022. Т. 25. № 2. С. 214–243. DOI: 10.31119/jssa.2022.25.2.9 EDN: ASRIAF

   Abramov R.N., Katechkina V.M. Social Aspects of Human-Robot Interaction: Experience of Experimental Research. *Zhurnal sotsiologii i sotsialnoi antropologii*. 2022. Vol. 25. No. 2. P. 214–243. DOI: 10.31119/jssa.2022.25.2.9 (In Russ.)

2. Автономное оружие: открытое письмо исследователей ИИ и роботов, 01.09.2017 // Future of Life Institute [электронный ресурс]. Дата обращения 15.03.2020. URL: https://futureoflife.org/open-letter-autonomous-weapons-russian/

   Open Letter on Autonomous Weapons, 01.09.2017. *Future of Life Institute*. Accessed 15.03.2020. URL: https://futureoflife.org/open-letter-autonomous-weapons-russian/ (In Russ.)

3. *Баррат Дж.* Последнее изобретение человечества: Искусственный интеллект и конец эры Homo sapiens / Пер. с англ. М.: Альпина нон-фикшн, 2015. — 304 с.

   Barrat J. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Transl. from Eng. Moscow: Alpina non-fikshn publ., 2015. 304 p. (In Russ.)

4. *Быков А.В.* Когнитивная и аналитическая перспективы в новой социологии морали: основания различения и ключевые особенности // Социологический журнал. 2024. Т. 30. № 1. С. 26–42. DOI: 10.19181/socjour.2024.30.1.2 EDN: FCOYMR

   Bykov A.V. The New Sociology of Morality: Cognitive and Analytical Perspectives. *Sotsiologicheskiy Zhurnal = Sociological Journal*. 2024. Vol. 30. No. 1. P. 26–42. DOI: 10.19181/socjour.2024.30.1.2 (In Russ.)

5. *Быков А.В.* Понятие морального сознания в социологической традиции // Социологический журнал. 2017. Т. 23. № 3. С. 26–43. DOI: 10.19181/socjour.2017.23.3.5362 EDN: ZIPDET

   Bykov A.V. The Concept of Moral Conscience in Sociological Tradition. *Sotsiologicheskiy Zhurnal = Sociological Journal*. 2017. Vol. 23. No. 3. P. 26–43. DOI: 10.19181/socjour.2017.23.3.5362 (In Russ.)

6. *Девятко И.Ф.* Проблема ориентации искусственного интеллекта на человеческие ценности (AI value alignment) и социология морали // Социологические исследования. 2023. № 9. С. 16–28. DOI: 10.31857/S013216250027775-5 EDN: QVKPLQ

   Devyatko I.F. AI Value Alignment and Sociology of Morality. *Sotsiologicheskie issledovaniya*. 2023. No. 9. P. 16–28. DOI: 10.31857/S013216250027775-5 (In Russ.)

7. *Кант И.* Критика практического разума / Пер. Н.М. Соколова // Кант И. Сочинения. В 8 т. / Под общ. ред. А.В. Гулыги. М.: ЧОРО, 1994. Т. 4. С. 373–479.

   Kant I. Critique of Practical Reason. Transl. by N.M. Sokolov. Kant I. *Works. In 8 Vols.* Ed. by A.V. Gulyga. Moscow: ChORO publ., 1994. Vol. 4. P. 373–479. (In Russ.)

8. *Кант И.* Основоположения метафизики нравов / Пер. Л.Д.Б. // Кант И. Сочинения. В 8 т. / Под общ. ред. А.В. Гулыги. М.: ЧОРО, 1994. Т. 4. С. 153–246.

   Kant I. Groundwork of the Metaphysics of Morals. Transl. by L.D.B. Kant I. *Works. In 8 Vols.* Ed. by A.V. Gulyga. Moscow: ChORO publ., 1994. Vol. 4. P. 153–246. (In Russ.)

9. *Катерный И.В.* Каузальные объяснения эффекта «зловещей долины» в робототехнике: теории и исследовательские данные // Качество и жизнь. 2017. № 4. С. 88–96 EDN: YOPGDD

   Katernyi I.V. Causal Explanations of the Uncanny Valley Effect in Robotics: Theories and Applied Research. *Kachestvo i zhizn*. 2017. No. 4. P. 88–96 (In Russ.)

10. *Корбут А.М.* Одомашнивание искусственного интеллекта: умные колонки и трансформация повседневной жизни // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 1. С. 193–216. DOI: 10.14515/monitoring.2021.1.1808 EDN: TKKKDP

    Korbut A.M. Domestication of Artificial Intelligence: Smart Speakers and Transformation of Everyday Life. *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsialnye peremeny*. 2021. No. 1. P. 193–216. DOI: 10.14515/monitoring.2021.1.1808 (In Russ.)

11. *Максимова А.С.* Настройка собеседника: интеракционные барьеры в телефонном разговоре с роботом // Приключения технологий: барьеры цифровизации в России / Л.В. Земнухова и др. М., СПб.: ФНИСЦ РАН, 2020. С. 95–132. DOI: 10.31119/978-5-89697-339-3

    Maksimova A.S. Configuring the Interlocutor: Interaction Barriers in a Telephone Conversation with a Robot. *Technology Adventures: Barriers to Digitalization in Russia*. Ed. by L.V. Zemnukhova, et al. Moscow, St. Petersburg: Federal Center of Theoretical and Applied Sociology of the Russian Academy of Sciences publ., 2020. P. 95–132. DOI: 10.31119/978-5-89697-339-3 (In Russ.)

12. *Нарьян С.К., Быков А.В.* Проблема моральной агентности акторов: перспективы социологического подхода в контексте теории «моральной диады» // Социологический журнал. 2022. Т. 28. № 1. С. 8–23. DOI: 10.19181/socjour.2022.28.1.8835 EDN: NIJSPT

Naryan S.K., Bykov A.V. The Problem of Moral Agency: Prospects of the Sociological Approach in the Context of the "Moral Dyad" Theory. *Sotsiologicheskiy Zhurnal = Sociological Journal*. 2022. Vol. 28. No. 1. P. 8–23. DOI: 10.19181/socjour.2022.28.1.8835 (In Russ.)

13. *Хонинева Е.А.* Гендер и дисплей: коммуникативные жанры и способы категоризации во взаимодействии с голосовыми ассистентами // Журнал социологии и социальной антропологии. 2017. Т. 20. № 5. С. 95–112. DOI: 10.31119/jssa.2017.20.5.6 EDN: ZVMKGF

Khonineva E.A. Gender and Display: Communicative Genres and Ways of Categorization in Interaction with Voice Assistants. *Zhurnal sotsiologii i sotsialnoi antropologii*. 2017. Vol. 20. No. 5. P. 95–112. DOI: 10.31119/jssa.2017.20.5.6 (In Russ.)

14. *Широкова М.* К роботам-курьерам отнеслись по-человечески, 30.11.2023 // Коммерсантъ [электронный ресурс]. Дата обращения 02.01.2025. URL: https://www.kommersant.ru/doc/6366826

Shirokova M. Delivery Robots Were Treated Like Human Beings, 30.11.2023. *Kommersant*. Accessed 02.01.2025. URL: https://www.kommersant.ru/doc/6366826 (In Russ.)

15. *Шталь Б.К., Шредер Д., Родригес Р.* Этика искусственного интеллекта: кейсы и варианты решения этических проблем / Пер. с англ.; Науч. ред. А. Павлов. М.: Издательский дом ВШЭ, 2024. — 200 с. DOI: 10.17323/978-5-7598-2981-2 EDN: WYWKGL

Stahl B.C., Schroeder D., Rodrigues R. Ethics of Artificial Intelligence. Case Studies and Options for Addressing Ethical Challenges. Trans. from Eng. Scientific ed. by A. Pavlov. Moscow: Publ. House of HSE, 2024. 200 p. DOI: 10.17323/978-5-7598-2981-2 (In Russ.)

16. Abend G. Thick Concepts and the Moral Brain. *European Journal of Sociology*. 2011. Vol. 52. No. 1. P. 143–172. DOI: 10.1017/S0003975611000051

17. Awad E., Dsouza S., Kim R., et. al. The Moral Machine Experiment. *Nature*. 2018. Vol. 563. P. 59–64. DOI: 10.1038/s41586-018-0637-6

18. Banks J. A Perceived Moral Agency Scale: Development and Validation of a Metric for Humans and Social Machines. *Computers in Human Behavior*. 2019. Vol. 90. P. 363–371. DOI: 10.1016/j.chb.2018.08.028

19. Banks J. From Warranty Voids to Uprising Advocacy: Human Action and the Perceived Moral Patiency of Social Robots. *Frontiers in Robotics and AI*. 2021. Vol. 8. Art. 670503. DOI: 10.3389/frobt.2021.670503

20. Banks J., Bowman N.D. Perceived Moral Patiency of Social Robots: Explication and Scale Development. *International Journal of Social Robotics*. 2023. Vol. 15. No. 1. P. 101–113. DOI: 10.1007/s12369-022-00950-6

21. Bartneck C., Keijsers M. The Morality of Abusing a Robot. *Paladyn, Journal of Behavioral Robotics*. 2020. Vol. 11. No. 1. P. 271–283. DOI: 10.1515/pjbr-2020-0017

22. Basl J., Sandler R. Three Puzzles Regarding the Moral Status of Synthetic Organisms. *Synthetic Biology and Morality: Artificial Life and the Bounds of Nature*. Ed. by G.E. Kaebnick, T.H. Murray. Cambridge, MA: MIT Press, 2013. P. 89–106. DOI: 10.7551/mitpress/9780262019392.003.0009

23. Bertoncini A.L.C., Serafim M.C. Ethical Content in Artificial Intelligence Systems: A Demand Explained in Three Critical Points. *Frontiers in Psychology*. 2023. Vol. 14. Art. 1074787. DOI: 10.3389/fpsyg.2023.1074787

24. Bigman Y.E., Gray K. People Are Averse to Machines Making Moral Decisions. *Cognition*. 2018. No. 181. P. 21–34. DOI: 10.1016/j.cognition.2018.08.003
25. Brożek B., Janik B. Can Artificial Intelligences Be Moral Agents? *New Ideas in Psychology*. 2019. Vol. 54. P. 101–106. DOI: 10.1016/j.newideapsych.2018.12.002
26. Cervantes J.-A., Lopez S., Rodriguez L.-F., et al. Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*. 2020. Vol. 26. No. 2. P. 501–532. DOI: 10.1007/s11948-019-00151-x
27. Choi T.R., Drumwright M.E. "OK, Google, Why Do I Use You?" Motivations, Post-Consumption Evaluations, and Perceptions of Voice AI Assistants. *Telematics and Informatics*. 2021. Vol. 62. Art. 101628. DOI: 10.1016/j.tele.2021.101628
28. Coeckelbergh M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*. 2020. Vol. 26. No. 4. P. 2051–2068. DOI: 10.1007/s11948-019-00146-8
29. Colombatto C., Fleming S. Folk Psychological Attributions of Consciousness to Large Language Models. *Neuroscience of Consciousness*. 2024. No. 1. Art. niae013. DOI: 10.1093/nc/niae013
30. DeGrazia D. Moral Status as a Matter of Degree? *Southern Journal of Philosophy*. 2008. Vol. 46. No. 2. P. 181–198 DOI: 10.1111/j.2041-6962.2008.tb00075.x
31. DeGrazia D. Robots with Moral Status? *Perspectives in Biology and Medicine*. 2022. Vol. 65. No. 1. P. 73–88. DOI: 10.1353/pbm.2022.0004
32. Dong M., Bocian K. Responsibility Gaps and Self-Interest Bias: People Attribute Moral Responsibility to AI for Their Own but not Others' Transgressions. *Journal of Experimental Social Psychology*. 2024. Vol. 111. Art. 104584. DOI: 10.1016/j.jesp.2023.104584
33. Formosa P., Ryan M. Making Moral Machines: Why We Need Artificial Moral Agents. *AI & SOCIETY*. 2021. Vol. 36. No. 3. P. 839–851. DOI: 10.1007/s00146-020-01089-6
34. Friedman C. Human-Robot Moral Relations: Human Interactants as Moral Patients of Their Own Agential Moral Actions Towards Robots. *Artificial Intelligence Research. SACAIR 2021. Communications in Computer and Information Science*. Ed. by A. Gerber. 2020. Vol 1342. P. 3–20. DOI: 10.1007/978-3-030-66151-9_1
35. Fritz A., Brandt W., Gimpel H., et al. Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI). *De Ethica*. 2020. Vol. 6. No. 1. P. 3–22. DOI: 10.3384/de-ethica.2001-8819.20613
36. Gamez P., Shank D.B., Arnold C., North M. Artificial Virtue: The Machine Question and Perceptions of Moral Character in Artificial Moral Agents. *AI & SOCIETY*. 2020. Vol. 35. No. 4. P. 795–809. DOI: 10.1007/s00146-020-00977-1
37. Hitlin S., Vaisey S. The New Sociology of Morality. *Annual Review of Sociology*. 2013. Vol. 39. No. 1. P. 51–68. DOI: 10.1146/annurev-soc-071312-145628
38. Ikari S., Sato K., Burdett E., et. al. Religion-Related Values Differently Influence Moral Attitude for Robots in the United States and Japan. *Journal of Cross-Cultural Psychology*. 2023. Vol. 54. No. 6–7. P. 742–759 DOI: 10.1177/00220221231193369
39. Korsgaard C. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996. 273 p. DOI: 10.1017/CBO9780511554476
40. Laakasuo M., Herzon V., Perander S., et al. Socio-Cognitive Biases in Folk AI Ethics and Risk Discourse. *AI and Ethics*. 2021. Vol. 1. No. 4. P. 593–610. DOI: 10.1007/s43681-021-00060-5

41. Mittelstadt B.D., Allo P., Taddeo M., et al. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*. 2016. Vol. 3. No. 2. Art. 205395171667967. DOI: 10.1177/2053951716679679
42. Moosavi P. Will Intelligent Machines Become Moral Patients? *Philosophy and Phenomenological Research*. 2023. Vol. 109. No. 1. P. 95–116. DOI: 10.1111/phpr.13019
43. O'Leary D.E. GOOGLE'S Duplex: Pretending to Be Human. *Intelligent Systems in Accounting Finance & Management*. 2019. Vol. 26. No. 1. P. 46–53. DOI: 10.1002/isaf.1443
44. Pause Giant AI Experiments: An Open Letter, 23.03.2023. *Future of Life Institute*. Accessed 09.01.2025. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/
45. Redaelli R. Different Approaches to the Moral Status of AI: A Comparative Analysis of Paradigmatic Trends in Science and Technology Studies. *Discover Artificial Intelligence*. 2023. Vol. 3. No. 1. Art. 25. DOI: 10.1007/s44163-023-00076-2
46. Scheessele M. A Framework for Grounding the Moral Status of Intelligent Machines. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Ed. by J. Furman, G. Marchant, H. Price, F. Rossi. N.-Y.: Association for Computing Machinery, 2018. P. 215–256 DOI: 10.1145/3278721.3278743
47. Shank D.B., DeSanti A., Maninger T. When Are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions after Individual and Joint Decisions. *Information, Communication & Society*. 2019. Vol. 22. No. 5. P. 648–663. DOI: 10.1080/1369118X.2019.1568515
48. Shevlin H. How Could We Know When a Robot Was a Moral Patient? *Cambridge Quarterly of Healthcare Ethics*. 2021. Vol. 30. No. 3. P. 459–471. DOI: 10.1017/S0963180120001012
49. Smiley L. The Legal Saga of Uber's Fatal Self-Driving Car Crash is Over. *Wired*. July 28, 2023. Accessed 25.12.2024. URL: https://www.wired.com/story/ubers-fatal-self-driving-car-crash-saga-over-operator-avoids-prison/
50. Sparrow R. Why Machines Cannot Be Moral. *AI & SOCIETY*. 2021. Vol. 36. No. 3. P. 685–693. DOI: 10.1007/s00146-020-01132-6
51. Sullivan Y.W., Fosso Wamba S. Moral Judgments in the Age of Artificial Intelligence. *Journal of Business Ethics*. 2022. Vol. 178. No. 4. P. 917–943. DOI: 10.1007/s10551-022-05053-w
52. The Nobel Prize in Chemistry 2024. Press-release, 09.10.2024. *The Royal Swedish Academy of Sciences*. Accessed 01.02.2025. URL.: https://www.nobelprize.org/prizes/chemistry/2024/press-release/
53. Tsamados A., Aggarwal N., Cowls J., et al. The Ethics of Algorithms: Key Problems and Solutions. *Ethics, Governance, and Policies in Artificial Intelligence*. Ed. by L. Floridi. Springer, 2021. P. 97–123. DOI: 10.1007/978-3-030-81907-1_8

Статья поступила в редакцию: 12.09.2024; поступила после рецензирования и доработки: 06.03.2025; принята к публикации: 13.03.2025.

Received: 12.09.2024; revised after review: 06.03.2025; accepted for publication: 13.03.2025.