Этическая мысль
2025. Т. 25. № 1. С. 36–49
УДК 17:004.8

Ethical Thought
2025, Vol. 25, No. 1, pp. 36–49
DOI: 10.21146/2074-4870-2025-25-1-36-49

*А.С. Зильбер, К.С. Зильбер*

# Бесстрастный холоп: кантианский подход к агентности искусственного интеллекта*

***Зильбер Андрей Сергеевич*** – кандидат философских наук. Балтийский федеральный университет им. И. Канта. Российская Федерация, 236041, г. Калининград, ул. А. Невского, д. 14.

ORCID 0000-0002-8317-4086
e-mail: azilber@kantiana.ru

***Зильбер Ксения Сергеевна*** – лаборант. Балтийский федеральный университет им. И. Канта. Российская Федерация, 236041, г. Калининград, ул. А. Невского, д. 14.

e-mail: kszilber@kantiana.ru

В статье исследуется вопрос о кантовской точке зрения на моральный статус искусственного интеллекта и перспективы применения элементов философии Канта для оценки ИИ как морального агента. Показано, что к этике Канта применима «стандартная» теория агентности Г. Франкфурта. Подход Канта к моральному статусу субъекта определяется природой агента – структурой способностей души, среди которых ведущую роль играют воля и разум, а также чувственность. Показана связь способностей души в системе Канта со свободой и моральной автономией. Раскрывается сложная двоякая функция чувственности в ограничении практической свободы и в воспитании свободы морального агента. С точки зрения философии Канта искусственный моральный агент невозможен с учетом сложности природы людей – и может выступать только квазиморальным агентом: внешне сходным с подлинными моральными агентами, но не сходным с ними по сути. Проблематичным

* Статья подготовлена при финансовой поддержке Министерства науки и высшего образования Российской Федерации, проект № 075-15-2019-1929 «Кантианская рациональность и ее потенциал в современной науке, технологиях и социальных институтах», реализуемый на базе Балтийского федерального университета им. И. Канта (Калининград).
This publication has been prepared with the financial support of the Ministry of Science and Higher Education of the Russian Federation, Project No. 075-15-2019-1929 “Kantian rationality and its potential in modern science, technology and social institutions”, implemented at the Immanuel Kant Baltic Federal University (Kaliningrad).

в ряде аспектов оказывается также статус легального агента и использование права вместо морали в качестве основы машинной этики. В то же время природа ИИ сулит, при условии обучения, возможности преодоления определенных ограничений человеческой природы. Рассматриваются элементы альтернативного функционального подхода, сфокусированного на вопросах ответственности, моральных ожиданий в сфере ИИ и общественных последствий применения ИИ. Итоговая основа для правового статуса ИИ выявляется в кантовской метафизике права: рабы или холопы, существа, имеющие обязанности, но не имеющие прав, что согласуется с ролью ИИ как строго ограниченного в своей свободе помощника.

***Ключевые слова:*** моральная агентность, Кант, искусственный интеллект, машинная этика, моральная автономия, практический разум, свобода воли, моральное чувство, ответственность

## Введение

Развитие систем искусственного интеллекта сделало актуальным вопрос о моральном статусе искусственных агентов. Термин «агентность», согласно теории, которая была представлена Гарри Франкфуртом [Frankfurt, 1971] и сегодня обрела статус «стандартной» [Schlosser, 2019], обозначает способность выполнять намеренное действие, которой обладают существа, имеющие желания, убеждения и намерения. В этой теории люди как агенты характеризуются специфической структурой воли: рефлексией по поводу своих мотивов, разделением и различением желаний первого и второго порядка. Это основные способности морального субъекта и ведущая тема в этике Канта, следовательно, вести речь о кантианском подходе к моральной агентности в целом возможно – ею могут обладать, по Канту, только разумные существа. В практическом отношении полноценная моральная агентность подразумевает способность объяснить свое поведение и нести ответственность за последствия своих действий [Shoemaker, 2011]. Нас интересуют обе точки зрения, которые можно условно назвать «теоретической» и «практической»: как формируется агентность и как она проявляется.

Существующие попытки определения кантианского подхода к статусу агентов с ИИ сходятся в том, что эти агенты не могут обладать полноценным моральным статусом [Федотова, 2023; Schönecker, 2022], но, возможно, смогут выполнять ограниченные функции в анализе морально значимой информации [Powers, 2006]. Мы намерены исследовать возможность и перспективы рассмотрения кантовских критериев моральности в контексте ИИ в аналитической трактовке – в частичном отрыве от системы его философии. Для этого требуется определить, как соотносятся: 1) устройство ИИ и кантовская концепция способностей души; 2) природа машинных вычислений и требования формализации, универсализации и автономии в этике Канта; 3) свобода агентов с ИИ и свобода человека в понимании Канта. Особый пункт аналитической трактовки образуется тем, что в кантовской этике важную роль играют критерии формальности и автономии, которые присутствуют в характеристиках роботов и компьютерных программ.

Понимание агентности в «стандартной» теории приблизительно соответствует определению ИИ как способности «выбора и принятия целесообразного решения при большом многообразии целей» [Толковый словарь, 1992, 247] – т.е. «общему ИИ», который сочетал бы уровень развития отдельных интеллектуальных способностей и широту их применения, а также адаптивность и способность к обучению [Сильный искусственный интеллект, 2021, 27]. Такой общий ИИ предполагается достижимым в обозримом будущем, в то время как создание «сильного» ИИ, притязающего на разумность с наличием сознания и самосознания, отодвинуто на дальний горизонт как сверхзадача. Перенос внимания от «сильного» к «общему» ИИ – одно из проявлений функционального подхода, в котором среди прочих выделяются вопросы практики применения искусственных агентов и тех моральных ожиданий, которые на них возлагаются обществом. В этой связи мы рассмотрим предложение наделить ИИ особым ограниченным моральным статусом «квазиморальный агент» [Глебова, Перова, 2023, 37], а также предложение строить машинную этику на основе права, а не морали, с учетом как природы агентов, так и эмпирической ситуации с признанием норм [Wright, 2022].

## Автономия агента: разум и воля

Человека делают моральным существом, по Канту, прежде всего разум и воля. На второе место после них с некоторыми оговорками можно поставить чувственность, ее противоречивая роль рассматривается в следующем пункте этой статьи. Остальные способности души играют вспомогательную роль, которую мы не будет рассматривать подробно. Кантианский подход к моральной агентности соединяет «теоретическую» и «практическую» точку зрения, поскольку в нем неотъемлемым критерием морального поступка объявляется мотив.

Волю Кант определяет как «способность действовать согласно *представлению* о законах, т.е. по принципам» [Кант, 1997б, 118]. Кант полагает, что волей обладают только разумные существа. Разум, в свою очередь, требуется «для выведения действий из законов» [Там же], таким образом, Кант готов отождествить его с волей, но у человека это тождество не полное: разум определяет волю недостаточно и она подвержена субъективным условиям – «известным мотивам» обеспечить себе счастье и благополучие [Там же, 119]. Субъективные побуждения направлены на конкретные (материальные) цели – обладание вещами, переживание эмоций, достижение состояний [Кант, 1997а, 21]. Главный моральный принцип, который исходит от разума, – поступать независимо от субъективных и преходящих побуждений чувственности [Там же, 27]. Это моральная автономия, свобода от определения воли чувственными побуждениями, конкретными предметами, случайными обстоятельствами, личными интересами [Там же, 29, 33]. Ввиду их частого расхождения с принципами морали, последние являются для человека долгом [Там же, 32].

Наличие желаний второго порядка, которые, согласно стандартной теории агентности, направлены на желания первого порядка, само по себе еще

не является моральностью. Рефлексия над желаниями первого уровня может иметь утилитарную цель. Если эта цель оказалась в согласии с долгом, то поступок не моральный, а легальный (внешне сообразный с долгом) [Кант, 1997а, 109]. Соблюдение моральных обязанностей из страха или склонности, без уважения к долгу – это гетерономия; соблюдение законов из страха или склонности, в одном только внешнем отношении, Кант полагал сферой права. В мотиве поступка невозможно удостовериться извне, он очевиден только самому агенту – субъекту, сознающему свои мотивы. Да и очевидность эта обманчивая: Кант подчеркивает склонность людей к самообману, к иллюзиям в отыскании моральных оправданий для своих поступков [Кант, 1997б, 95]. На полную моральную чистоту рассчитывать не приходится и вовсе не нужно [Кант, 1997а, 531], но долг человека подразумевает стремление к очищению своей воли, стремление к совершенствованию как процесс прояснения своих мотивов [Там же, 579, 619].

Согласно философии Канта, моральной агентностью обладает всякое разумное живое существо, поскольку оно имеет способность осознавать моральный закон и поступать морально [Там же, 395], а вместе с ней и безусловное нравственное достоинство [Кант, 1997б, 199]. Позволяет ли природа ИИ сделать его субъектом морального рассуждения, который свободен от ограничений, обусловленных природой человека? Для ИИ веления долга не требуются, поскольку ИИ не имеет телесной природы, а значит, не имеет искушений, склонностей, чувственных побуждений. Можно предположить, что вычислительная мощность в сочетании с отсутствием склонностей даст машинам преимущество в последовательном применении универсальных и формальных принципов к конкретным случаям с четкими размеченными критериями. Однако к этому есть значительные препятствия.

Во-первых, вопреки надеждам Канта на простоту и легкость дедуктивного вывода, современные моральные философы показывают, что попытки развернуть категорический императив морали как алгоритм для принятия решений – выявляют серьезные трудности этого предприятия, обзор которых образует отдельную большую тему за рамками нашего исследования (см.: [Чалый, 2024]). Фрагмент смежной темы – о возможности заложить в ИИ принципы права – рассматривается далее.

Во-вторых, дело в самой сущности разума. Его функция, по Канту, состоит в том, чтобы давать принципы единства правил рассудка, направляя его на поиск систематического единства законов природы [Кант, 2006, 357]. Разум нацелен на поиск условий для познания рассудка [Там же, 364]. ИИ не вырабатывает собственные принципы и условия для мышления как познания, поэтому для ИИ больше подходит функция рассудка как способности мыслить. Кроме того, ИИ не обладает «представлениями», т.е. ментальными репрезентациями, а значит, и намерениями. Свобода ИИ в вынесении моральных суждений ограничена рассудком и не выходит на уровень разума.

Вдобавок смысл той свободы, которую Кант видел в человеке, в позитивном измерении состоит в способности формулировать законы, значимые для всех разумных существ – и соответствующие достоинству разумного существа. В робототехнике, как отмечает В.Ю. Перов, речь идет о другой свободе

и другой автономии: степенью свободы называют переменные для определения движения тела в пространстве [Перов, 2024, 411]; автономностью робототехнических устройств называют «способность выполнять задачи по назначению на основе текущего состояния и восприятия внешней среды без вмешательства человека» [Там же, 412].

### Эмоции и ответственность

Беспристрастность позволяла бы избегать решений, диктуемых случайными желаниями, личными интересами и склонностями, – однако Кант вовсе не проповедовал аскетизм. Люди по своей природе имеют, по Канту, «виды на блаженство» – стремление к счастью [Кант, 1997б, 127; Кант, 1997а, 335], и способствовать своему счастью тоже долг, но только опосредованным образом – в той мере, в какой быть счастливым означает быть способным совершать любые поступки, в том числе моральные, а также обладать стойкостью для противостояния искушениям [Кант, 1997а, 531]. Содержание морального сознания составляют как борьба, так и союзы побуждений разума и чувственности. Единственное чувство, необходимое в моральной автономии, – особое чувство уважения к долгу, имеющее неэмпирическое происхождение [Там же, 73].

В сознании людей чувства выступают не только препятствием для соблюдения правил поведения, но и побудителем к моральным или легальным поступкам. Оценивать соотношение положительной и отрицательной стороны этого влияния можно по-разному. Д. Шенекер доказывает, что в этике Канта чувства и эмоции выступают обязательным компонентом формирования морального сознания [Schönecker, 2022]. За пределами этики Канта констатации этой связи звучат не менее решительно. Отмечается, что именно эмоциональный опыт придает окраску нашим представлениям о вещах, что именно эмоции позволяют нам понимать разницу между допустимым и нежелательным поведением. Например, сильный эмоциональный дефицит, даже при отсутствии когнитивных нарушений, ведет к трудностям в понимании разницы между безусловными моральными правилами и ограниченными конвенциональными правилами [Damasio, 1994].

Таким образом, холодные, чисто механические и рациональные рассуждения сами по себе недостаточны для формирования моральной мотивации – чувственность в тесной связке с мышлением вместе образуют своего рода когнитивно-эмоциональное оснащение живого существа как морального агента. В кантианском подходе можно признать, что значимость моральных ценностей если не определяется, то как минимум закрепляется через чувства одобрения или неодобрения, удовлетворения или отвращения. Современные исследователи психики это называют «эффективной связью» ценностей с эмоциональной жизнью [Deonna, Teroni, 2012, 106]. В данном контексте речь идет не столько о чувствах как источнике информации о внешнем мире и состоянии своего тела, сколько о сфере моральных чувств: от любви и симпатии до злобы и зависти – хотя они тоже тесно связаны с телесной природой, которой не имеет ИИ, как слабый, так и общий.

Искусственные агенты на основе ИИ, которые можно назвать модульными практическими искусственными агентами (МПИА), не имеют потребностей и желаний, и не испытывают искушения отклониться от того, что диктуют его моральные расчеты, и лишены себялюбия. Ошибочные решения МПИА происходят от неправильной оценки ситуации. В докладе экспертов Совбеза ООН по Ливии в 2021 г. сообщалось о применении в боевых действиях автономных роботов, принимающих решения о поражении живых целей без участия оператора [Заключительный доклад, 2021, 20]. Если, например, боевой дрон уничтожил мирного жителя, который был распознан как враг, то это нельзя считать аморальным. На боевой операции неуместно подозревать МПИА в «дружественном огне» по своим под видом ошибки ради сведения счетов с личным врагом, т.к. образ врага формируется при условии восприятия себя как личности. Не имея самосознания, МПИА не имеет совести – не может чувствовать вину и раскаиваться в своих ошибках. Не имея рефлексивности, он неспособен оправдать свои действия или извинить их.

Итак, кантианскому моральному агенту требуется понимание различия между сущим и должным, а также между тем, что правильно, и тем, что неправильно. Робот без самосознания, по словам С. Шовье, похож на «апатичного святого» [Chauvier, 2016]. Кс. Барандайран вместе с соавторами доказывал в 2009 г., что минимальная агентность требует не столько обладания ментальными состояниями, сколько адаптивной регуляции слияния агента со средой и метаболического самоподдержания [Barandiaran et al., 2009, 382] – то есть наличия биологической природы живого существа. Для искусственных систем такой критерий выглядит очень проблематичным. Развитие ИИ на основе новой формы жизни – это не только усложнение задачи, но и довольно пугающая перспектива, о которой писали фантасты.

Статус морального агента включает моральную ответственность, и ее тоже связывают с моральными чувствами. Об этом рассуждают не только в контексте кантовской этики [Федотова, 2023], но и в современном общефилософском контексте (см.: [Глебова, Перова, 2023]). Среди значимых свойств учитывается та их совокупность, которая включает ограниченную автономность ИИ и неспособность осознать ценностное основание правил, а также отсутствие личностного начала (памяти о своем собственном опыте).

С.В. Глебова и Н.В. Перова [Там же, 37] показывают, что отказать ИИ в полноценной моральной агентности – половинчатое решение, которого недостаточно для определения его морального статуса по причине моральной значимости ситуаций применения ИИ. Ярчайшим примером в их обзоре оказывается применение нейросетей в сфере информационных услуг, а именно «прекращение работы нейросетей в случае отрицания холокоста или оскорбления определенных групп населения» [Там же, 36]. Вдобавок от ИИ «ожидается еще и соблюдение ряда нравственных установок» – хотя Глебова и Перова отмечают, что языковые модели ИИ руководствуются не прямыми формулировками этических принципов, а «стоп-словами» [Там же]. Можно считать, что дополнительная настройка ограничений в ИИ по сути аналогична практике, когда производитель отзывает у пользователей определенную партию автомобилей, в которой обнаружились критические угрозы безопасности

человека. С этой точки зрения, в случае нейросетей проблема состоит только в том, что ожидания человека от ИИ завышены, что доверие к ИИ необоснованно велико, в то время как высокие требования качества и безопасности предъявляются к широкому кругу сложных технических устройств. Однако Глебова и Перова предлагают рассматривать ИИ как «актора, которому приписывается ряд нормативных установок», и выделить для ИИ особый статус: «квазиморальный» агент, поскольку он относительно самостоятельно оперирует ограниченным набором инструментов, а также поскольку процесс взаимодействия ИИ с моральными агентами отличается от простого применения техники и может быть назван моральной коммуникацией [Глебова, Перова, 2023, 37]. Приставка «квази-» в этом статусе означает наличие внешнего сходства (и – можно добавить – преобладание положительных результатов) одновременно с несоответствием внутренней сути; квазиморальный агент отличается от полноценного агента больше, чем ограниченный моральный агент (дети и душевнобольные). Такое понимание, при всех его нюансах, на наш взгляд, согласуется с кантианским подходом, для которого главное именно во внутренней сути или природе агента.

В кантианском подходе ИИ ограничен ролью помощника человека. В этой связи заслуживает внимания предложение Н.А. Ястреб сосредоточить внимание не на возможностях ИИ или изначальных целях его создания, но на анализе влияния ИИ на человека и задать следующие вопросы: «Почему мне нравится использовать эту систему; какие мои слабости она использует, какие недостатки преодолевает; какие ценности поддерживает, как какие нивелирует и, наконец, какие установки и стереотипы эксплуатирует? Искусственный интеллект не столько создает новые этические проблемы и парадоксы, сколько актуализирует, расширяет и делает заметным то, что уже свойственно человеку и присутствует в обществе» [Ястреб, 2024, 449]. Это предложение развивать кодексы ИИ с учетом крупномасштабных эффектов воздействия этой технологии на людей.

### Агентность моральная или правовая?

После установления невозможности *моральных* рассуждений для общего ИИ, ввиду отсутствия у него разума, воли, сознания, моральных чувств, – встает вопрос о том, как соотносятся возможности ИИ с кантовским определением *легальных* рассуждений. Легальность поступка по определению не принимает во внимание мотивацию [Кант, 2014, 57]. Правда, у Канта речь идет не об отсутствии воли и сознания: в легальных поступках они присутствуют, но определяющее основание воли не имеет значения. Принцип права у Канта – обеспечение внешней свободы тех, кому эта свобода требуется для поисков своего личного счастья, которое является целью каждого человека согласно его природе. Иными словами, свобода не для тех зомби, которые фигурируют в современной философии сознания, а для существ, осознающих свои мотивы и интересы, как разумные, так и неразумные. Можно подозревать большинство людей в том, что их мотивы совершения добрых поступков чаще всего только

легальны, однако, с кантовской точки зрения, это не лишает их человеческого достоинства. Машины, способные соблюдать нормы морали чисто внешним образом, могли бы стать добропорядочными агентами, схожими с людьми, и при этом было бы неважно, стремятся ли они к некоему машинному аналогу человеческого счастья. Итак, следует ли считать рассуждения ИИ о морали по сути легальными рассуждениями на том основании, что ИИ способен применять моральные критерии к оценке действий и событий, но не обладает моральным сознанием?

А.Т. Райт предложил формировать машинную этику именно на основе норм права, а не морали – сконцентрироваться на создании «правовых» (rightful)[1] машин [Wright, 2022]. Сферу права он считает базовой и четко определенной. По Канту, система публичного права имеет приоритет перед личным этическим суждением. Основную проблему разработки «моральных» машин Райт видит в том, что этические обязанности в большинстве своем позитивные (предписания) и по ним возможно многообразие личных предпочтений, которые зачастую несовместимы, в то время как большинство юридических обязанностей – «негативны», это запреты, или, иными словами, четкие ограничения [Ibid., 229].

Для иллюстрации своего подхода Райт рассматривает две версии известной проблемы вагонетки [Ibid., 230–231]. Первая версия – проблема с точки зрения водителя, который попадает в ситуацию, когда жертвы неизбежны и выбор состоит только в том, сколько их будет, больше или меньше. Это ситуация конфликта между юридическими обязанностями, однако Райт видит из нее юридический выход: соблюдение основной обязанности водителя по обеспечению безопасного вождения. Предпринять бездействие, продолжив ехать в своей полосе независимо от последствий, – будет невыполнением этой обязанности. Если по возможности уменьшить число пострадавших и жертв – обязанность будет соблюдена, таким образом, Райт выступает за решение спасти больше жизней, насколько это способен сделать водитель.

Вторая версия – столкнуть ли с моста одного толстяка, чтобы остановить вагонетку и таким образом спасти несколько других людей? В этом случае, по Райту, следует исходить из того, что юридическое право человека на жизнь включает в себя право не жертвовать своей жизнью или, иными словами, «не быть принужденным к смерти ради спасения других» [Ibid., 230]. Запрет на убийство Райт рассматривает как обязанность, закрепленную в публичном праве, а спасение жизни и здоровья людей он полагает в данной ситуации этическим долгом [Ibid., 231]. Принципы справедливости, – продолжает Райт, – как правило, запрещают нарушение прав одного человека для достижения большего блага, такого как спасение многих людей [Ibid.].

Этот подход по избеганию конфликтов между обязательствами путем опоры на основные права имеет очевидные плюсы. В праве более четко, чем в морали, определяются критерии соблюдения предписаний, а также ответственность за их нарушения и за последствия действий, которая позволяет оценить риски. Согласование внешней свободы людей можно считать сравнительно менее трудной задачей, чем моральное воспитание – хотя, как показывает

[1] Ю.С. Федотова [Федотова, 2023] переводит rightful как «законный».

история и философия права, в этой задаче есть свои сложности, и даже очень большие. Кант обращал внимание на то, что правовой долг – безусловный, и поэтому имеет приоритет над долгом моральным, например долгом любви, и чувство благоволения нельзя делать основанием для ущемления чьих-либо прав [Кант, 2024а, 210]. В текстах Канта встречается рассуждение о том, что проявлениями морального прогресса должен будет стать прогресс легальности [Кант, 2024б, 337]: и если законопослушные машины будут способствовать этой трансформации, то для людей это станет дополнительным фактором в воспитании морального образа мыслей.

Слабые места в подходе Райта также присутствуют и касаются как проблем реализации на уровне алгоритмов, так и природы агентов. Во-первых, его решение, по сути, требует абстрактного рассуждения и затрагивает сферу ценностей. Показать искусственному интеллекту, что такое «безопасное вождение», – означает обучать нейросеть на примерах, показывающих, какое вождение на практике более безопасное. Во-вторых, хотя Райт утверждает, что основные права, такие как право на жизнь, можно считать предметом всеобщего одобрения, все же позитивное право представлено различными версиями в разных государствах. Некоторые государства не относят право на жизнь к основным правам либо допускают прекращение его действия в случае эвтаназии или при вынесении приговора суда о смертной казни. В этом контексте можно усмотреть следующую вилку альтернатив: либо исходить из Всеобщей декларации прав человека, а в дополнение к ней – искать компромисс между системами позитивного права, либо закладывать в программу кантовские принципы чистого права и готовиться к возможным коллизиям между этими принципами и реальной практикой, принятой в государствах.

В подходе Райта не ставится вопрос о правовом статусе самих искусственных агентов – роботов и нейросетей. Можно было бы именовать их «легальными» агентами, поскольку это понятие более узкое, чем понятие «правовое». Человек в кантовской системе права обладает не только обязанностями, но и правами, и принцип права не только в обеспечении внешней свободы. Право означает правомочие принуждать, носитель прав является потенциальным субъектом и объектом принуждения [Кант, 2014, 91]. Наделить роботов правами означало бы считаться с их взглядами и нуждами, быть готовыми ради них поступиться своей человеческой свободой. Однако поскольку, как мы показали, наличие собственно взглядов и нужд у роботов не предполагается ввиду их природы, то не возникает и вопроса о принуждении и наделении их правами. В системе Канта предусмотрено место для существ, которые имеют обязанности, не имея никаких прав, – это холопы или рабы, это люди, лишенные личности [Там же, 117]. Именно таким мог бы быть правовой статус «легальных агентов», согласно их природе. Люди получают такой статус только «по суду и по праву», становясь «орудием чужой воли», которое можно «отчуждать как вещь, и использовать его по своему усмотрению (только не с низменными целями), и распоряжаться (располагать) его силой» [Там же, 357], что на практике означало также и распоряжение жизнью, например, когда рабов могли доводить до смерти через истощение их сил [Там же, 359].

## Заключение

Кантианский подход к моральной агентности основан на агентности человека как существа, обладающего разумом, волей и чувственностью. Именно это сочетание дает свободу как способность формулировать общезначимые принципы и следовать им, также именно эта связка способностей позволяет нести ответственность и испытывать моральные чувства, включая среди прочих и ключевое для морали чувство уважения к долгу. Сформулированное в начале нашего исследования предположение о том, что отсутствие у роботов чувственности, биологической телесности (источников гетерономии воли) может быть перспективным для моральной агентности – в ходе детального рассмотрения кантовского подхода не подтвердилось. Все три способности работают во взаимодействии, и физическая чувственность как способность, нарушающая чистоту мотивов, вместе с тем играет важную роль и в формировании воли и высших моральных чувств.

Агенты со «слабым» искусственным интеллектом отличаются тем, что воплощают отдельные узконаправленные физические и интеллектуальные способности лучше человека, в то время как люди отличаются от машин именно широтой своих способностей. Сочетанием разума, воли и чувственности обладают сознательные биологические организмы, от которых остается далеким даже гипотетический «общий» ИИ из (предположительно) довольно близкого будущего, который, как ожидается, должен обладать способностями существенно более широкими, чем «слабый», но все же не предполагается носителем сознания.

Тем не менее аналитическое рассмотрение машинного воплощения отдельных способностей, из которых складывается моральная агентность, представляет интерес как перспектива дополнения кантовской этики в ее приложении к современной технике. Эту перспективу мы оценивали, опять же, преимущественно с кантовских позиций. Правда, в нашем исследовании не выявлено никаких впечатляющих итогов такого рассмотрения: 1) свобода морального рассуждения ИИ ограничена рассудком и не выходит на уровень разума; 2) отсутствие высших моральных чувств лишает ИИ понимания морального достоинства; 3) искусственным агентам проблематично приписывать даже легальное поведение, поскольку оно тоже предполагает деятельность сознания и наличие интересов.

Предложение Глебовой и Перовой ввести статус «квазиморального» агента с учетом тех возможностей, которыми обладают нейросети и роботы, а также тех ожиданий, которые на них возлагаются, – по сути согласуется с кантовским подходом. Квазиморальный агент имеет сходство с моральным агентом по своему проявлению, но не по своей сути. В том же значении можно было бы употребить приставку «псевдо-», имеющую, правда, несколько более негативный, чем позитивный, оттенок. Предложение А.Т. Райта принять право вместо морали в качестве основы машинной этики проблематично ввиду того, что право содержит ничуть не меньше абстрактных понятий, чем мораль, и имеет различные воплощения в системах позитивного права.

Однако именно в сфере права удается найти ту роль, которая предусмотрена в метафизике Канта и при этом близка к функции современных агентов

на основе ИИ – это роль раба или холопа: существа с обязанностями, но без прав. Рассуждать о том, преобладают ли в действиях такого агента легальные поступки, – затея неблагодарная и малополезная. Важно, что этот статус позволяет в одностороннем порядке реализовать концепцию права как «правомочия принуждать», избежав при этом пугающего для людей ответного принуждения – необходимости считаться с волей роботов на том основании, что они носители прав. Остается только, во избежание недоразумений, объяснить роботам и нейросетям, что применительно к ним «холоп» или «раб» звучит ничуть не оскорбительно.

## Impassive Serf: The Kantian Approach to Artificial Intelligence Agency

***Andrey S. Zilber***

Immanuel Kant Baltic Federal University. 14 A. Nevskogo str., Kaliningrad, 236041, Russian Federation.

ORCID 0000-0002-8317-4086
e-mail: azilber@kantiana.ru

***Kseniya S. Zilber***

Immanuel Kant Baltic Federal University. 14 A. Nevskiy str., Kaliningrad, 236041, Russian Federation.

e-mail: kszilber@kantiana.ru

The article examines the issue of Kant's point of view on the moral status of artificial intelligence and the prospects for applying elements of Kant's philosophy to evaluate AI as a moral agent. It is shown that the "standard" theory of agency by G. Frankfurt is applicable to Kant's ethics. Kant's approach to the moral status of the subject is determined by the nature of the agent – the structure of the abilities of the soul, among which the leading role is played by will and reason, as well as sensuality. The connection of the soul's abilities in Kant's system with freedom and moral autonomy is shown. Authors reveal the complex dual function of sensuality in limiting practical freedom and in fostering the freedom of a moral agent. From the point of view of Kant's philosophy, an artificial moral agent is impossible given the complexity of human nature – and can only act as a quasi-moral agent: superficially similar to genuine moral agents, but not similar to them in essence. The status of a legal agent and the use of law instead of morality as the basis of machine ethics are also problematic in a number of aspects. At the same time, the nature of AI promises, subject to training, the possibility of overcoming certain limitations of human nature. The elements of an alternative functional approach focused on issues of responsibility, moral expectations in the field of AI and the social consequences of the use of AI are considered. The final basis for the legal status of AI is revealed in the Kantian metaphysics of law: slaves or serfs, beings with duties but no rights, which is consistent with the role of AI as a strictly limited helper.

***Keywords:*** moral agency, Kant, artificial intelligence, machine ethics, moral autonomy, practical reason, free will, moral sense, responsibility

## Литература / References

*Кант И.* Критика чистого разума. 2-е изд. // *Кант И.* Соч. на нем. и рус. яз. Т. 2 (1) / Под ред. Б. Тушлинга, Н.В. Мотрошиловой. М.: Наука, 2006.

Kant, I. "Kritika chistogo razuma", 2nd ed. [Critique of Pure Reason], in: I. Kant, *Sochinenia na nem. i rus. jaz.* [Works in Russian and German], Vol. 2 (1), eds. B. Tushling, N.V. Motroshilova. Moscow: Nauka Publ., 2006. (In Russian)

*Кант И.* Критика практического разума // *Кант И.* Соч. на нем. и рус. яз. Т. 3 / Под ред. Э.Ю. Соловьева, А.К. Судакова. М.: Московский философский фонд, 1997а. С. 277–733.

Kant, I. "Kritika prakticheskogo razuma" [Critique of Practical Reason], in: I. Kant, *Sochinenia na nem. i rus. jaz.* [Works in German and Russian], Vol. 3, eds. E.Ju. Solov'ev, A.K. Sudakov. Moscow: Moscow Filosofskij Fond Publ., 1997a, pp. 277–733. (In Russian)

*Кант И.* Основоположение к метафизике нравов // *Кант И.* Соч. на нем. и рус. яз. Т. 3 / Под ред. Э.Ю. Соловьева, А.К. Судакова. М.: Московский философский фонд, 1997б. С. 39–275.

Kant, I. "Osnovopolozhenie k metafizike nravov" [Groundwork of the Metaphysics of Morals], in: Kant, I. *Sochinenia na nem. i rus. jaz.* [Works in German and Russian], Vol. 3, eds. E.Ju. Solov'ev, A.K. Sudakov. Moscow: Moskovskij filosofskij fond Publ., 1997б, pp. 39–275. (In Russian)

*Кант И.* Метафизика нравов. Часть первая. Метафизические первоначала учения о праве // *Кант И.* Соч. на нем. и рус. яз. Т. 5. Ч. 1. М.: Канон+ РООИ «Реабилитация», 2014. С. 19–474.

Kant, I. "Metafizika nravov. Chast' pervaja. Metafizicheskie pervonachala uchenija o prave" [Metaphysics of Morals. Metaphysical First Principals of the Doctrine of Right], in: I. Kant, *Sochinenia na nem. i rus. jaz.* [Works in German and Russian], Vol. 5, Part. 1. Moscow: Kanon+ ROOI "Reabilitacija" Publ., 2014, pp. 19–474. (In Russian)

*Кант И.* К вечному миру / Пер. М. Ровбо // *Кант И.* Сочинения по антропологии, философии политики и философии религии в 3 т. Т. 2. Калининград: Изд-во БФУ им. И. Канта, 2024а. С. 159–211.

Kant, I. "K vechnomu miru" [Toward Perpetual Peace], trans. by M. Rovbo, in: I. Kant, *Sochinenija po antropologii, filosofii politiki i filosofii religii* v 3 t. [Works on Anthropology, Philosophy of Politics and Philosophy of Religion in 3 vols.], Vol. 2. Kaliningrad: IKBFU Press Publ., 2024a, pp. 159–211. (In Russian)

*Кант И.* Спор факультетов / Пер. А.С. Зильбера // *Кант И.* Сочинения по антропологии, философии политики и философии религии в 3 т. Т. 2. Калининград: Изд-во БФУ им. И. Канта, 2024б. С. 250–361.

Kant, I. "Spor fakul'tetov" [The Conflict of the Faculties], trans. by A.S. Zilber, in: I. Kant, *Sochinenija po antropologii, filosofii politiki i filosofii religii* v 3 t. [Works on Anthropology, Philosophy of Politics and Philosophy of Religion in 3 vols.], Vol. 2. Kaliningrad: IKBFU Press Publ., 2024б, pp. 250–361. (In Russian)

*Глебова С.В., Перова Н.В.* Условия моральной агентности: моральный агент, ограниченный моральный агент, квазиморальный агент // Дискурс. 2023. Т. 9. № 4. С. 29–40.

Glebova, S.V., Perova, N.V. "Uslovija moral'noj agentnosti: moral'nyj agent, ogranichennyj moral'nyj agent, kvazimoral'nyj agent" [Moral Agency Conditions: Moral Agent, Limited Moral Agent, Quasi-Moral Agent], *Discourse*, 2023, Vol. 9, No. 4, pp. 29–40. (In Russian)

ГОСТ Р 60.0.0.4 – 2019/ ИСО 8373:2012. Роботы и робототехнические устройства. Термины и определения. М.: Стандартинформ, 2019.

*GOST R 60.0.0.4 — 2019/ ISO 8373:2012 Roboty i robototehnicheskie ustrojstva. Terminy i opredelenija* [GOST standard R 60.0.0.4 – 2019/ ISO 8373:2012 Robots and Robotic Devices. Terms and Definitions]. Moscow: Standartinform Publ., 2019. (In Russian)

Заключительный доклад Группы экспертов по Ливии, учрежденной резолюцией 1973 (2011) Совета Безопасности. S/2021/229. Представлен 8 марта 2021 г. URL: https://documents.un.org/doc/undoc/gen/n21/037/74/pdf/n2103774.pdf (дата обращения: 07.08.2024).

*Zakljuchitel'nyj doklad Gruppy jekspertov po Livii, uchrezhdennoj rezoljuciej 1973 (2011) Soveta Bezopasnosti. S/2021/229*. [Letter dated 8 March 2021 from the Panel of Experts on Libya Established pursuant to resolution 1973 (2011) addressed to the President of the Security Council] [https://documents.un.org/doc/undoc/gen/n21/037/74/pdf/n2103774.pdf, accessed on 07.08.2024] (In Russian)

*Перов В.Ю.* Этика искусственного интеллекта: некоторые проблемы междисциплинарной терминологии // Наука, технологии и ценности в неустойчивом мире: сборник научных статей / Научн. ред. и сост. И.Т. Касавин, Н.А. Ястреб, Л.В. Шиповалова. М.: Русское общество истории и философии науки, 2024. С. 409–413.

Perov, V.Ju. "Etika iskusstvennogo intellekta: nekotorye problemy mezhdisciplinarnoj terminologii" [Ethics of Artificial Intelligence: Some Problems of Interdisciplinary Terminology], *Nauka, tehnologii i cennosti v neustojchivom mire: sbornik nauchnyh statej* [Science, Technology and Values in an Unstable World: A Collection of Scientific Articles], eds. I.T. Kasavin, N.A. Jastreb, L.V. Shipovalova. Moscow: Russkoe obshhestvo istorii i filosofii nauki Publ., 2024, pp. 409–413. (In Russian)

Толковый словарь по искусственному интеллекту / Сост. А.Н. Аверкин, М.Г. Гаазе-Рапопорт, Д.А. Поспелов. М.: Радио и связь, 1992.

*Tolkovyj slovar' po iskusstvennomu intellektu* [Explanatory Dictionary of Artificial Intelligence], eds. A.N. Averkin, M.G. Gaaze-Rapoport, D.A. Pospelov. Moscow: Radio i svjaz' Publ., 1992. (In Russian)

*Потапов А.В.* (ред.) Сильный искусственный интеллект: на подступах к сверхразуму. М.: Альпина Паблишер, 2021.

Potapov, A.S. (ed.) *Sil'nyj iskusstvennyj intellekt: na podstupah k sverhrazumu* [Strong Artificial Intelligence: on the Way to Superintelligence]. Moscow: Al'pina Pablisher Publ., 2021. (In Russian)

*Федотова Ю.С.* Проблема возможности существования искусственного морального агента в контексте практической философии И. Канта // Кантовский сборник. 2023. Т. 42. № 4. С. 225–239.

Fedotova, Yu.S. "Problema vozmozhnosti iskusstvennogo moral'nogo agenta v kontekste prakticheskoi filosofii Kanta" [The Problem of the Possibility of an Artificial Moral Agent in the Context of Kant's Practical Philosophy], *Kantian Journal*, 2023, Vol. 42, Issue 4, pp. 225–239. (In Russian)

*Чалый В.А.* Первая формула категорического императива: от универсализма к фаллибилизму // Иммануил Кант и моральная философия / Под ред. А.В. Разина. Калининград: Изд-во БФУ им. И. Канта, 2024. С. 383–423.

Chaly, V.A. "Pervaja formula kategoricheskogo imperativa: ot universalizma k fallibilizmu" [The First Formula of the Categorical Imperative: from Universalism to Fallibilism], *Immanuil Kant i moral'naja filosofija* [Immanuel Kant and Moral Philosophy], ed. by A.V. Razin. Kaliningrad: IKBFU Press Publ., 2024, pp. 383–423. (In Russian)

*Ястреб Н.А.* Критический подход в этике искусственного интеллекта // Наука, технологии и ценности в неустойчивом мире: сборник научных статей / Научн. ред. и сост. И.Т. Касавин, Н.А. Ястреб, Л.В. Шиповалова. М.: Русское общество истории и философии науки, 2024. С. 448–450.

Jastreb, N.A. "Kriticheskij podhod v jetike iskusstvennogo intellekta" [Critical Approach to the Ethics of Artificial Intelligence], *Nauka, tehnologii i cennosti v neustojchivom mire: sbornik nauchnyh statej* [Science, Technology and Values in an Unstable World: A Collection of Scientific Articles], eds. I.T. Kasavin, N.A. Jastreb, L.V. Shipovalova. Moscow: Russkoe obshhestvo istorii i filosofii nauki Publ., 2024, pp. 448–450. (In Russian)

Barandiaran, X.E., Paolo, E. di, Rohde, M. “Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action”, *Adaptive Behavior*, 2009, Vol. 17, Issue 5, pp. 367–386.

Chauvier, S. “Why Individuality Matters”, *Individuals Across Sciences*, eds. A. Gay, T. Pradeu. New York: Oxford UP, 2016, pp. 25–45.

Damasio, A.R. *Descartes’ Error: Emotion, Reason, and the Human Brain*. New York: Avon Books, 1994.

Deonna, J.A., Teroni, F. *The Emotions. A Philosophical Introduction*. London; New York: Routledge, 2012.

Frankfurt, H. “Freedom of the Will and the Concept of a Person”, *Journal of Philosophy*, 1971, Vol. 68, Issue 1, pp. 5–20.

Powers, T.M. “Prospects for a Kantian Machine”, *IEEE Intelligent Systems,* 2006, Vol. 21, Issue 4, pp. 46–51.

Schlosser, M. “Agency”, *Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta (Winter 2019 Edition) [https://plato.stanford.edu/entries/agency/, accessed on 05.08.2024].

Schönecker, D. “Kant’s Argument from Moral Feelings: Why Practical Reason Cannot be Artificial”, *Kant and Artificial Intelligence*, eds. H. Kim, D. Schönecker. Berlin; Boston: De Gruyter, 2022, pp. 169–188.

Shoemaker, D. “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility”, *Ethics*, 2011, Vol. 121, Issue 3, pp. 602–632.

Wright, A.T. “Rightful Machines”, *Kant and Artificial Intelligence,* eds. H. Kim and D. Schönecker. Berlin; Boston: De Gruyter, 2022, pp. 223–238.