УДК 17

ПРОБЛЕМА ВОЗМОЖНОСТИ СУЩЕСТВОВАНИЯ ИСКУССТВЕННОГО МОРАЛЬНОГО АГЕНТА В КОНТЕКСТЕ ПРАКТИЧЕСКОЙ ФИЛОСОФИИ И. КАНТА

Ю. С. Федотова[1]

Вопрос о возможности существования искусственного морального агента предполагает обсуждение целого ряда проблем, поднятых И. Кантом в рамках практической философии и не исчерпавших своего эвристического потенциала до наших дней. Прежде всего это проблема соотношения морального закона и свободы. Так как разумное существо полагает свою волю независимой от внешних влияний, воля оказывается как подчиненной нравственному закону, так и автономной. Моральность и свобода предстают взаимосвязанными через независимость от внешнего. Соответственно, если действия искусственного интеллекта (ИИ) определяются чем-то или кем-то внешним по отношению к нему (человеком), то он действует не морально и не свободно, а гетерономно. Вследствие отсутствия у ИИ автономии и, соответственно, доступа к моральному закону у него нет и не может быть морального понимания, исходящего из морального закона. Другим следствием является отсутствие у него чувства долга, которое следовало бы из морального закона. Таким образом, моральное действие становится невозможным для искусственного морального агента, поскольку у него нет автономии и морального закона, морального понимания и чувства долга. Вывод состоит в том, что, во-первых, ИИ не только не может быть моральным с точки зрения Канта, но и не должен быть таковым, так как включение какого-либо морального принципа будет предполагать необходимость его выбора человеком, что сделает сам выбор принципа аморальным. Во-вторых, хотя воля как таковая у ИИ отсутствует, что с первого взгляда делает невоз-

[1] Московский государственный университет им. М. В. Ломоносова,
Россия, 119991, Москва, Ломоносовский проспект, д. 27, корп. 4.
Поступила в редакцию: 01.08.2023 г.
doi: 10.5922/0207-6918-2023-4-12

THE PROBLEM OF THE POSSIBILITY OF AN ARTIFICIAL MORAL AGENT IN THE CONTEXT OF KANT'S PRACTICAL PHILOSOPHY

Yu. S. Fedotova[1]

The question of whether an artificial moral agent (AMA) is possible implies discussion of a whole range of problems raised by Kant within the framework of practical philosophy that have not exhausted their heuristic potential to this day. First, I show the significance of the correlation between moral law and freedom. Since a rational being believes that his/her will is independent of external influences, the will turns out to be governed by the moral law and is autonomous. Morality and freedom are correlated through independence from the external. Accordingly, if the actions of artificial intelligence (AI) are determined by something or someone external to it (by a human), then it does not act morally and freely, but heteronomously. As a consequence of AI's lack of autonomy, and thus lack of access to the moral law, it does not and cannot have a moral understanding that proceeds from the moral law. Another consequence is that it has no sense of duty, which would follow from the moral law. Thus, moral action becomes impossible for the AMA because it lacks autonomy and moral law, moral understanding and sense of duty. It is concluded that, first, AMA not only cannot be moral, but should not be that, since the inclusion of any moral principle would imply the necessity for the individual to choose it, making the choice of the principle itself immoral. Second, although AI has no will as such, which prima facie makes not

[1] Lomonosov Moscow State University.
27-4 Lomonosovsky Prospekt, GSP-1, Moscow, 119991, Russia.
Received: 01.08.2023.
doi: 10.5922/0207-6918-2023-4-12

можными моральные и легальные действия, он все же может поступать легально в смысле соответствия юридическому закону, так как несет в себе квазиволю человека. Таким образом, необходимо, чтобы при создании ИИ соблюдались не моральные принципы, а юридическое законодательство, для которого приоритетны свобода и права человека.

Ключевые слова: *Кант, автономия, свобода, моральный закон, ИИ, искусственный моральный агент, категорический императив, моральное понимание*

only moral but also legal action impossible, it can still act legally in the sense of conforming to legal law, since AI carries a quasi-human will. Thus, it is necessary that the creation of AI should be based not on moral principles, but on legal law that prioritises human freedom and rights.

Keywords: *Kant, autonomy, freedom, moral law, AI, artificial moral agent, categorical imperative, moral understanding*

Введение

Стремительное развитие искусственного интеллекта (ИИ) неизбежно поднимает вопрос о том, смогут ли подобные технологии принимать моральные решения, за которые мы сможем наложить на них моральную ответственность; или иначе — возможен ли искусственный моральный агент (ИМА)? Некоторые исследователи считают, что разработка моральной машины (здесь и далее используется как синоним ИИ) и является целью машинной этики. Так, например, М. Андерсон и С.Л. Андерсон (Anderson, Anderson, 2007a; 2007б), Дж. Мур (Moor, 2006), Н. Бостром (Bostrom, 2014), К. Аллен (Allen et al., 2000; 2005; 2006) полагают, что нельзя исключать такое развитие ИИ, которое приведет к возможности совершения моральных поступков и принятию исключительно точных моральных решений. Вопрос о возможности (и даже необходимости) ИМА дает новую актуальность практической философии Канта, так как она может предложить ответ — такую точку зрения, которую необходимо учитывать при обсуждения этого вопроса.

В статье будет представлен тезис о невозможности ИМА в свете практической философии Канта, так как кантовское понимание моральности подразумевает исключительно свободное, конечное, разумное и чувственное существо, рассматривающее себя не только как часть природы, но и как часть интеллигибельного мира, где и находится идея свободы.

В первой части работы будет показано, что для того, чтобы обладать статусом морального агента, по Канту, нужно обладать моральным законом, который является «единственным фактом чистого разума, который возвещает о себе таким образом как изначально законодательный разум» (AA 05, S. 31; Кант, 1997a, с. 351), и которым может обладать только разумное существо, так как моральный закон предполагает наличие автономной воли, не зависящей от внешнего и имеющей причину в самой себе, что также является одной из формулировок категорического императива. Моральность связывается с понятием свободы и автономии: полагая себя моральным, человек полагает себя в то же время свободным; а полагать себя свободным, одновременно не полагая себя моральным, невозможно, поскольку моральный закон есть условие возможности хотя бы какого-то усвоения свободы. Значит, если ИИ находится исключительно в мире природы и его действия всегда направляются волей человека, то его действия никогда не имеют причину в самих себе и, следовательно, ИИ никогда не имеет автономной воли и одновременно не может быть моральным. Во второй части будет рассмотрено одно из следствий обладания автономией, вытекающих из ин-

терпретации кантовской этики Барбарой Херман, — моральное понимание (moral understanding)[2] ситуации (Herman, 1993), которым не обладает ИИ. Термин «моральное понимание» в контексте данного аргумента используется в смысле, который в него вкладывает Б. Херман. В третьей части рассматривается еще одно следствие автономии — наличие морального чувства долга, также отсутствующего у ИИ. Делается вывод, что у ИИ отсутствуют моральный закон, автономия и, как следствие, моральное понимание и моральное чувство долга, а значит, у него не может быть статуса морального агента, что влечет за собой невозможность морального поступка, но дает возможность рассуждать об ИИ с юридической точки зрения.

1. Моральный закон, свобода, автономия и ИИ

В «Основоположении к метафизике нравов» и «Критике практического разума» Кант определяет соотношение морали и свободы, которые оказываются взаимосвязанными понятиями. Согласно его рассуждению, каждое разумное существо, имеющее волю, то есть такое, которое полагает себя способным создавать предметы или определять их и рассматривает себя независимым от внешних влияний, должно иметь идею свободы и рассматривать себя свободным, чтобы причина его предметов и действий была в нем самом, и только так он сможет считать волю своей: «Разум должен рассматривать самого себя как творца своих принципов, независимо от чуждых влияний, следовательно, как практический разум или как воля разумного существа он должен представляться самому себе свободным, т.е. воля разумного существа только под идеей свободы может быть его собственной волей, и, следовательно, в практическом отношении она должна быть приписана всем разумным существам» (AA 04, S. 448; Кант, 1997б, с. 227), при этом «как разумное, следовательно, принадлежащее к интеллигибельному миру существо, человек никогда не может мыслить причинность своей собственной воли иначе, как обращаясь к идее свободы; ибо независимость от определяющих причин чувственного мира (каковую разум необходимо должен всегда приписывать самому себе) есть свобода. С идеей же свободы неразрывно связано понятие *автономии*, а с последним — всеобщий принцип нравственности, который в идее точно также лежит в основе всех действий *разумных* существ, как закон природы в основе всех явлений» (AA 04, S. 452—453; Кант, 1997б, с. 241). Свободная (автономная) воля и воля, подчиненная нравственным законам, по сути, оказываются одним и тем же, ведь и та и другая должны быть законодательствующими для самих себя и независимыми

[2] Херман выделяет несколько синонимичных понятий, в том числе *moral understanding* (моральное понимание) и *moral knowledge* (моральное знание). *Моральное понимание*, как и *знание*, используется для обозначения того факта, что субъект обладает пониманием основных моральных норм общества, которым он был обучен, и осознанием моральной проблемности ситуации перед началом осуществления процедуры категорического императива (КИ). Херман пишет: «КИ не может быть эффективным практическим принципом суждения, если только агент не обладает каким-то моральным пониманием (*moral understanding*) своих действий прежде, чем он начинает использовать процедуру КИ» (Herman, 1993, p. 77), и «полезно думать о моральном знании (moral knowledge), необходимом для кантовских агентов, как о знании своего рода морального правила. Давайте назовем их "правилами моральной значимости". Приобретенные как элементы морального воспитания, они структурируют восприятие ситуации агента таким образом, что то, что он воспринимает, есть мир с моральными характеристиками» (Ibid.). Херман также выделяет синонимичные понятия *moral sensitivity*, *moral perception* и *moral sensibility*, *moral awareness* (моральные чувствительность, восприятие, чувственность, осознанность), которыми обладает субъект, начинающий процедуру КИ. Это развитая способность субъекта распознавать ситуацию как морально проблематичную, и она является необходимой частью моральной агентности субъекта. Если субъект не воспринимает ситуацию как моральную, он не будет применять правила. Понятие *moral sensibility* Херман вводит для того, чтобы отделить его от понятия *emotional sensitivity*, поскольку моральная чувственность кантовского морального агента формируется через моральное знание правил. Моральная чувственность ставится выше эмоциональной чувствительности.

от внешних влияний. Кант делает вывод, что самозаконодательство воли для самой себя есть выражение категорического императива[3] (и принципа нравственности), поэтому свободная воля становится эквивалентом воли, подчиненной нравственному закону (AA 04, S. 446—447; Кант, 1997б, с. 223). Откуда получается, что свободная воля и моральность взаимосвязаны.

При этом моральный закон является *ratio cognoscendi* свободы, благодаря которому мы только и можем схватить, но никогда не объяснить ее существование, так как свободы нет в мире явлений, она находится в мире вещей в себе и не может быть наблюдаема. Сама свобода становится *ratio essendi* морального закона, так как без свободы он бы не мог существовать. Моральность предполагает свободу, а свобода влечет за собой моральность. При этом свобода (идея) и автономия (воля) определяются, как уже было сказано, через независимость от внешнего и через самозаконодательство. Если мы все же оказываемся под влиянием внешних причин, то мы уже не действуем морально, а лишь по побуждению или по склонности, тогда наша воля является не автономной, а гетерономной, так как определяется не внутренним, а внешним: «Природная необходимость была гетерономией действующих причин, так как каждый результат был возможен только по тому закону, что нечто другое определяло действующую причину к причинности, а чем же другим может быть свобода воли, если не автономией, т.е. свойством воли быть самой для себя законом?» (AA 04, S. 446—447; Кант, 1997б, с. 223). Из этого можно сделать вывод, что, хотя человек и обладает возможностью к автономии, это еще не значит, что он свободен, ведь в случае если он находится под внешними влияниями, его воля все еще не свободна. Можно также заметить, что речь у Канта идет о разумном конечном существе, чья воля только и может не согласовываться с моральным законом ввиду внешних отвлечений, в то время как бесконечное существо, например Бог, не испытывает подобных затруднений. Далее перед Кантом становится проблема сочетания автономии воли человека и детерминированного мира природы. Он решает данный вопрос так: человек, будучи существом природным и находясь в сфере явлений, определен временем, с этой стороны детерминирован. Однако, рассматривая себя как вещь в себе, которую создает Бог, человек приходит к тому, что «каждый поступок и вообще… в сознании его интеллигибельного бытия есть не что иное, как следствие, но отнюдь не определяющее основание причинности его как *ноумена*» (AA 05, S. 97—98; Кант, 1997а, с. 545), при этом человек не зависит и от Бога, так как тот создает его как вещь в себе, но не является причиной его поступков. То есть нужно принять, что Бог создает человека свободным и тот существует в двух сферах одновременно: мысля себя как вещь в себе, он свободен, но как феномен он детерминирован природой. Человек также есть единственное из известных подобных существ, а значит, и единственное, кому доступны моральный закон и автономия.

В сфере явлений, по Канту, целиком существуют животные, над которыми он возвышает людей. В связи с этим важно отметить, что воля имеет отношение к человеку, так как она есть «способность или создавать предметы, соответствующие представлениям, или хотя бы определять самое себя для их создания… т.е. свою причинность» (AA 05, S. 15; Кант, 1997а, с. 313), что в своем определении связывает волю с автономией (своя причинность) и исключает из рассмотрения животных, над которыми Кант человека возносит: «…он, кроме того, обладает разумом еще и для более высокого предназначения, а именно для того, чтобы не только принимать в соображение также и то,

[3] «Но положение: воля во всех действиях есть сама для себя закон — означает только принцип: не поступать иначе, как по той максиме, которая может иметь себя предметом также в качестве всеобщего закона. Но это есть как раз формула категорического императива и принцип нравственности; таким образом, свободная воля и воля, подчиненная нравственным законам, есть одно и то же» (AA 04, S. 446—447; Кант, 1997б, с. 223).

что́ есть доброе или злое само по себе… но и совершенно отличать эту оценку от первой и делать ее высшим условием первой» (AA 05, S. 61—62; Кант, 1997a, с. 435). Таким образом, обладая автономией, человек имеет практический разум и тем самым качественно отличается от других существ.

Кант также пишет о том, что если бы человек имел причину своих поступков не в самом в себе, а в Боге, то он был бы «марионеткой или автоматом Вокансона, сделанным и заведенным высшим мастером всех искусных творений» (AA 05, S. 101; Кант, 1997a, с. 555). Автоматы Вокансона — механические игрушки, среди которых были также экземпляры в виде людей, например флейтиста. Из слов Канта можно заключить, что он уже тогда задумывался о том, что такие человекоподобные механизмы могли бы выглядеть разумными, поэтому он добавляет, что даже если бы человек был такой мыслящей игрушкой, то он никогда по-настоящему не был бы свободным, потому что эта свобода была бы «лишь обманом, так как спонтанность может быть названа так только сравнительно, ибо, хотя ближайшие причины, определяющие его движения, и длинный ряд этих причин, восходящих к своим определяющим причинам, были бы внутренними, все-таки последняя и высшая причина находилась бы целиком в чужой власти» (AA 05, S. 101; Кант, 1997a, с. 555). Следуя данному рассуждению, можно вывести, что любое существо, которое имеет свою причину не в себе, а в чем-то другом (как животное определено только эмпирически, а механизм своим мастером), не может быть одновременно и свободным и моральным, так как и то и другое подразумевает необходимость самостоятельного законодательства для своей воли.

Нетрудно увидеть, как подобное рассуждение можно перенести в сферу ИИ. «Воля» машины на самом деле есть воля человека, который ее разработал, и она целиком и полностью им определяется, так же как механическая игрушка заводится мастером. Цели, которые может потенциально ставить себе ИИ, являются целями, поставленными ему человеком и нужными ему для воплощения именно человеческих желаний. Даже если принять гипотетическую возможность существования сильного ИИ, способного, допустим, к моральному суждению и действию, эту способность в него будет закладывать человек для решения какой-то своей цели, а это всегда материальные цели, носящие вторичный характер, потому что желают люди, но не машины. О. Улген отмечает, что как таковой воли у ИИ нет, поэтому он не сможет имитировать автономию (Ulgen, 2017, p. 76). Добавлю, что ИИ постоянно нуждаются в корректировке человеком в случаях, если программа работает не так, как надо, поэтому если ИИ вдруг ведет себя иначе, чем предполагается, это не проявление воли, а лишь ошибка, которую исправит человек.

Интересно отметить, что П. Стросон в работе «Свобода и обида» пишет, что люди используют по отношению к психически нездоровым людям объективную установку. Согласно Стросону, те, кто применяют подобную установку, рассматривают другого человека как объект, и это не позволяет им испытывать обиду за действия того, кто при иных условиях был бы признан моральным агентом. Моральная ответственность не накладывается, так как мы считаем их действия полностью детерминированными какими-то обстоятельствами (Стросон, 2020, с. 210). Подобную же объективную установку можно увидеть и у Канта, когда он говорит о животных и вещах, которые находятся в детерминированном мире природы и не обладают свободой. Причем человек также способен опуститься до уровня животного, если позволяет себе определяться внешними обстоятельствами. Поэтому можно понимать ИИ как инструмент, по отношению к которому мы применяем объективную установку, не рассматривая его в качестве морального агента, в то время как моральная ответственность падет на человека, его разработавшего, так как именно человек обладает автономией.

Некоторые исследователи, однако, наоборот, считают, что свобода воли не связывается с моральной агентностью (и даже просто агентностью). Так, например, Л. Флориди и Дж. Сандерс (Floridi, Sanders, 2004) пишут, что агентность (еще не моральная) определяется интерактивностью (interactivity) — способностью влиять на другого или окружающую среду и испытывать ответное влияние); автономией (autonomy) как возможностью самостоятельно изменять свое состояние, в том числе независимо от влияния окружающей среды, но, по факту, не свободой в кантовском смысле, так как подобное определение автономии не учитывает ее положительного момента — самостоятельного и независимого создания собственных правил; адаптивностью (adaptability) — обучаемостью или способностью изменять правила, ответственные за изменение состояния. И искусственный интеллект, как и человек, по мнению авторов, удовлетворяет этим условиям. Однако же Флориди и Сандерс ничего не пишут об интенциональности, когда говорят о влиянии на других и окружение, так что интерактивность не учитывает желание и направленность этой возможности взаимного влияния (при этом они упоминают об интенциональности, но в связи с «моральной агентностью», и отбрасывают ее как случайность). Второй взятый ими критерий — автономия — понимается не в смысле свободы и принятия самостоятельных, совершенно независимых решений (в том числе от воли другого агента), а в смысле способности к внутреннему изменению по собственным правилам, приводящим к перемене собственного состояния; однако, если мы учитываем наличие влияния чужой воли, то ИИ, будучи созданием человека, правила которому определяет именно последний, под критерий уже не попадает. Способность к адаптации, то есть к обучению, влияющему на изменение, довольно широка и подходит не только к человеку или ИИ, хотя авторы рассматривают все критерии в совокупности. Таким образом, Флориди и Сандерс вводят три условия для определения агентности, на чем строят дальнейшие рассуждения, по которым оказывается, что необязательно обладать свободой, чтобы быть агентом. При этом агентность саму по себе они отделяют от моральной агентности, а «моральная» ее часть позже добавляется через критерий причинения морального добра / зла. Вопрос об агентности оказывается решенным авторами на основании подбора удобных критериев; остается решить, кто является моральным агентом.

Для этого Флориди и Сандерс вводят понятие «уровень абстракции», означающий, что мы должны воспринимать окружающие объекты не так, как они есть перед нами, а абстрагируясь от всего лишнего, как, например, через камеру наблюдения, которая выводит лишь необходимые данные, удаляя все ненужное. Это определенный набор ограниченных наблюдений. Учитывая условие уровня абстракции, мы становимся свидетелями такой ситуации: один агент убивает пациента, другой его лечит, один из агентов — человек, другой — машина с ИИ, но наблюдатель не знает, кто есть кто, так как наблюдает через «камеру», показывающую лишь совершенные действия убийства и лечения. С точки зрения авторов, оба «агента» будут моральными, при условии что их действия причинили моральное зло или моральное добро (Floridi, Sanders, 2004, p. 15), при этом не принимается во внимание свобода воли и интенциональность у одного и их отсутствие у другого, что указало бы на то, что один агент поступает по собственному желанию и выбору, а другой — согласно программе (и не мог поступить иначе).

Одна из проблем заключается в определении морального действия как «возможности причинения морального зла и добра» (Ibid.). В этом случае причинение зла или добра определяется через вред или пользу, без учета качества воли агента. Это значит, что абсолютно любое действие подобного агента, которое причиняет вред / пользу, может быть эквивалентно моральному добру или злу, даже если агент не собирался причинять ни добро, ни зло — например, если он сделал что-то и случайным образом принес кому-то пользу, причинил зло под действием аффекта / болезни или причинил вред с целью сделать добро. Авторы также пишут, что подобное определение морально-

сти не является консеквенциалистским (Ibid.), однако, на мой взгляд, оно как раз фокусируется исключительно на последствиях действий (причинение добра / зла, то есть пользы / вреда), не принимая во внимание мотивы агентов, ведь один действует согласно своей программе, не испытывая при этом никаких чувств и ориентируясь на цель, заданную ему кем-то, а другой — исходя либо из долга, либо из склонности, которые предполагают изначальную способностью поступить иначе. Кроме того, непонятно, как быть с действиями, которые были сделаны ради причинения вреда, но оказались полезными, ведь, по логике авторов, их надо будет назвать моральным добром. Но даже если учитывать условия абстракции эксперимента, то мы знаем, что машина с ИИ, если только в ней не было заводской ошибки, могла быть только той, кто лечил пациента, потому что такова ее программа, о которой мы знаем, ведь иначе она не находилась бы в больнице, так как в машине, которая намеренно убивала бы людей в больнице, не было бы смысла. В таком случае у нас не возникнет сложностей с определением того, кто что делал в подобной ситуации, и тогда одно действие мы сможем назвать моральным (или, скорее, аморальным), а другое — нет, так как одно (лечение пациента) выполнялось согласно программе, а другое (убийство) — нет. При этом если бы и человек, и ИИ лечили пациента, то, по определению авторов, нужно было бы назвать оба действия моральным добром, хотя оба лишь выполняли свою работу. Другими словами, проблема эксперимента упирается в авторское определение морального действия, возведенного в статус аксиомы, но подобные рассуждения можно строить до бесконечности, не принимая во внимание их односторонность.

Другой проблемный момент данного эксперимента заключается в том, что Флориди и Сандерс намеренно вводят уровни абстракции, чтобы подогнать возможность морального действия ИИ. Авторы убирают весь контекст ситуации, считая, что и то и другое действие можно назвать моральным, потому что и то и другое причинило «моральное добро» и «моральное зло». Вынесение контекста за скобки намеренно упрощает мораль и моральные проблемы, позволяя искусственно ввести ИИ в категорию моральной агентности. Если бы мы даже не знали, что одно из действующих лиц в комнате — ИИ, и назвали бы оба действия моральными при искусственном условии уровня абстракции, то мы бы исправили свое суждение, узнав, что один из агентов был ИИ. Таким образом, эксперимент намеренно смоделирован так, чтобы не учитывать существование мотивов. Более того, моральный выбор — в реальности или в мысленном эксперименте — совершается человеком при условии различных деталей и контекста, а не при уровне абстракции, поэтому моральные дилеммы и вызывают сложность, и когда мы возвращаем вынесенные за скобки детали, проблема актуализируется вновь.

Работа Р. Тонкенса интересна тем, что автор говорит именно о кантовском ИМА и о возможности его создания, приходя к отрицательному выводу, который основывается на том, что у него отсутствует автономия, в противовес тому что пишут Флориди и Сандерс. Создание подобных ИМА также, по утверждению Тонкенса, аморально, потому что создание ИМА будет противоречить второй формулировке императива (Tonkens, 2009, p. 429), согласно которой человек никогда не должен относиться к человечеству в своем или чужом лице только как средству, но всегда как к цели. Согласно автору, если ИМА — средство человека, так как оно создается исключительно для достижения человеческих целей, то мы поступаем аморально, уже пытаясь создать подобного агента. Когда мы говорим об ИИ и ИМА, речь всегда идет о том, чтобы создать его исключительно на пользу человеку, будь то разработка какой-то программы или решение моральных вопросов за человека; ИА или ИМА всегда есть только и только средство, но никогда не цель. Значит ли это, что создавать ИИ в принципе аморально? Нет, если мы не пытаемся намеренно его морализировать, ведь создание механизмов как таковых не является аморальным.

Можно заметить, что те авторы, которые поднимают вопросы о моральности ИИ, как правило, делают это из страха, что ИИ будет поступать аморально, однако если для морали необходима одновременно и свобода, то это подразумевало бы возможность совершения зла, так как автономия состоит именно в независимости от склонностей, которые должны представлять активный интерес. По Тонкенсу, для того чтобы создать кантовского ИМА, нужно было бы как минимум позволить ему совершать аморальные вещи, что, конечно же, не является целью тех авторов, которые хотели бы создания ИМА (Ibid., p. 431). Необходимость преодоления собственных склонностей и постоянная проверка себя подразумевает, что человек может стать зависимым, теряя свою свободу. Таким образом, если мы пытаемся создать именно кантовский ИИ — целиком и полностью автономный (что, как уже было показано выше на примере марионетки Вокансона, невозможно) и моральный, — это также означает, что человек, создающий ИИ, должен будет заложить в него возможность причинения вреда человеку (и многое другое), которую ИИ должен будет как-то собственными усилиями преодолевать: у него должна быть возможность поступить иначе. Однако ИМА, по задумке его сторонников, должен быть создан как раз для того, чтобы ИИ не поступал аморально. Другими словами, создание ИМА противоречило бы самой цели его создания. Таким образом, проблема скорее не в том, что попытки создания ИМА противоречили бы категорическому императиву, как пишет Тонкенс, а в том, что они были бы бессмысленными. Тонкенс также считает, что искать моральные основания для ИМА необходимо не в кантовской практической философии, однако, на мой взгляд, это показывает, что желание создать ИМА в принципе невозможно исполнить до тех пор, пока ИИ несет в себе волю человека, а если иное невозможно, то невозможен и ИМА.

Существует другой взгляд на необходимость создания ИИ: Дж. Уайт в противовес Тонкенсу считает, что создание кантовского ИМА не просто возможно, но и является человеческой обязанностью, так как ИИ может себя развить в морального агента (White, 2022, p. 666). По мнению Уайта, ИМА должны получить «достоинство, уважение и даже (человеческие) права» (Ibid., p. 672), однако проблеме отсутствия автономии у ИИ он не уделяет внимания. Если создание действительно автономного и, как неизбежное следствие, морального ИМА невозможно, то нельзя заключить, что ИИ может развиться в морального агента, как и нельзя вывести необходимость того, что человек обязан его развивать в такового. Д. Калверлей также пишет о возможности рассмотрения «небиологических машин» с юридической точки зрения и считает, что они могут быть юридическими лицами (legal persons). По его мысли, в случае если программа каким-то образом (сама) пытается получить информацию нелегальным способом по своей воле (Calverley, 2008, p. 533), например чтобы порадовать человека, тем самым она проявляет некоторую автономию, так что закон мог бы расценивать ее действия и рассматривать ее как ответственную. Тем не менее сложно представить себе программу, которая сама решает прибегнуть к вредоносным действиям, если это в ней изначально не прописано или не упущено, так как именно программист был тем, кто писал программу и не предусмотрел подобный исход, поэтому непонятно, почему ответственность должна переноситься на кого-либо другого (и сам автор ставит под вопрос (хотя и временный), возможно ли осуществление подобного рода действий у небиологических систем, однако смотрит на это довольно позитивно, так как считает, что теоретически это не невозможно). Даже если произошла некая ошибка и сбой в работе ИИ, то, скорее всего, проблема в том, что его вовремя не скорректировали. Эта ошибка легко убирается волей человека и приводится в соответствие с ней при последующей корректировке, а значит, она не является существенной. Д. Джонсон (Johnson, 2006) считает, что, хотя компьютерные программы не обладают агентностью, они все же могут являться «моральными сущностями» (moral entities) и подлежать моральной оценке, поскольку человек, создавая каку-

ю-то технологию, делает это с намерением, и если программа была разработана с целью украсть / навредить, то она не может быть расценена как «нейтральная» сама по себе. Разница между тем, что пишет Джонсон, и тем, что пишет Калверлей, заключается в том, что Калверлей не учитывает в своем примере проявления автономии и намерения создателя программы. Рассуждения Джонсон, на мой взгляд, охватывают более полную картину, так как предлагают моральное оценивание программы в свете того, с какой целью ее создавали и как ее использовали, то есть в совокупности, хотя автор и наделяет сам механизм интенциональностью. Можно также отметить, что, принимая во внимание действия и создателя, и программы, и пользователя, Джонсон частично выступает против некоторых уровней абстракции Флориди и Сандерса, заявляя, что не во всех случаях уровни абстракции релевантны относительно морального оценивания, так как необходимо учитывать создателей программ и контекст их использования в социальных практиках (Johnson, 2006, p. 198). Добавлю, что как раз связь созданной программы со своим создателем, а также с тем, для какой цели она была создана, отсылает к необходимости соблюдения законов при создании подобных программ, о чем будет сказано в п. 4.

2. Моральное понимание и ИИ

Другой важный аспект принятия моральных решений, в соответствии с концепцией Б. Херман (Herman, 1993), — это моральное понимание (moral understanding) ситуации, которое неизбежно связано с автономией воли. Херман считает, что категорическому императиву, который является моральным рассуждением, изначально предшествует определенное моральное понимание ситуации агентом, которое, в свою очередь, формируется посредством моральной тренировки, структурирующей определенное моральное восприятие у агента. Таким образом, прежде чем стать моральным агентом, человек должен пройти тренировку, что очень напоминает подход Витгенштейна к языковым играм, связанным с формами жизни и обучением. Если ребенку несколько раз объяснили, что причинять вред — плохо, то он должен считать это плохим, так же как если бы ребенку объяснили, что такое «пять», «красное» и «яблоко». В случае если ребенок не понимает, что причинять вред — плохо, согласно Херман, это лишь моральная патология, к чему можно отнести психические расстройства, не позволяющие потенциальному агенту осознать и понять моральную ситуацию. Подобный подход будет означать, что мы учимся моральности так же, как обучаемся языку, и получаем статус морального агента только в случае успешной тренировки в восприятии ситуаций с моральной точки зрения. По такой логике, успешность моральной тренировки будет зависеть от того, распознает ли агент, подходит случай под моральные категории или нет. Л. Нагль поддерживает точку зрения Херман о «герменевтической чувствительности субъекта к контексту предполагаемого поступка» (Нагль, 2022, с. 79) и делает заключение о том, что «алгоритмы ИИ… неспособны *полностью* "оценить" сложную социальную фактуру "ситуации"» и «неспособны самостоятельно определить и тем более доказать *обязательность* выполняемых ими законов, так как являются гетерономными получателями "команд"» (Там же, с. 81).

И все же, на мой взгляд, если моральности необходимо учить, как об этом пишет Херман, то обучение зависит от контекста, что отсылает к возможности разного определения моральности ситуаций, потому что в зависимости от социального и политического контекста определения блага и зла могут разниться. Однако это противоречит универсальности и общеобязательности морального закона, на чем Кант делал особый акцент. Более того, если моральности сначала необходимо обучить, чтобы у человека появилась способность распознать моральные ситуации, то не значит ли это, что его воля будет изначально находиться под влиянием и не будет автономной?

Ведь если мы применяем кантовский подход, то осознание моральности ситуации должно происходить не из-за внешних причин, а исключительно из внутренних. Моральный закон как таковой, хотя и не может быть познан конкретно, также познается у Канта внутренне. В случае если закон дается чем-то внешним, то воля автоматически перестает быть автономной, так как она определена внешним влиянием.

Тем не менее, хотя обучение моральным правилам и моральной ответственности не совсем соответствует кантовскому подходу, все же нельзя отрицать, что агент изначально должен распознавать ситуацию как моральную, правда, исходя из внутреннего осознания. Более того, он станет прибегать к категорическому императиву только тогда, когда изначально сомневается в правильности своего выбора или нуждается в моральном оправдании. Херман считает, что человек не использует императив каждый день и при каждом своем выборе, однако он становится таким моральным принципом, который приходит на помощь тогда, когда агент не в состоянии принять решение. Таким образом, для того чтобы агент использовал категорический императив, он должен испытывать сомнение и как бы изначально понимать, что то, что он хотел бы выбрать, возможно, является морально недопустимым, и, нуждаясь в некотором моральном поощрении, использовать категорический императив. Если посмотреть на то, как Кант отвергает оправдание лжи, применяя категорический императив, то подобный вывод, в ходе которого мы приходим как бы к противоречию, строится на логическом рассуждении. Собирающийся лгать человек, изначально не испытывающий сомнений на этот счет, вряд ли бы стал проделывать эту процедуру. Соответственно, поскольку ИИ должен будет испытывать сомнения в правильности своего выбора (что не только невозможно, но и будет подразумевать, что он может поступить аморально), это уничтожит смысл создания ИМА в принципе.

В некоторых случаях категорический императив может использоваться неосознанно для оправданий и морально недопустимых действий, как бы убеждая агента в том, что его действия правильны. О подобном использовании императива пишет Дж. Батлер в своей работе «Сила ненасилия». Согласно Батлер, применению императива предшествует размышление, предполагающее воображение мира, в котором мы не хотим жить и действительно делаем то, что лишь собираемся сделать. Этот воображаемый мир влияет на наше восприятие ситуации, которое «не до конца управляется сознанием» (Батлер, 2022, с. 92). Например, подобное использование императива можно увидеть в логике расистов, которые оправдывают выстрел в спину безоружному человеку с другим цветом кожи тем, что они вообразили, что уже живут в мире, в котором не хотели бы жить, и люди с отличающимся цветом кожи изначально, по их мнению, представляют угрозу. Таким образом, чтобы избежать опасности, которую они придумали себе в голове (расистский фантазм), они решают выстрелить в безоружного человека как бы в качестве самозащиты. Категорический императив здесь не только используется не совсем осознанно, поскольку на него влияет расистский фантазм об опасности, но и оправдывает подобное решение, ведь человек в этом примере искренне готов жить в мире, где убийство на почве ненависти является законом, так как он уже воображает, что не живет в идеальном для себя мире. «Правильное» (то есть как его задумывал Кант) использование категорического императива не предполагает искаженной веры агента в то, что он и так готов жить в таком мире, где люди лгут или причиняют вред другим людям, поэтому можно дополнить рассуждение Херман и Нагля, добавив, что, хотя применению императива должно предшествовать моральное понимание, это относится только к «правильному» истолкованию категорического императива, где моральный агент действительно искренне сомневается в моральном выборе и не имеет искаженных представлений.

3. Чувство долга и ИИ

Моральный выбор у Канта сопровождается не просто пониманием и рассуждением, но и моральным чувством. Я понимаю, что вредить другим плохо, не просто исходя из сухого факта, что другому человеку нанесен вред и это почему-то должно категоризироваться мной как «плохой» поступок (это умение можно было бы и натренировать), но потому? что моральный закон, вызывая во мне уважение и трепет, запрещает так поступать, и я обязана его послушаться, если только моя воля при этом автономна. Кант также пишет, что моральный закон причиняет и страдание, так как смиряет и унижает самомнение человека: «Мы можем a priori усмотреть, что моральный закон как определяющее основание воли ввиду того, что он наносит ущерб всем нашим склонностям, должен породить чувство, которое может быть названо страданием» (AA 05, S. 72—73; Кант, 1997a, с. 468—469), при этом он не только с объективной стороны определяет поступок, но и «есть вместе с тем и субъективное определяющее основание, т.е. побуждение к этому поступку, так как он оказывает влияние на чувственность субъекта и возбуждает чувство, которое содействует влиянию закона на волю» (AA 05, S. 75—76; Кант, 1997a, с. 475). Следует отметить, что моральное чувство не предшествует моральному закону, а появляется как его действие, поэтому если нет морального закона, который базируется на полагании свободы, не может быть и морального чувства. В то же время уважение, которое вызывает моральный закон, имеется только у чувственных и конечных разумных существ, значит, у Бога, как существа, свободного от чувственности, не может быть к нему уважения. Кроме того, уважение люди могут испытывать «только к людям и никогда — к вещам» (AA 05, S. 76; Кант, 1997a, с. 479), так как на примере людей мы, с одной стороны, видим исполнение закона, а с другой — сами себя стыдим на этом основании. Причем к вещам Кант также относит и животных, к которым можно испытывать либо склонности, либо аффекты. Можно полагать, что ИИ, будучи подобной же вещью в терминах Канта, так же может вызывать у нас только склонности, любовь или, что больше всего проявляется у технопессимистов, страх. Таким образом, моральное рассуждение сопровождается не просто пониманием ситуации, но и моральными чувствами, которые отсутствуют у ИИ даже в сильной интерпретации, поскольку являются следствием внутреннего осознания морального закона, что возможно лишь у тех чувственных разумных и конечных существ, которые находятся не только в сфере явлений, но и в сфере вещей в себе.

Л. Беносси и С. Бернекер настаивают, что аргумент об отсутствии эмоциональной жизни у роботов не работает, так как существуют люди, которые не испытывают эмоций или морального чувства, но тем не менее несут моральную ответственность и обладают статусом морального агента, например социопаты, в то время как дети, хотя и имеют эмоции и чувства, до определенного момента не рассматриваются как морально ответственные агенты; из этого авторы делают вывод, что наличие эмоциональной жизни может быть необязательным для того, чтобы обладать статусом морального агента (Benossi, Bernecker, 2022, p. 150). И действительно, одного только чувства недостаточно, однако авторы не учитывают, что оно становится прямым следствием наличия морального закона, из которого вытекает рассмотрение своей воли как свободной. Следуя Канту, невозможно не обладать моральным чувством, сопровождающим моральное действие, и быть автономным, поэтому, хотя моральное чувство не является достаточным условием, все же оно необходимо. Без морального чувства категорический императив был бы простым подведением под правило, однако это не было бы применением именно категорического императива, о чем писали Херман и Нагль. Как нельзя научить моральному пониманию ситуации, так нельзя научить обладать внутренним чувством. Оно приходит из осознания в себе морального закона и ощущения уважения к нему, для чего необ-

ходимо не простое знание формулы императива, которая сама по себе ничего не дает, но моральное понимание его значения и использования, а также доступ к моральному закону, которого нет в мире явлений (Schönecker, 2022, p. 178).

4. Законное, не моральное действие ИИ

Хотя машины неспособны к моральному действию, на мой взгляд, они вполне способны на легальное. Правда, это несколько расходится с кантовским взглядом, поскольку легальный (юридически законный) поступок предполагает волю, определяемую «только посредством чувства… следовательно, совершается *не ради закона*» (AA 05, S. 71; Кант, 1997a, с. 465). То есть для легального поступка необходимо иметь волю и связанное с материальностью чувство, например страха, который толкает на совершение поступка. Так как воля, согласно Канту, связана с наличием разума, данного человеку для какой-то высшей, отличной от животной цели, и способности морального понимания зла и добра, то она может принадлежать только человеку. Другими словами, при легальном поступке всегда остается возможность моральности, если бы человеческий агент повел себя не из чувственности, связанной с материальностью, которая убирает любую возможность моральности, а из долга. Поэтому у ИИ невозможен ни легальный в указанном смысле, ни моральный поступок. Получается, что ИИ не способен ни к моральному, ни к легальному поступку ввиду отсутствия у него какой бы то ни было воли. Однако мне представляется, что он все же должен быть способен на легальный поступок, поскольку, хотя воля у самого ИИ отсутствует, она может в нем присутствовать, будучи выражением воли человека. Другими словами, несмотря на то что философия Канта не подразумевает никакого другого (не только морального, но и легального) агента, кроме разумного конечного существа — человека, ИИ все же может и должен поступать легально, так как, выражая волю людей, его поведение должно соответствовать законодательству. Поскольку действия ИИ всегда обусловлены чужой волей, то вопросы могут возникнуть уже к качеству вложенной в него человеческой воли. Соответственно, если ИИ вынужден принимать решения в неоднозначных для человека ситуациях (пусть даже это решение, кому жить, а кому умереть, в проблеме вагонетки), важно, чтобы он выбирал не на основании предубеждений, перекочевавших в него из человека, а исходя из соответствия решения закону. К. Аллен, задаваясь вопросом вагонеток и беспилотных машин, утверждает, что проблема должна с необходимостью решаться при помощи создания ИМА, которые должны «чтить частную жизнь, поддерживать общественные этические стандарты, защищать гражданские права, индивидуальную свободу и дальнейшее процветание людей» (Allen, 2006, p. 13). Однако эти требования возможно соблюдать без учета морали (которой искусственный агент не может обладать ввиду отсутствия свободы), лишь соблюдая закон. А.Т. Райт, рассуждая о проблеме беспилотной вагонетки, также делает вывод о том, что машина должна вести себя в первую очередь законно (rightful) (Wright, 2022, p. 228), так как это решит большинство проблем, которые могут возникнуть с ИИ, а значит, нет необходимости искать моральный принцип для ИИ или создавать ИМА.

Э. Шмидт, осмысляя этот же вопрос, приходит к выводу, что, по Канту, программист не должен прописывать, кого конкретно и на каких условиях должен переезжать автомобиль в случае, если столкнется с подобным выбором: ведь ни водитель (в случае с вагонеткой), ни посторонний наблюдатель не являются ответственными за убийство, если они выбирают не поворачивать рычаг, так как изначально вагонетка идет по выбранному пути, поэтому именно поворот рычага будет считаться неоправдываемым в моральном смысле убийством (Schmidt, 2022, p. 198). Кроме того, выделяя широкий (более свободный в исполнении) и узкий (более строгий) долг, чтобы избежать между

ними конфликта, Шмидт заключает, что в данном случае узкий долг «не убить» перевешивает широкий долг «спасти» (Ibid.), поэтому программист не должен решать, кого спасти, ведь это уже само по себе будет аморальным. Можно прийти к выводу, что включение этического момента в работу машин может быть недопустимым, так как потребует от человека сделать конкретные выборы в чью-то пользу, и вряд ли кто-то захочет быть в меньшей группе людей, которую можно будет переехать в случае необходимости. Из этого следует, что создатель не должен выбирать никакого иного принципа при создании ИИ, кроме соответствующего юридическому законодательству и в первую очередь соблюдающего права и свободы каждого человека. Конечно, здесь встает проблема того, что законодательства разных стран могут отличаться и разные программы на основе ИИ могут быть законными в одной стране, но считаться недопустимыми в другой. Однако этот вопрос затрагивает более широкую проблему возможности универсального законодательства относительно создания ИИ и потенциального усложнения международных отношений в контексте использования разных принципов при создании ИИ, что возвращает нас к необходимости контроля над созданием искусственных технологий, симулирующих человеческий интеллект. Такой контроль необходим не потому, что ИИ может поступить аморально по отношению к человеку, но потому, что он может быть использован людьми не в самых благих целях — особенно с учетом того, что большие компании вряд ли упустят возможность развития и монетизации новых технологий. Тем самым актуализируется целый ряд вопросов: вопрос авторских прав при создании ИИ-генерированных работ, вопрос сбора личной информации (в связи с возможностью утечки данных), вопрос еще более агрессивной таргетированной рекламы и другие проблемы активно развивающихся технологий.

Заключение

Обозначенная выше проблема возможности искусственного морального агента подвела нас к тому, что подобный агент с точки зрения Канта невозможен, потому что кантовская философия предполагает именно разумного (человеческого) агента, способного как к моральным, так и к легальным поступкам ввиду необходимого наличия у разумного существа воли, которая сама по себе у ИИ отсутствует. Было показано, что ИИ не может обладать свободой и автономией воли и, следовательно, моральностью, поскольку находится в детерминированном мире природы и его действия во всем определяются волей человека, в силу чего он не способен к самозаконодательству и во всем зависит от своего разработчика. Не имея моральности и автономии, ИИ не действует из долга, исходящего из уважения к моральному закону, так как не может иметь доступа к таковому и не может морально понимать ситуацию. Кантовский ответ не терпит искусственной агентности в морали еще и потому, что к ней имеет доступ только тот, кто находится не только в сфере явлений, как животное, но тот, кто также существует и в мире вещей в себе. Таким образом, ИИ, находясь исключительно в сфере чувственно-воспринимаемого, не имеет причины своих действий только в себе, следовательно, не может быть автономным и моральным. Тем не менее, хотя ИИ не может быть моральным, он может и должен действовать согласно юридическим законам, поскольку, не обладая собственной волей, является носителем чужой. Это подводит к проблеме контроля за созданием ИИ и соответствием его действий в первую очередь законам, а не моральному принципу, а также ставит вопрос о современных вызовах относительно контроля над ИИ. Можно дополнительно отметить, что данная проблема поднимает вопрос о соотношении автономии и гетерономии человека в современном мире с учетом развития не только ИИ, но и в целом цифровых технологий, влияющих на принятие человеком решений, что может оказаться особенно значимым для политической сферы.

Список литературы

Батлер Дж. Сила ненасилия: сцепка этики и политики / пер. с англ. И. Кушнаревой. М. : Изд. дом ВШЭ, 2022.

Кант И. Критика практического разума // Соч. на нем. и рус. яз. М. : Московский философский фонд, 1997а. Т. 3. С. 279—733.

Кант И. Основоположение к метафизике нравов // Соч. на нем. и рус. яз. М. : Московский философский фонд, 1997б. Т. 3. С. 41—275.

Нагль Л. Цифровые технологии: размышления о различии между инструментальной рациональностью и практическим разумом // Кантовский сборник. 2022. Т. 42, № 1. С. 60—88.

Стросон П. Свобода и обида / пер. с англ. Е. Логинова // Финиковый Компот. 2020. № 15. С. 204—221. doi: 10.24412/2587-9308-2020-15-204-221.

Allen C., Varner G., Zinser J. Prolegomena to Any Future Artificial Moral Agent // Journal of Experimental and Theoretical Artificial Intelligence. 2000. Vol. 12. P. 251—261.

Allen C., Smit I., Wallach W. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches // Ethics and Information Technology. 2005. Vol. 7. P. 149—155. doi: 10.1007/s10676-006-0004-4.

Allen C., Wallach W., Smit I. Why Machine Ethics? // IEEE Intelligent Systems. 2006. Vol. 21, № 4. P. 12—17.

Anderson M., Anderson S. L. The Status of Machine Ethics: a Report From The AAAI Symposium // Minds and Machines. 2007a. Vol. 17. P. 1—10. doi: 10.1007/s11023-007-9053-7.

Anderson M., Anderson S. L. Machine Ethics: Creating an Ethical Intelligent Agent // AI Magazine. 2007б. Vol. 28. № 4. P. 15—26.

Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford : Oxford University Press, 2014.

Benossi L., Bernecker S. A Kantian Perspective on Robot Ethics // Kant and Artificial Intelligence. Berlin ; Boston : De Gruyter, 2022. P. 147—169.

Calverley D. J. Imagining a Non-Biological Machine as a Legal Person // AI & SOCIETY. 2008. Vol. 22, № 4. P. 523—537.

Floridi L., Sanders J. W. On the Morality of Artificial Agents // Minds and Machines. 2004. Vol. 14, № 3. P. 1—29.

Herman B. The Practice of Moral Judgement. L. : Harvard University Press, 1993.

References

Allen, C., Smit, I. and Wallach, W., 2005. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7, pp. 149-155. https://doi.org/10.1007/s10676-006-0004-4.

Allen, C., Varner, G. and Zinser, J., 2000. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12, pp. 251-261. https://doi.org/10.1080/09528130050111428.

Allen, C., Wallach, W. and Smit, I., 2006. Why Machine Ethics? *IEEE Intelligent Systems*, 21(4), pp. 12-17. https://doi.org/10.1109/MIS.2006.83.

Anderson, M. and Anderson, S.L., 2007a. The Status of Machine Ethics: A Report from the AAAI Symposium. *Minds and Machines*, 17, pp. 1-10. https://doi.org/10.1007/s11023-007-9053-7.

Anderson, M. and Anderson, S.L., 2007b. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28(4), pp. 15-26. https://doi.org/10.1609/aimag.v28i4.2065.

Bostrom, N., 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Benossi, L. and Bernecker, S., 2022. A Kantian Perspective on Robot Ethics. In: H. Kim and D. Schönecker, ed. 2022. *Kant and Artificial Intelligence*. Berlin & Boston: De Gruyter, pp. 147-169. https://doi.org/10.1515/9783110706611-005.

Butler, J., 2022. *Sila nenasiliya: Stsepka etiki i politiki* [*The Force of Nonviolence: An Ethico-Political Bind*]. Translated from English by I. Kushnareva. Moscow: HSE Press. (In Rus.)

Calverley, D. J., 2008. Imagining a Non-Biological Machine as a Legal Person. *AI & Society*, 22(4), pp. 523-537. https://doi.org/10.1007/s00146-007-0092-7.

Floridi, L. and Sanders, J. W., 2004. On the Morality of Artificial Agents. *Minds and Machines*, 14(3), pp. 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Herman, B., 1993. *The Practice of Moral Judgement*. London: Harvard University Press.

Johnson, D.G., 2006. Computer Systems: Moral Entities but Not Moral Agents. *Ethics and Information Technology*, 8, pp. 195-204. https://doi.org/10.1007/s10676-006-9111-5.

Kant, I., 1997a. The Critique of Practical Reason. In: I. Kant, 1997. *Sochineniya v 4-h tomah na nemetskom i russkom yazykakh* [*Works in 4 Volumes in German and Russian Languages*]. Volume 3. Moscow: Moskovskii filosofskii fond, pp. 279-733. (In Rus.)

Kant, I., 1997b. Groundwork of the Metaphysics of Morals. In: I. Kant, 1997. *Sochineniya v 4-h tomah na nemetskom i russkom yazykakh* [*Works in 4 Volumes in German and Russian Languages*]. Volume 3. Moscow: Moskovskii filosofskii fond, pp. 41-275. (In Rus.)

Moor, J. H., 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21(4), pp. 18-21. https://doi.org/10.1109/MIS.2006.80.

Johnson D. G. Computer Systems: Moral Entities But Not Moral Agents // Ethics and Information Technology. 2006. № 8. P. 195—204. doi: 10.1007/s10676-006-9111-5.

Moor J. H. The Nature, Importance, And Difficulty of Machine Ethics // IEEE Intelligent Systems Special Issue on Machine Ethics. 2006. Vol. 21, № 4. P. 18—21.

Schönecker D. Kant's Argument from Moral Feelings: Why Practical Reason Cannot Be Artificial // Kant and Artificial Intelligence. Berlin ; Boston : De Gruyter, 2022. P. 169—189.

Schmidt E. E. Kant on Trolleys and Autonomous Driving // Kant and Artificial Intelligence. Berlin ; Boston : De Gruyter, 2022. P. 189—223.

Tonkens R. A Challenge for Machine Ethics // Minds and Machines. 2009. Vol. 19, № 3. P. 421—438. doi: 10.1007/s11023-009-9159-1.

Ulgen O. Kantian Ethics In The Age of Artificial Intelligence and Robotics // QIL, Zoom-in. 2017. Vol. 43. P. 59—83.

White J. Autonomous Reboot: Kant, The Categorical Imperative, and Contemporary Challenges For Machine Ethicists // AI & Society. 2022. Vol. 37, № 2. P. 661—673.

Wright A. T. Rightful Machines // Kant and Artificial Intelligence. Berlin ; Boston : De Gruyter, 2022. P. 223—239.

Об авторе

Юлия Сергеевна ***Федотова***, философский факультет, Московский государственный университет им. М. В. Ломоносова, Москва, Россия.

E-mail: fedotovajs@gmail.com

Для цитирования:

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

doi: 10.5922/0207-6918-2023-4-12

Nagl, L., 2022. Digital Technology: Reflections on the Difference between Instrumental Rationality and Practical Reason. *Kantian Journal*, 41(1), pp. 60-88. http://dx.doi.org/10.5922/0207-6918-2022-1-3.

Schönecker, D., 2022. Kant's Argument from Moral Feelings: Why Practical Reason Cannot Be Artificial. In: H. Kim and D. Schönecker, ed. 2022. *Kant and Artificial Intelligence*. Berlin & Boston: De Gruyter, pp. 169-189. https://doi.org/10.1515/9783110706611-006.

Schmidt, E. E., 2022. Kant on Trolleys and Autonomous Driving. In: H. Kim and D. Schönecker, ed. 2022. Kant and Artificial Intelligence. Berlin & Boston: De Gruyter. pp. 189-223. https://doi.org/10.1515/9783110706611-007.

Strawson, P., 2020. Freedom and Resentment. Translated into Russian by E. Loginov. *Date Palm Compote*, 15, pp. 204-221. https://doi.org/10.24412/2587-9308-2020-15-204-221. (In Rus.)

Tonkens, R., 2009. A Challenge for Machine Ethics. *Minds and Machines*, 19(3), pp. 421-438. https://doi.org/10.1007/s11023-009-9159-1.

Ulgen, O., 2017. Kantian Ethics in the Age of Artificial Intelligence and Robotics. *QIL, Zoom-in*, 43, pp. 59-83. Available at: <http://www.qil-qdi.org/wp-content/uploads/2017/10/04_AWS_Ulgen_FIN.pdf> (Accessed 23.07.2023).

White, J., 2022. Autonomous Reboot: Kant, the Categorical Imperative, and Contemporary Challenges for Machine Ethicists. *AI & Society*, 37(2), pp. 661-673. https://doi.org/10.1007/s00146-020-01142-4.

Wright, A. T., 2022. Rightful Machines. In: H. Kim and D. Schönecker, ed. 2022. Kant and Artificial Intelligence. Berlin & Boston: De Gruyter, pp. 223-239. https://doi.org/10.1515/9783110706611-008.

The author

Yulia Sergeevna ***Fedotova***, Faculty of Philosophy, Lomonosov Moscow State University, Moscow, Russia.

E-mail: fedotovajs@gmail.com

To cite this article:

Fedotova, Yu. S., 2023. The Problem of the Possibility of an Artificial Moral Agent in the Context of Kant's Practical Philosophy. *Kantian Journal*, 42(4), pp. 225-239. (In Rus.)

http://dx.doi.org/10.5922/0207-6918-2023-4-12