УДК 141; 179
DOI: 10.15372/PS20240506
EDN: LAMTQY

ЛИЗА БЕНОССИ, СВЕН БЕРНЕКЕР. КАНТИАНСКИЙ ВЗГЛЯД НА РОБОЭТИКУ[1]

(Перевод А. С. Зайковой)

Каким условиям должен удовлетворять робот, чтобы считаться моральным агентом? Должны ли роботы стать моральными агентами, или же человечество должно полностью сохранить для себя агентность и индивидуальность? Допустимо ли препятствовать развитию нравственных качеств роботов? В статье эти вопросы рассматриваются с нейтральной и кантианской точки зрения. Что касается первого вопроса, то мы утверждаем, что роботы не способны соответствовать кантианским стандартам моральной агентности. На второй и третий вопросы ответить труднее, отчасти потому, что нейтральная точка зрения не дает четкого вердикта. Мы утверждаем, что кантианская позиция позволяет нам выдвинуть правдоподобный ответ. Идея заключается в том, что помешать роботам достичь нравственной личности морально допустимо, поскольку наше намерение соответствует уважению человеческой жизни и ее разумной природы.

Ключевые слова: робоэтика, агентность, кантианская точка зрения, нейтральная точка зрения, моральный агент, роботы

LISA BENOSSI & SVEN BERNECKER. A KANTIAN PERSPECTIVE ON ROBOT ETHICS

(Translated by A. S. Zaykova)

What conditions does a robot have to satisfy to qualify as a moral agent? Should robots become moral agents, or should humanity fully retain agency and personhood for itself? Is it permissible to prevent robots from developing moral agency? This paper examines these questions from a viewpoint-neutral and a Kantian perspective. Regarding the first question, we argue that the Kantian standards for moral agency could not possibly be met by robots. The second and third questions are more difficult to answer, in part because the viewpoint-neutral perspective does not provide a clear verdict. We argue that it is a feature of the Kantian perspective to propose a plausible answer. The idea is that preventing robots from achieving moral

[1] Benossi, Lisa & Bernecker, Sven (2022). 5 A Kantian Perspective on Robot Ethics. In Hyeongjoo Kim & Dieter Schönecker (eds.), Kant and Artificial Intelligence. De Gruyter. pp. 145–168. https://www.degruyter.com/document/doi/10.1515/9783110706611-005/html. Перевод с англ. А.С. Зайковой.

personality is morally permissible, insofar as our intention is consistent with the respect of human life and its rational nature.

Keywords: robot ethics, agency, Kantian viewpoint, neutral viewpoint, moral agent, robots

1. Введение

Термин *«робоэтика»* может использоваться в разных значениях. Он может относиться к профессиональной этике робототехников, к моральному кодексу, запрограммированному в роботах, к способности роботов рассуждать этически, а также к моральным вопросам, связанным с проектированием и разработкой роботов. В данной работе используется именно последнее применение термина. В частности, нас интересует вопрос о том, допустимо ли препятствовать развитию у роботов моральной агентности.

Что такое роботы? Это простой вопрос: мы знаем, что робот использует датчики для выявления элементов окружающей среды, программное обеспечение для ее осмысления и механизмы управления для взаимодействия с ней. Датчики нужны для получения информации из окружающей среды. Реактивное поведение (подобно рефлексу растяжения у человека) не требует глубоких когнитивных способностей, но для того, чтобы робот мог выполнять важные задачи автономно, необходимо соответствующее программное обеспечение («бортовой интеллект»), кроме того, нужны механизмы (приводы) для того, чтобы робот мог воздействовать на окружающую среду [6]. Однако помимо этих трюизмов существует множество разногласий по поводу того, какое определение для термина «робот» является наиболее подходящим.

В данной статье мы придерживаемся общего определения, в рамках которого робот должен иметь датчики, «бортовой интеллект» и механизмы управления. Это определение не включает в себя виртуальных или программных роботов (так называемых «программных ботов»). Полностью дистанционно управляемые машины также не являются роботами, поскольку они не думают сами за себя. В нашем понимании робот должен думать и принимать решения самостоятельно. Таким образом, обычные наземные мины или калькуляторы не являются роботами. Робот думает в том смысле, что он может использовать информацию от датчиков и набор правил (запрограммированных или выученных), чтобы принимать некоторые решения автономно. Робот принимает решения самостоятельно, если он способен работать в определенной среде в течение некоторого времени без какого-либо внешнего контроля.

Мы склонны представлять себе роботов как артефакты, сделанные из гаек и болтов, как, например, беспилотные автомобили. Однако

роботы могут быть созданы и из органических материалов. Недавно ученые использовали живые клетки лягушки, чтобы получить так называемых *ксеноботов*, которые могут двигаться к цели и лечить себя после порезов. Эти новые живые машины не являются ни традиционными роботами, ни какими-либо известными видами животных. Это живые, но программируемые организмы [48].

Необходимым свойством роботов – живых и неживых – является интеллект. Учитывая, что роботы являются артефактами, интеллект, о котором идет речь, – это искусственный интеллект (ИИ). Принято различать два вида ИИ – сильный и слабый. Слабый ИИ ориентирован на достижение конкретных целей, предназначен для выполнения отдельных задач, и «умен» только при выполнении конкретной задачи, на которую он запрограммирован. В качестве примеров слабого ИИ можно привести Siri от Apple, роботы-дроны и беспилотные автомобили. Для того чтобы система демонстрировала сильный ИИ, она, напротив, должна действовать как мозг. Вместо того, чтобы классифицировать вещи в соответствии с заданными критериями, она должна использовать для обработки данных кластеризацию и сопоставление. Сильный ИИ должен не только решать конкретные задачи, но и имитировать человеческий интеллект и поведение, обладая способностью обучаться и применять интеллектуальные навыки для решения широкого круга задач. Принято считать, что сильный ИИ включает в себя сознание, автономность, разум, представление знаний, способность чувствовать, планировать, учиться и обобщать, а также способность общаться на естественном языке [31, pp. 1–3, 1020–40, 1044–6]. Поскольку сильный ИИ мог бы думать, понимать и действовать таким образом, что в любой конкретной ситуации он неотличим от человека, он бы с легкостью прошёл тест Тьюринга. Причина, по которой мы используем сослагательное наклонение, заключается в том, что сильного ИИ в настоящее время не существует.

Мы не будем утверждать возможность создания сильного ИИ, но предположим, что роботы с сильным ИИ («роботы» для краткости) возможны, и исследуем три следующих вопроса. Во-первых, может ли робот квалифицироваться как *моральный агент* в смысле разумного существа, способного сознательно выбирать свои собственные жизненные цели, а не служить простым средством для достижения целей других людей? Во-вторых, если роботы могут претендовать на роль моральных агентов или обладать некоторыми чертами, необходимыми для этого, обязаны ли мы работать над созданием таких роботов? Другими словами, обязаны ли мы помочь роботам достичь полной моральной субъектности? В-третьих, если мы не обязаны создавать

роботов, являющихся моральными личностями, существуют ли ограничения на то, что мы можем делать с роботами? Когда и в какой степени допустимо эксплуатировать роботов?

Значимость этих вопросов очевидна. Если с нашей помощью роботы смогут эволюционировать в моральных агентов, люди должны решить, стоит ли им позволить этому случиться. Если бы роботы стали личностями, то они, предположительно, имели бы те же права и обязанности, что и мы. А если у роботов были бы те же права, что и у нас, мы бы не смогли поручать им такую скучную, грязную и опасную работу, какую ни один человек не захотел бы выполнять. Помогая роботам стать личностями, мы расширяем их возможности, но в то же время снижаем их полезность, потому что эксплуатировать личность – это аморально. Чем больше возможностей у роботов, тем менее они полезны, и наоборот. Мы можем попытаться обойти эту дилемму, удерживая роботов на стадии развития ниже уровня личности. Мы могли бы запретить разработку роботов с сильным искусственным интеллектом. Однако такая стратегия порождает дополнительные моральные сомнения: допустимо ли с моральной точки зрения не помогать роботам или активно препятствовать их полноценному развитию?

Моральная философия Канта устанавливает критерии моральной агентности и личности, в соответствии с которыми весьма сомнительно, что роботы когда-либо будут квалифицироваться как моральные агенты. С точки зрения Канта, моральная личность в конечном итоге предполагает доступ к моральному закону и автономной воле. Чтобы робот стал моральным агентом, он не только должен быть способен свободно принимать решения и действовать так, чтобы это выглядело морально, но и его практический разум должен быть автономным. Робот должен был бы управлять собой. Более того, практическая философия Канта предлагает критерий для решения вопроса о том, допустимо ли препятствовать роботу стать моральной личностью. Моральная допустимость действия, по мнению Канта, зависит от того, можно ли волевым усилием сделать данную максиму универсальным законом, а также от того, уважаем ли мы разум и разумную природу.

В разделе 2 объясняется, почему принято считать, что роботы не способны к развитию моральной агентности. В разделе 3 излагается концепция моральной агентности Канта и утверждается, что, учитывая эту концепцию, крайне маловероятно, что роботы могут стать моральными агентами. В разделе 4 ставится вопрос о том, предпочтительно ли иметь роботов, являющихся моральными агентами, и почему. Предположим, что нам не выгодно, чтобы роботы становились моральными агентами: допустимо ли препятствовать развитию у них мо-

ральной агентности? В разделе 5 этот вопрос рассматривается с точки зрения Канта. Раздел 6 содержит некоторые заключительные замечания.

2. Роботы и моральная ответственность

Два стандартных возражения против приписывания роботу с сильным ИИ полноценной моральной агентности заключаются в том, что у него нет двух ключевых компонентов, необходимых для принятия моральных решений, – эмоциональной «внутренней» жизни и свободы воли. Давайте рассмотрим эти возражения по очереди.

Согласно аристотелевской традиции, у человека есть два разных вида систем принятия решений – инстинктивная (иррациональная) и когнитивная (разумная). Инстинктивная система принятия решений связана с эмоциями и является общей для высших млекопитающих. Она является частью когнитивной системы S1, которая работает быстро, интуитивно и – в основном – бессознательно. По большому счету, люди не считаются (в полной мере) морально ответственными за действия, совершенные на основе инстинктивной системы. Хотя большая часть человеческой деятельности обусловлена инстинктивной системой, мы также можем принимать решения на основе сознательного рассуждения. Когнитивная система, которая является частью мышления второго типа, позволяет нам представлять различные возможные варианты будущего и выбирать курс действий, основываясь на наших ценностях и вероятном исходе рассматриваемого действия. Моральная ответственность обычно относится к действиям, обусловленным когнитивной системой. Высшие млекопитающие и люди, не являющиеся агентами, такие как младенцы и люди с тяжелыми когнитивными нарушениями, не несут моральной ответственности за свои действия именно потому, что их решения действовать не являются результатом когнитивного размышления (cognitive deliberative system).

Некоторые отказывают роботам в моральной агентности, потому что у роботов нет эмоциональной внутренней жизни, необходимой для принятия решений. Это сомнительный шаг по двум связанным причинам. Во-первых, для того чтобы человек мог нести моральную ответственность за свои действия, ему необходима функционирующая система, способная принимать решения на основе размышления. Однако если роботы и не похожи на нас, то, предположительно потому, что у них отсутствует наша инстинктивная система, а не потому, что у них нет чего-то похожего на когнитивную систему принятия решений. Следовательно, это различие между людьми и роботами, похоже, не влияет на возможность моральной агентности. Во-вторых, люди с

дисфункциональной или отсутствующей эмоциональной внутренней жизнью (например, психопаты) все же могут нести моральную (и юридическую) ответственность за свои действия[2], в то время как те, кто имеет нормальные эмоциональные реакции, но не может рационально обдумывать свои действия (например, младенцы и умственно отсталые), не могут.

Другая стандартная причина отказа роботам в моральной агентности связана не с отсутствием у них эмоций, а с отсутствием у них свободы воли, то есть способности в аналогичных обстоятельствах принять иное решение. Идея заключается в том, что роботы не свободны, потому что их выбор является результатом детерминированного алгоритма.

Много написано о том, совместим ли каузальный детерминизм со свободой воли. Компатибилисты утверждают, что свободный выбор может быть обусловлен метафизической (но не физической) цепью событий. Канта часто понимают как своего рода компатибилиста, поскольку он проводит различие между законом причинности в феноменальном мире и законом свободы в мире ноуменальном. Интересно отметить, что для описания компатибилизма Кант использует жесткие выражения, такие как «жалкая уловка» и «мелочный педантизм» (KpV, AA 05:96). Однако в этих отрывках под компатибилизмом Кант понимает утверждение, что моральные поступки могут быть свободными в детерминированном мире постольку, поскольку они исходят изнутри нас, а не вынуждаются извне. Следовательно, похоже, Кант прав в том, что этого предложения недостаточно, чтобы объяснить, как могут сосуществовать причинность и свобода. В любом случае, он однозначно считает, что моральная деятельность требует свободы воли. Поэтому далее мы рассматриваем связанные вопросы: могут ли роботы быть моральными агентами и могут ли роботы обладать свободой воли.

Гарри Франкфурт (1969) разработал контрпримеры к принципу альтернативных возможностей, который гласит, что агент несет моральную ответственность за действие только в том случае, если он мог поступить иначе [10]. Рассмотрим следующий случай в стиле Франкфурта. Блэк хочет, чтобы Джонс убил Смита. Блэк создал устройство для манипулирования мозговыми процессами Джонса, чтобы Блэк мог определить, что Джонс решит убить Смита. Блэк вмешивается в процесс принятия решения Джонсом только тогда, когда Блэк недоволен

[2] Скэнлон [32, 287–290] и Талберт [41] утверждают, что агентов, полностью лишенных способности к моральному пониманию, все равно можно обвинять, если они обладают широкими разумными компетенциями.

тем, как Джонс собирается принять решение. Предположим, что Джонс сам решает убить Смита и убивает его. У Джонса нет другого выхода, кроме как сделать то, что хотел от него Блэк; независимо от того, сделает он это по собственной воле или из-за вмешательства Блэка, он убьет Смита. Многие философы считают, что Джонс несет ответственность за убийство Смита. Тем не менее, кажется, что Джонс не мог избежать убийства Смита. Когда Джонс сам убивает Смита, он несет моральную ответственность. Его ответственность не зависит от того, что Блэк затаился на заднем плане, готовый вмешаться, поскольку это вмешательство не вступает в игру. Джонс несет моральную ответственность за свой поступок, но он не мог поступить иначе. Для наших целей это означает, что даже если роботы не свободны, они все равно могут быть морально ответственными. А согласно Канту, моральная ответственность зависит от того, считаем ли мы агента ответственным за его действия.

3. Кант о роботах и моральной агентности

Известно, что моральная философия Канта является примером логоцентризма, поскольку она строится вокруг разума и разумных существ.

Например, моральные понятия и моральные законы необходимы и априорны (GMS, AA 04:408), и поэтому они, как утверждается, действительны для всех разумных существ. Осознание того, что моральные понятия должны быть действительны для всех разумных существ, Кант считает большим новшеством в теории морали. Предыдущие попытки создания морали потерпели неудачу, поскольку они основывались либо на эмпирических соображениях, либо на общем понятии воли. В отличие от этого, он основывает мораль на концепции чистой воли, которая является общей для всех разумных существ (GMS, AA 04:390, 4:407). Более того, моральные понятия вытекают из самого понятия разумного существа (GMS, AA 04:412)[3]. Что такое разумные существа и как моральные предписания опираются на их разумность? В самом простом смысле разумные существа – это существа, наделенные практическим разумом и способностью к желанию, которая опре-

[3] Согласно GMS, AA (04:412), разумные существа характеризуются способностью действовать в соответствии с репрезентациями законов. В «Основах» практический разум как способность выводить действия из законов приравнивается к воле. Однако в MS Кант отличает волю в строгом смысле слова, не имеющую определяющего основания, от воли в той мере, в какой она может определять выбор. Последняя также отождествляется с практическим разумом (06:213).

деляется самим практическим разумом. По мнению Канта, люди представляют собой особый вид разумных существ в той степени, в которой их воля зависит как от практического разума, так и от чувственных желаний и склонностей. Следовательно, для человека предписания практического разума могут вступать в противоречие с чувственными склонностями. Таким образом, моральный закон обязывает нас и становится императивным. Другие разумные существа, такие как Бог и ангелы, являются божественными или святыми волями, которые уже находятся в полном соответствии с моральным законом (GMS, AA 04:414). Разницу между человеческой и святой волей можно проиллюстрировать следующим образом. Чтобы поступать нравственно, мы должны действовать из долга, а не просто в соответствии с долгом. Что касается людей, то мы никогда не можем быть уверены в том, что наши волевые действия исходят из долга, поскольку на нас могли повлиять наши недобрые побуждения, такие как самолюбие (GMS, AA 04:406). В отличие от этого, понятия долга не применимы к воле Бога, которая обязательно и без исключений согласуется с моральным законом (GMS, AA 04:414).

Теория морали Канта сосредоточена вокруг разумных существ, но его фактическое изложение моральных обязанностей и прав в «Метафизике нравов» сосредоточено на человеческих существах. В частности, Кант, похоже, одобряет различие между прямыми и косвенными обязанностями. Первые применимы только к отношениям между людьми, либо как обязанности по отношению к себе, либо как обязанности по отношению к другим людям. Что отличает отношения между людьми от отношений к другим существам, так это то, что они содержат как прямые права, так и прямые обязанности (MS, AA 06:442). Более того, некоторые обязанности, такие как обязанность избегать преднамеренного разрушения прекрасного в природе и обязанность избегать ненужного насилия по отношению к животным (MS, AA 06:443), кажутся обязанностями по отношению к неразумным и нечеловеческим существам. Однако, по мнению Канта, эти обязанности являются косвенными: это обязанности по отношению к нам самим, которые касаются неразумных и нечеловеческих существ. Таким образом, кажется, что в конечном счете у нас есть только прямые обязанности уважать человечество в себе и в других людях. Эта идея основана на том, что только люди как разумные существа принимают моральное законодательство и при этом являются объектами возможного опыта.

Что касается неразумных существ, таких как нечеловеческие животные и природа, то мы обязаны уважать их ради нас самих, а не ради

них[4]. Причина, по которой мы должны избегать разрушения прекрасного в природе, заключается в том, что это ослабит нашу способность любить что-то независимо от наших собственных целей и интересов (MS, AA 06:443). А причина, по которой мы не должны жестоко обращаться с неразумными животными, заключается в том, что мы можем стать равнодушнее к страданиям других людей (MS, AA 06:443). Что касается разумных существ, которые находятся в полном согласии с моральным законом, то у нас нет обязанностей по отношению к ним, поскольку они не являются объектами нашего возможного опыта (MS, AA 06:242, 06:444). Бог – верховный глава морального законодательства, но поскольку Бог не является объектом опыта, он не является субъектом ни обязанностей, ни прав (GMS, AA 04:433). Однако мы все равно должны верить в Бога как в практический долг перед собой.

Давайте на минуту предположим, что роботы могут быть разумными существами, то есть существами, наделенными практическим разумом и волей, которая может быть определена практическим разумом. Могут ли роботы быть примером святой воли, так что их воления автоматически согласуются с предписаниями практического разума? Это кажется сомнительным по двум причинам. Во-первых, считается, что такие разумные существа, как Бог и ангелы, не являются объектами нашего возможного опыта. Роботы, однако, воспринимаются нами. Тем не менее, это может быть случайным фактом, что боги и ангелы подпадают как под категории святой или божественной воли, так и под категории существ, которые не могут быть для нас объектами возможного опыта. Во-вторых, и это самое важное: несмотря на то, что роботы не обладают когнитивной системой S1, они демонстрируют поведение, подобное инстинкту. Вполне возможно, и это является насущной проблемой в нынешних дискуссиях о самоуправляемых автомобилях, что роботы могут демонстрировать конфликт между предписаниями практического разума, такими как защита жизни людей, и другими правилами более низкого уровня, такими как оптимизация комфорта или тому подобное. Далее мы утверждаем, что если роботы могут быть моральными агентами, то они должны быть схожи с человеческими существами. Из этого следует, что, если они могут быть моральными агентами, они должны быть восприимчивы к понятиям

[4] Взятые вместе утверждение о том, что у нас есть только прямые обязанности перед людьми, и утверждение о том, что мы должны уважать природу и нечеловеческих животных ради нас самих, составляют так называемый «непрямой взгляд». Корсгаард замечает, что эти два элемента взглядов Канта не обязательно должны быть объединены в пару. Например, мы можем иметь обязанности только по отношению к другим людям, но ценить природу и животных ради них самих [19, 116].

долга. Мы будем утверждать, что роботы не могут считаться разумными существами, поскольку они не обладают автономией.

Во введении к «Метафизике нравов» Кант предлагает предварительное определение морального агента или личности как того, кому можно вменить действия (MS, AA 06:223), и кто может рассматриваться как автор (*causa libera*) действия (06:227)[5]. Это, по-видимому, отражает общепринятое представление о том, что действие человека может быть морально неправильным только в том случае, если он контролирует это действие (см. [30]). Моральная личность требует способности действовать в соответствии с общим законом, или волей (MS, AA 06:224; GMS, AA 04:412). Для разумных существ это сводится к двум особенностям: во-первых, «свобода разумного существа [состоит в том, чтобы находиться] под действием моральных законов» (MS, AA 06:223) и, во-вторых, человек подчиняется только тем законам, которые он сам себе устанавливает (MS, AA 06:223)[6]. Давайте рассмотрим эти две особенности по очереди.

И люди, и животные обладают способностью создавать объекты своих представлений и желаний (MS, AA 06:211). Однако выбор животных (*arbitrium brutum*) полностью определяется чувственными побуждениями и импульсами (MS, AA 06:213; 27:344), в то время как выбор человека может зависеть от склонностей, не будучи полностью ими определенным. Выбор человека свободен в той мере, в какой он определяется чистым разумом (MS, AA 06:213). Известно, что в «Основах метафизики нравственности» и в Критике чистого разума Кант называет свободной саму волю, а в «Метафизике нравов» он вводит понятие «свободный выбор» (Willkür). Согласно «Метафизике

[5] Этому предварительному определению моральной агентности не следует придавать чрезмерного значения, поскольку неясно, как оно будет применяться к святой воле и Богу. Тем не менее, оно полезно в качестве предварительного взгляда на концепцию Канта о моральной агентности для человеческих существ.

[6] Личность – это одна из трех предиспозиций человеческой природы. Кант различает в самих людях нашу животную природу, нашу человеческую природу и нашу личность. Животная природа включает в себя наши естественные желания и чувственные побуждения. Человечность – это способность ставить перед собой произвольные цели. Личность – это разумная способность устанавливать законы и подчиняться им (06:26, 7:321-324). Как справедливо замечает Вуд [47, p. 189], человечность состоит из технической способности устанавливать для себя произвольные цели и практической тенденции гармонизировать наши цели в единое целое, называемое счастьем. Этот элемент становится решающим, когда мы рассматриваем содержание обязанностей по отношению к себе и к другим людям, представленных в «Метафизике нравов». Например, мы обязаны заботиться о счастье других людей. Понимая этот долг, мы должны помнить, что люди – разумные существа, и ценить их разум, но также и то, что их человечность порождает потребность в счастье. Возможно, другие разумные сущности, подобные святой воле, не имеют таких потребностей.

нравов», воля, строго говоря, не является ни свободной, ни ограниченной, она «не имеет определяющего основания» (MC, AA 06:213). Независимо от несколько различных концепций свободы воли в моральном опусе Канта, воля характеризуется как «способность действовать в соответствии с представлением законов, т. е. в соответствии с принципами» (GMS, AA 04:412, 04:427; см. MS, AA 06:213). Проще говоря, воля – это каузальная способность, присущая разумным существам, направлять свой выбор целей с помощью принципов или суждений о том, что есть благо [14, xvi].

На первый взгляд может показаться, что определение воли как способности действовать в соответствии с представлением законов легко применимо к роботам. В конце концов, прямой способ описать поведение роботов – это заявить, что они действуют на основе правил, выученных или запрограммированных. Вряд ли роботы способны действовать на основе своих *репрезентаций* законов или правил, а не просто слепо следовать этим законам и правилам. Точно так же нечеловеческие животные действуют на основе правил, задаваемых им инстинктами и желаниями, но у них отсутствует практический разум, который является источником законов. Далее мы объясним, что значит действовать с практическим разумом или без него в контексте концепции морали Канта.

Чтобы понять концепцию свободного выбора Канта, полезно кратко изложить его понятие свободы. С одной стороны, Кант формулирует позитивную концепцию свободы в терминах «способности чистого разума быть самому себе практичным» (MS, AA 06:214) или в терминах «внутреннего законотворчества разума» (MS, AA 06:227). С другой стороны, негативная концепция свободы как способности действовать без какой-либо внешней причины требует, чтобы мы были *трансцендентально* свободны. Трансцендентальная свобода – это способность действовать, не будучи детерминированной внешними причинами и естественными законами, такими как причинность (KpV, AA 05:29). Это, по-видимому, является условием практической свободы, понимаемой как автономия (см. [8], ср. KpV, AA 05:29). Известно, что в Основах метафизики нравов III Кант пытается обосновать мораль с помощью понятия свободы. Позже, в Критике практического разума, Кант возвращается к вопросу о взаимоотношениях между моралью и свободой[7]. Там он утверждает, что свобода – это ratio essendi (лат.

[7] Вопрос о том, отличается ли и если отличается, то насколько, взгляд на свободу и мораль, изложенный в «Основах», от взгляда, выраженного в «Критике практического разума», является предметом споров. Шёнекер (1999) отстаивает мнение, что Кант действительно изменил свое мнение между двумя работами [33]. В отличие от него, Элли-

основание бытия – прим. переводчика) морального закона, но мы узнаем о своей свободе только благодаря моральному закону (KpV, AA 05:6 n). Он утверждает, что наш моральный опыт, ограниченный моральным законом, является «фактом разума», поскольку он не может быть выведен из других данных нашего разума, таких как сознание нашей свободы (KpV, AA 05:31)[8].

Независимо от того, обеспечивает ли моральный закон нашу практическую свободу или наоборот, свобода является важнейшим компонентом концепции морали Канта. В Критике практического разума Кант доходит до утверждения, что без свободы, понимаемой как автономия воли, мы были бы подобны автоматам или роботам (KpV, AA 05:101). Если бы мы не были действительно свободны в этом смысле, то и наше сознание своей спонтанности было бы иллюзией. Поскольку даже если бы наш когнитивный механизм казался интерсубъективным и самопричинным, в конечном итоге всеми нашими действиями руководила бы «чужая рука». Это замечание подчеркивает приоритет того, что мы на самом деле автономны, над нашим осознанием этой автономии.

Кант объясняет нашу практическую свободу в терминах автономии нашего практического разума (KpV, AA 05:31; MS, AA 06:227)[9]. Практический разум – это то же самое, что и широкая концепция воли. Сказать, что практический разум автономен, означает, что люди связаны только теми законами, которые они сами себе устанавливают (GMS, AA 04:432). Другими словами, каждое разумное существо должно рассматривать себя как дающего универсальные законы через максимы своей воли (GMS, AA 04:432). Важность понятия автономии в практической философии Канта трудно переоценить. Автономия объясняет не только свободу, но и внутреннее достоинство человека. Все согласны с тем, что автономия играет центральную роль в прак-

сон (2011) утверждает, что в концепции свободы Канта нет «радикальных изменений» [2, 297, n. 41]. Недавно Пулс (2016) также отстаивал мнение о том, что в G III и KpV наблюдается существенное сходство [25].

[8] Существует много споров о том, как следует понимать разговор о «факте разума», см. в частности, [21, 16, 35, 43].

[9] Понятие автономии, используемое Кантом, существенно отличается от понятия автономии, используемого в ИИ и представленного в разделе 1. В ИИ робот считается автономным, если он не полностью зависит от своих предыдущих знаний, но способен интегрировать информацию, полученную из собственных представлений. Для достижения этой цели важно, чтобы роботы, основанные на знаниях, могли обучаться на основе своих собственных представлений [31, p. 39, p. 236]. Кроме того, представленное в разделе 5 понятие «автономии предпочтений» также отличается от автономии как в кантовском смысле, так и в смысле ИИ. Автономия предпочтений заключается в способности человека и нечеловеческих животных иметь предпочтения, основанные на их потребностях и импульсах, и инициировать действия (см. [47, 200; 28, 84–6]).

тической философии Канта, однако существуют разногласия по поводу того, что именно означает для нашего разума обладать моральным законом[10].

В данной работе мы опираемся на концепцию автономии, отстаиваемую Кляйнгельдом и Виллашеком [17]. Они утверждают, что наш разум автономен не в том смысле, что он сам дает себе моральный закон, на котором основаны все конкретные моральные законы, а в том смысле, что мы являемся источником обязательной силы морального закона. Мы являемся законодателями для самих себя в той мере, в какой мы делаем закон действительным для нас. Теперь посмотрим, могут ли роботы считаться моральными агентами при таком понимании автономии. Заметим, что если субъект, будь то человек или робот, не может быть автономным в смысле обеспечения нормативной силы для морального закона, то он не может считаться автономным в сильном смысле вклада в создание морального закона. Нет никакого релевантного смысла, в соответствии с которым роботы могли бы быть поняты как дающие себе моральный закон, потому что независимо от того, заложен ли он в них жестко или создан путем эмпирического обучения, он все равно в конечном счете исходит от «чужой руки» (KpV, AA 05:101). Тем не менее, даже в соответствии с концепцией автономии как источника обязательной силы закона, роботы не могут считаться автономными, поскольку они не способны сделать моральный закон действительным для себя. Следовательно, невозможно рассматривать роботов как разумных агентов.

Если, как мы предполагаем, роботы должны считаться разумными существами, подобными людям, в том смысле, что их воля может противоречить предписаниям практических соображений, роботы должны будут удовлетворять дополнительным требованиям, чтобы считаться разумными агентами. Упомянутые до сих пор дополнения к морали относятся к людям как к разумным существам. Кант, однако, также представляет нам список моральных чувств, которые основаны на понятии долга (MS, AA 06:399). В «Метафизике нравов» чувство уважения, которое уже обсуждалось в «Основах» и в «Критике практического разума» (в разделе о мотивах чистого практического разума), расширяется до четырех чувств: *нравственное чувство, совесть, любовь к ближнему и уважение к себе (самоуважение*). В целом

[10] Рит [Reath] (1994) предложил анализ автономии по аналогии с политикой: государство свободно в той мере, в какой законы, связывающие граждан, являются результатом их действий, например, путем голосования [29]. Конструктивистское прочтение см. [26], [18], [24], [9]. О реалистической традиции см. [3], [47], [20], [12]. См. также [39] и [4].

нравственные чувства характеризуются тем, что они не предшествуют ни желанию, ни представлению закона. Напротив, удовольствия, проистекающие из склонностей, могут предшествовать желанию и побуждению. Они характеризуются в терминах «эстетической восприимчивости к понятиям долга (уважения)» (MS, AA 06:399). Эти чувства далее характеризуются как «естественные душевные задатки, на которые оказывают воздействие понятия долга» (MS, AA 06:399). Кант решительно утверждает, что не может быть обязанности иметь их, но только обязанность культивировать их. Действительно, благодаря этим чувствам мы осознаем обязательства, содержащиеся в моральном законе[11]. Другими словами, можно сказать, что эти чувства представляют собой способ воздействия морали на нас, то есть на разумных существ, которые, как мы, склонны отступать от предписаний морального закона.

Давайте вкратце рассмотрим, как эти четыре чувства могут подтолкнуть нас в направлении нравственного закона. *Нравственное чувство* – это чувство удовольствия или неудовольствия, которое зависит от нашего сознания того, что наши действия согласны или не согласны с нравственным законом (MS, AA 06:399). *Совесть* характеризуется через метафору «внутреннего суда» (MS, AA 06:438). Практический разум судит и осуждает наши поступки, предоставляя объективные правила для нашего поведения. Но именно способность суждения выносит конкретные суждения, которые имеют отношение к совести. Это субъективные суждения, касающиеся не того, что объективно является нашим долгом, а скорее того, была ли та или иная максима, ведущая к действию, представлена практическому разуму. Таким образом, в рамках судебной метафоры внутреннего суда Кант приписывает роль обвинителя самому практическому разуму. Совесть, напротив, – это наша внимательность к голосу этого внутреннего судьи (MS, AA 06:401) или сознание этого внутреннего суда (MS, AA 06: 438). Иными словами, совесть – это естественная предрасположенность со стороны чувства, которая позволяет нам прислушиваться к вердиктам этого внутреннего суда, а значит, позволяет нам судить о том, соответствуют ли наши поступки нашим обязанностям или нет, вызывая угрызения совести или радость (MS, AA 06:440). Более того, благодаря этому процессу мы можем приписывать себе действия. (MS, AA 06:838-9). *Любовь к ближнему* – третье нравственное чувство, представленное в «Метафизике нравов». Оно связано с обязанностью быть благо-

[11] Дитер Шёнекер предполагает, что эти чувства играют важную роль в нашем знании о моральном законе [36].

желательным по отношению к другим в смысле прямой помощи им в их материальном благополучии и косвенной помощи им в их моральном благополучии (06:393-94). Кант уточняет, что благожелательность не должна основываться на практической любви к ближним. Если бы мы действовали таким образом, то поступали бы просто по склонности и могли бы перестать помогать другим, как только перестали бы к этому склоняться. Любовь характеризуется в терминах *amor complacientiae (*лат. любовь к самоуспокоенности, самодовольству – прим. переводчика*)*, которая представляется как непосредственное наслаждение, возникающее в результате нашего стремления к нравственному совершенству[12]. Естественно предположить, что такая любовь требует некоторого чувства принадлежности к обществу (см. также [5]; [11]; [45]). Наконец, *уважение к себе (или самоуважение)* – это чувство по отношению к себе, которое помогает нам уважать человечность в себе[13]. Как и уважение в «Основах», разумные критерии нравственности – это всего лишь воздействие морального закона на нас.

На первый взгляд может показаться, что разумные критерии Канта эквивалентны требованию к роботам иметь внутреннюю эмоциональную жизнь, о котором говорилось в разделе 2. Однако обратите внимание, что важнейшим элементом изложения Канта является то, что разумные критерии – это способ, которым моральный закон и понятие долга связывают нас. Следовательно, если роботы могут быть не согласны с практическим разумом, то и у них должен быть какой-то способ, чтобы моральный закон связывал их и оказывал свое воздействие на них как на субъектов закона[14].

Чтобы выяснить, могут ли роботы считаться кантовскими моральными агентами, мы представили важнейшие черты теории морали Канта. Кроме того, мы рассмотрели два возможных варианта того, как роботы могут квалифицироваться в качестве разумных агентов. Либо они подобны святой воли, и в этом случае им не хватает автономии и свободы воли; либо это подобие человеческой воли, и в этом случае

---

[12] Любовь к ближнему часто интерпретируется в терминах благожелательности. Против этой общей тенденции и для более полного описания любви как *amor complacientiae* см. Шенекер [34].

[13] В (GMS, AA 04:401 n), уважение характеризуется следующим образом: «Когда я познаю нечто непосредственно как закон для себя, я познаю с уважением, которое означает лишь сознание того, что моя воля подчинена закону без посредства других влияний на мои чувства. Непосредственное определение воли законом и сознание этого определения называется уважением, так что уважение рассматривается как действие закона на субъект, а не как причина, этого закона».

[14] Аналогичное замечание см. [37].

им не хватает автономии, свободы воли и разумных критериев морали. В любом случае роботы не могут считаться моральными агентами в кантовском смысле. Мы не представляем, что нужно сделать, чтобы создать робота, отвечающего условию Канта о моральной способности.

4. Целесообразность моральных роботов

В предыдущем разделе мы увидели, что, учитывая кантовскую точку зрения, весьма сомнительно, что роботы когда-либо станут моральными агентами. Однако на данный момент давайте оставим это за скобками, и вместо этого зададимся вопросом, хотим ли мы создать роботов с моральными способностями. Целесообразно ли для нас иметь моральных роботов?

Есть как минимум три соображения в пользу моральных роботов [7]. Во-первых, роботы с моральными способностями станут более социально полезными и интегрированными в нашу жизнь, чем роботы без моральных способностей. Например, роботы с моральными способностями могли бы использоваться в качестве медсестер для пациентов с высокоинфекционными заболеваниями и в космических исследованиях. Во-вторых, в некоторых областях (например, медицина, военное дело, автономные транспортные средства) было бы безответственно внедрять роботов, если бы они не обладали моральными способностями в той или иной форме. В-третьих, поскольку роботы менее неоднозначны в своих моральных суждениях и менее непостоянны и изменчивы в своих моральных чувствах, они могут помочь нам в принятии моральных решений. Например, при принятии решений о распределительном и уголовном правосудии мы, как правило, увязаем во множестве моральных переменных и интересов, и нам трудно эффективно сбалансировать эти интересы при принятии решений. Благодаря своей простоте и более строгому следованию правилам роботы могут помочь нам пробиться сквозь моральный шум. Принятие моральных решений с помощью роботов может быть более быстрым, последовательным, справедливым и, в конечном счете, более безопасным (например, в случае с автомобилями без водителя). Еще одно место, где роботы могут оказаться полезными, – это суд присяжных. Обычно трудно найти беспристрастных присяжных для громких судебных процессов. Именно в этом случае беспристрастный, но морально компетентный присяжный в виде робота был бы очень полезен.

С другой стороны, есть и веские причины не желать, чтобы роботы обретали моральную агентность. Для начала, позволив роботам обрести моральную способность, мы лишим себя возможности эксплуатировать их. Как уже говорилось во введении, роботы в настоящее

время используются для выполнения настолько скучной, грязной и опасной работы, что ни один человек не захочет ее выполнять. Если бы роботы стали моральными агентами, они имели бы те же права, что и мы, и, следовательно, нам бы пришлось использовать их по-другому.

Еще одна причина, по которой мы не хотим, чтобы роботы развивали моральные способности, заключается в том, что сценарии, в которых мы больше всего нуждаемся в компетентных этических агентах, включают в себя моральную двусмысленность и необходимость понимания контекста. Неоднозначные ситуации – это ситуации, в которых требуется суждение и нет единственно правильного ответа. Яркий пример неоднозначной ситуации – поле боя. Вероятно, мы не хотим наделять моральных роботов способностью принимать самостоятельные решения об убийстве людей. Но даже в менее серьезных ситуациях моральный робот может причинить вред, не до конца понимая всю сложность ситуации. Шарки приводит пример робота-бармена, который подает взрослым клиентам столько алкоголя, сколько они хотят. Она пишет: «Но как робот может принимать правильные решения о том, когда похвалить ребенка или ограничить его деятельность, без морального понимания? Точно так же, как робот может обеспечить хороший уход за пожилым человеком, не понимая его потребностей и последствий своих действий? Даже робот, обслуживающий бар, может быть поставлен в ситуацию, в которой ему придется принимать решения о том, кого следует или не следует обслуживать, и что является приемлемым поведением, а что – нет» [40, p. 293].

5. Кант о допустимости воспрепятствовать превращению роботов в моральных агентов

Если бы роботы стали моральными агентами, у нас были бы четкие моральные обязанности по отношению к роботам, и они имели бы те же права, что и мы, или, по крайней мере, некоторые минимальные права, такие как право на «жизнь»[15]. Если бы нам по-прежнему разрешалось использовать их труд, мы должны были бы относиться к ним как к целям, а не просто как к средствам. В разделе 3 мы увидели, что понятие моральной личности Канта неприменимо к роботам. Это ценный результат, ведь если бы роботы обладали моральной личностью, у них были бы права и обязанности по отношению к самим себе. Следовательно, мы не смогли бы заставить их жертвовать собой ради защиты человеческих жизней в критических ситуациях. А ведь это одно из важнейших преимуществ при взаимодействии человека и робота.

[15] О техническом понятии жизни у Канта см. (6:211).

Далее мы рассмотрим два актуальных вопроса о роботах и их развитии. Во-первых, мы должны понять, вправе ли мы с моральной точки зрения препятствовать роботам в их становлении в качестве моральных личностей. Этот вопрос ведет нас на неизведанную территорию, поскольку, насколько нам известно, Кант не затрагивал в явном виде вопрос о том, можем ли мы помешать существу, которое может обладать практическим разумом или некоторыми предпосылками разума, достичь полной моральной индивидуальности. Более того, этот вопрос является полностью гипотетическим. В разделе 3 мы установили, что, согласно концепции морали Канта, роботы не могут быть моральными агентами. Теперь мы рассмотрим, возможно ли, чтобы они стали моральными личностями или, по крайней мере, разделяли некоторые аспекты нашей разумной природы.

Prima facie вопрос о том, допустимо ли препятствовать роботам в достижении моральной личности, аналогичен вопросу о том, допустимо ли с точки зрения морали препятствовать ребенку или человеку с когнитивными нарушениями в становлении в качестве моральных личностей. Характеристика разумных существ, данная в разделе 3, подразумевает, что наделенность способностью к практическому разуму является *conditio sine qua non* – неотъемлемым условием для моральной личности. Однако мы также видели, что, по мнению Канта, для того чтобы быть моральным агентом, требуется нечто большее, чем просто потенциал быть разумными. Некоторые исследователи Канта даже утверждают, что, учитывая критерии Канта, дети и люди с когнитивными недостатками и нарушениями не могут считаться полноценными моральными личностями ([28]; [46]; [22]; [23]; см. также [13]). Эта точка зрения перекликается с нашей интуицией, согласно которой таких агентов нельзя винить за их аморальные поступки. Однако Кант, по-видимому, привержен идее, что мы должны позволить детям и людям с нарушениями познавательной деятельности достичь полной разумности и, следовательно, моральной способности. Или, по крайней мере, Кант утверждает, что на нас лежит широкая, несовершенная обязанность не вмешиваться в нравственное развитие других человеческих существ. Например, эта широкая, несовершенная обязанность может выражаться в обеспечении того, чтобы их материальные условия не нанесли ущерб их моральному статусу (MS, AA 06:394)[16]. Более того, в приложении к «Метафизике нравов», где Кант

[16] Формулировка Канта в MS, AA 06:394 совершенно не соответствует рассматриваемому случаю, поскольку он пишет: «Я обязан воздерживаться от того, что, учитывая природу человека, может склонить его к поступку, за который его совесть может впоследствии мучить его, воздерживаться от того, что называется скандалом». В случае с

излагает свои взгляды на преподавание этики, ясно говорится, что добродетели «можно и нужно учить» (MS, AA 06:477). Вуд [46, p. 198] даже предполагает, что, учитывая хрупкость детей и слабоумных, уважение к разумной природе может даже требовать от нас защищать их и отдавать приоритет их развитию.

При строгом толковании понятия «разумное существо», согласно которому разумными являются только существа, обладающие полным и действительным практическим разумом, очевидно, что роботы не могут стать моральными личностями, и, следовательно, вопрос о том, должны ли мы допускать их моральное развитие, не возникает. В последние годы исследователи Канта утверждают, что моральная система Канта требует расширения, чтобы включить и удовлетворительно объяснить наши обязанности не только по отношению к детям и когнитивно неполноценным людям, но и по отношению к нечеловеческим животным и природе [46]; [19]. Эти интерпретации пересматривают понятие разумности Канта, показывая, что для последовательного понимания системы Канта требуется некоторая форма разума, которая должна быть общей для всех этих категорий.

Вуд (1998) развивает идею «потенциального разума» или «инфраструктуры рациональности» [46]. Он утверждает, что Кант, по-видимому, придерживается принципа персонификации: согласно второй формулировке морального закона, формуле человечности (FH), мы должны уважать разумную природу, которая первична в нас самих или в других[17]. Вуд предполагает, что последовательное изложение этики

человеческими существами, которые еще не полностью развили свои разумные и моральные способности, такими как дети, можно предположить, что дети еще не могут в полной мере ощутить чувственные последствия нашего сознания обязанностей и, следовательно, испытывать боль совести. Тем не менее, это лишь кажущаяся проблема: во-первых, дети, предположительно, уже чувствуют некоторую форму раскаяния за свои действия; во-вторых, они могут или потенциально могут достичь полной моральной устойчивости и, следовательно, чувствовать раскаяние за свои действия. Второй пункт применим даже к людям, которые, возможно, никогда не достигнут полного морального состояния и полной разумности, потому что их потенциальный разум, если он полностью актуализируется, заставит их чувствовать угрызения совести, следовательно, они потенциально подвержены угрызениям совести.

[17] Кант, как известно, предлагает три основные формулировки морального закона в «Основах», причем две из них имеют варианты (см. также [38, pp. 122-172]). Первая формулировка, называемая Формулой всеобщего закона (FUL), звучит следующим образом: «Действуй только в соответствии с той максимой, посредством которой ты можешь в то же время желать, чтобы она стала всеобщим законом» (GMS, AA 04:421). Вторая формулировка, называемая Формулой человечности (FH), гласит: Поступай так, чтобы ты всегда относился к человечеству и в своем лице, и в лице всякого другого также как к цели и никогда не относился бы к нему только как к средству» (GMS, AA 04:429). Третья формулировка, названная Формулой автономии, предписывает «идею воли каждого разумного существа как воли, дающей универсальный закон» (GMS, AA 04:431). Эти три формулировки с их вариантами теоретически эквивалентны: все они выражают мо-

Канта требует отказа от принципа персонификации и понимания FH в терминах уважения разумной природы как таковой. Более того, он утверждает, что обязательства Канта по справедливому обращению с животными предполагают аналогию между разумом, с которой мы сталкиваемся в человеческих существах, и «инфраструктурой рациональности», с которой мы сталкиваемся в животных. В его представлении моральные соображения касаются не только тех, кто полностью разумен, но и всех тех, кто потенциально обладает разумной природой. Понятие потенциального разума включает в себя людей, которые практически обладают разумом или обладали им в прошлом, а также тех, кто имеет части разумной природы или ее необходимые условия [46, 200–1]. Таким образом, дети, когнитивно неполноценные люди и животные достойны морального рассмотрения. Вуд [46, 200] и Реган [28, 84–6] утверждают, что животные обладают «автономией предпочтений», поскольку у них есть предпочтения и способность инициировать действия. Вуд представляет автономию предпочтений как необходимое предварительное условие для моральной автономии и как фундаментальную составляющую нашей разумной природы. Другими словами, с этой точки зрения, животные обладают необходимой «инфраструктурой рациональности». Когда мы проявляем излишнюю жестокость по отношению к животным, мы проявляем неуважение к той части разумной природы, которую мы разделяем с животными. Эта часть нашей разумной природы, по-видимому, совпадает с нашей животной частью, способностью действовать на основе естественных импульсов и желаний. Поэтому как разумно несовершенные люди, так и нечеловеческие животные имеют права.

Согласно предположению Корсгаард [19], взгляды Канта об обязанностях по отношению к животным подразумевают, что люди и нечеловеческие животные аналогичны в соответствующем аспекте. Корсгаард утверждает, что мы можем различить два смысла выражения «самоцели». С одной стороны, человек является самоцелью, поскольку он может придать силу закона своим требованиям и практическим суждениям, участвуя в моральном законодательстве. С другой стороны, человеческие существа являются самоцелью, понимаемой как источник легитимных нормативных требований – требований, которые должны быть признаны всеми разумными агентами. Согласно последней концепции самоцели, животные являются самоцелью в той же мере, что и мы. Согласно взглядам Корсгаард, животные способны

ральный закон в разных аспектах. Тем не менее, они не эквивалентны практически. Например, в «Метафизике нравов», которая представляет собой содержание этики Канта, а не просто ее основание, FH имеет явный приоритет (см. [14, xxxi-ii]).

распознавать явления как хорошие и плохие. Мы разделяем эту черту животности, и от нас морально требуется уважать этот статус как у людей (MS, AA 06:420 и далее, 06:452), так и у нечеловеческих животных.

Взгляды Вуда и Корсгаард гораздо более широки, чем стандартный взгляд. В своих рассуждениях они приходят к тому, что животные являются подходящими объектами моральных размышлений, даже если обязанности, касающиеся благополучия животных, являются обязанностью по отношению к нам самим. Наши чувства благодарности по отношению к животным, которые служили нам, уместны только потому, что животные похожи на нас, т. е. они обладают чем-то, что мы должны уважать в других людях. Тем не менее, даже согласно этим либеральным интерпретациям концепции Канта о получателях моральных прав, мы не обязаны позволять моральное и разумное развитие роботов. Во-первых, отметим, что даже согласно Корсгаард мы не должны иметь никаких обязанностей по моральному развитию животных. В конце концов, мы разделяем с ними чувство самоцели, согласно которому самоцель является источником нормативных требований. Животные не участвуют в моральном законодательстве и никогда не смогут в нем участвовать. Во-вторых, разумно несовершенные люди явно обладают практическим разумом потенциально или виртуально, а также необходимыми предпосылками практического разума. По мнению Вуда, животные демонстрируют «инфраструктуру рациональности» и этим заслуживают наше уважение. Роботы, однако, лишены всех конститутивных элементов практического разума: они не наделены практическим разумом ни потенциально, ни виртуально, а также не демонстрируют автономии предпочтений, свойственных животным. Роботы, кажется, не дотягивают даже до нашей животной природы. Роботы не испытывают боли, желаний и естественных побуждений так, как это делают люди и животные. Автономия предпочтений может быть условием моральной личности (для людей) только в том случае, если разумные побуждения животных не навязаны им или не запрограммированы в них. Роботы не являются теми вещами, которые самостоятельно способны определять хорошее и плохое. Даже если бы роботы были способны демонстрировать боль и поведение, подобное желанию, это все равно было бы прямым или косвенным результатом нашего собственного программирования. Это не было бы частью разума, которым роботы обладают независимо от нас и который мы разделяем с ними. Таким образом, мы не обязаны роботам моральными соображениями, которые, по мнению Вуда и Корсгаард, мы должны животным.

Здесь целесообразно использовать тест Канта на универсальность. Действие морально допустимо, если мы можем пожелать, чтобы оно стало всеобщим законом (GMS, AA 04:402, 04:421-3). Мы не должны убивать людей, потому что мы не можем даже представить себе мир, в котором убийство станет всеобщим законом: в нем не останется ни одного человека. Мы должны помогать другим, потому что мы не можем представить себе мир, в котором никто не помогает другим. Совершенные, узкие обязанности вытекают из невозможности представить себе, что какая-либо максима не станет всеобщим законом. Несовершенные, широкие обязанности проистекают из неспособности волевым усилием превратить данную максиму в универсальный закон (GMS, AA 04:421-5, см. также MS, AA 06:390-4). Мы не можем помешать человеку развить свою нравственную личность, потому что это уничтожило бы человечество и его способность к нравственным поступкам. Препятствовать человеку в достижении нравственной личности означало бы не уважать разумную природу этого человека, и это было бы равносильно неуважению к нашему собственному величайшему нравственному совершенству, то есть способности действовать по долгу (MS, AA 06:392). И все же мы можем представить себе и пожелать мир, в котором роботы будут лишены возможности становления моральными личностями. Роботы, как мы утверждали, не могут стать моральными существами, и вполне законно, что мы не хотим, чтобы это произошло.

Когда Кант дает нам полное изложение наших обязанностей в «Метафизике нравов», он часто опирается на формулу человечности, а не на формулировку всеобщего закона или закона природы. Поэтому было бы полезно перестроить нашу дискуссию на основе формулы человечности: «Поступай так, чтобы ты всегда относился к человечеству и в своем лице, и в лице всякого другого также как к цели и никогда не относился бы к нему только как к средству» (GMS, AA 04:429). Причина, по которой морально недопустимо мешать человеку достичь нравственной индивидуальности, заключается в том, что такая максима нарушала бы права и ценности человечества. Мы бы унизили нашу собственную разумную природу. Аналогичным образом, предотвращение достижения роботами моральной индивидуальности морально допустимо, если наше намерение согласуется с уважением к человеческой жизни и ее разумной природе.

Исходя из всего, что мы до сих пор доказывали, может показаться, что кантовская этика позволяет нам делать с роботами все, что заблагорассудится. Но это не так. Этическая система Канта может предложить рекомендации по морально допустимому использованию роботов. Создавая роботов, мы должны не убивать других людей, уважать

человечность в себе и в других и т. д. В сущности, система обязанностей, которую де-факто сформулировал Кант, должна регулировать наши эксперименты с роботами.

Предположим, что мы решили использовать автономных роботов (согласно определению в разделе 1) в военных действиях для уничтожения вражеской армии или мирных жителей. Одна из очевидных причин, по которой это может показаться предпочтительным по сравнению с войнами, ведущимися только людьми, заключается в том, что убийство людей часто приводит к травматическим последствиям, посттравматическому стрессовому расстройству и другим психологическим состояниям. Тем не менее, с кантовской точки зрения, такой путь противоречит морали, или мы так думаем. Если мы позволяем автономному роботу уничтожать жизнь, мы тем самым игнорируем и не уважаем достоинство человеческих жизней. Как утверждает Ульген [42], в таком сценарии относительная цель, например, защита солдата от посттравматического стрессового расстройства, ставится выше фундаментального принципа человечности как объективной ценности. Следовательно, создавая роботов, убивающих людей, мы нарушаем формулу человечности. Аналогично, использование роботов не должно наносить ущерб или нарушать материальное благополучие других людей. Если бы мы решили создать роботов, которые сделали бы других людей ненужными или низвели бы их до состояния рабства, мы бы низвели абсолютную ценность человечности до простого средства.

6. Заключение

Мы показали, что концепция моральной агентности Канта предлагает ответы на некоторые из основных вопросов робоэтики. Мы рассмотрели три таких вопроса: могут ли роботы квалифицироваться как моральные агенты, целесообразно ли, чтобы роботы развивали моральную способность и допустимо ли препятствовать развитию моральной способности у роботов.

С кантианской точки зрения перспективы превращения роботов с сильным ИИ в моральных агентов сомнительны. Но, несмотря на низкую вероятность того, что роботы когда-нибудь станут моральными агентами в кантовском смысле, остается вопрос о том, целесообразно ли это и, если нет, допустимо ли препятствовать развитию моральной способности у роботов. Мы показали, что даже при либеральной интерпретации метафизики морали Канта роботы с сильным ИИ не являются подходящим объектом моральных размышлений ради них самих. Для нас морально допустимо препятствовать достижению роботами моральной индивидуальности, поскольку они не являются

разумными агентами в практическом смысле и не разделяют нашу животную природу. Это, однако, не означает, что кантовская этика не накладывает моральных ограничений на разработку или использование роботов. Наше использование роботов регулируется теми обязанностями, которые мы имеем по отношению к самим себе. Таким образом, при разработке роботов с сильным искусственным интеллектом мы должны помнить об уважении разумной природы, как нашей собственной, так и других людей[18].

Литература

Все ссылки на работы Канта даны по Gesammelte Schriften, Ausgabe der Preußischen Akademie der Wissenschaften (Berlin: de Gruyter, 1902 ff.). Для отдельных работ используются следующие сокращения:

GMS Grundlegung zur Metaphysik der Sitten/Groundwork of the Metaphysics of Morals/Основы метафизики нравов

KpV Kritik der praktischen Vernunft/Critique of Practical Reason/Критика практического разума

MS Metaphysik der Sitten/Metaphysics of Morals/Метафизика нравов

Полный список литературы приводится в разделе *References* для сокращения объема статьи (прим. переводчика).

References

All references to Kant's works are to Kant's Gesammelte Schriften, Ausgabe der Preußischen Akademie der Wissenschaften (Berlin: de Gruyter, 1902 ff.). The following abbreviations of individual works are used:

GMS Grundlegung zur Metaphysik der Sitten/Groundwork of the Metaphysics of Morals/Основы метафизики нравов

KpV Kritik der praktischen Vernunft/Critique of Practical Reason/Критика практического разума

MS Metaphysik der Sitten/Metaphysics of Morals/Метафизика нравов

1. *Abney, K.* (2012). Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed. In: Patrick Lin/Keith Abney/George A. Bekey (Hrsg.): Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge/MA: MIT Press, 35–52.

2. *Allison, H.E.* (2011). Kant's Groundwork for the Metaphysics of Morals: A Commentary. Oxford: Oxford University Press.

3. *Ameriks, K.* (2003). On Two Non-Realist Interpretations of Kant's Ethics. In: Karl Ameriks (Hrsg.): Interpreting Kant's Critiques. Oxford: Oxford University Press, 263–283.

[18] Мы благодарны Эндрю Киму и Дитеру Шенекеру за комментарии к черновому варианту этой статьи. Кроме того, мы благодарим членов CONCEPT (Кельнского центра современной эпистемологии и кантовской традиции), участвовавших в работе Кантовской читательской группы, за плодотворное обсуждение этики Канта. Свен Бернекер хотел бы выразить благодарность за финансовую поддержку, оказанную ему профессорской премией Александра фон Гумбольдта. Лиза Бенosси благодарит Фонд Александра фон Гумбольдта за щедрую поддержку, оказанную ей во время пребывания в Кельне.

4. *Bacin, S. & Sensen, O.* (Hrsg.) (2018). The Emergence of Autonomy in Kant's Moral Philosophy. Cambridge: Cambridge University Press.

5. *Bauer, K.* (2018). Cognitive Self-Enhancement as a Duty to Oneself: A Kantian Perspective. Southern Journal of Philosophy, 56(1), 36–58.

6. *Bekey, G.A.* (2005). Autonomous Robots. Cambridge/MA: MIT Press.

7. *Danaher, J.* (2019). Should we create artificial moral agents? A Critical Analysis. URL: https://philosophicaldisquisitions.blogspot.com/2019/09/should-we-create-artificialmoral. html, visited on 6.7.2021

8. *Düsing, K.* (1993). Spontaneità e Libertà Nella Filosofia Pratica di Kant. In: Studi Kantiani, 6, 23–46.

9. *Engstrom, S.P.* (Hrsg.) (2009). The Form of Practical Knowledge: A Study of the Categorical Imperative. Cambridge/MA: Harvard University Press.

10. *Frankfurt, Harry* (1969). Alternate Possibilities and Moral Responsibility. In: Journal of Philosophy, 66(23), 829–839.

11. *Jaarsma, P., Gelhaus, P. & Welin, S.* (2012). Living the Categorical Imperative: Autistic Perspectives on Lying and truth Telling–between Kant and Care Ethics. In: Medicine, Health Care and Philosophy, 15(3), 271–277.

12. *Kain, P.* (2004). Self-legislation in Kant's Moral Philosophy. In: Archiv für Geschichteder Philosophie, 86(3), 257–306.

13. *Kain, P.* (2009). Kant's Defense of Human Moral Status. In: Journal of the History of Philosophy, 47(1), 59–101.

14. *Kant, I.* (1999). Practical Philosophy / Trans. and ed. Mary J. Gregor. Cambridge: Cambridge University Press.

15. *Kant, I.* (2002). Critique of Practical Reason / Trans. by Werner S. Pluhar. Indianapolis: Hackett Publishing.

16. *Kleingeld, P.* (2010). Moral Consciousness and the 'Fact of Reason'. In: Jens Timmerman, Andrew Reath (Hrsg.). A Critical Guide to Kant's 'Critique of Practical Reason'. Cambridge: Cambridge University Press, 55–72.

17. *Kleingeld, P. & Willaschek, M.* (2019). Autonomy without Paradox: Kant, Self-Legislation and the Moral Law. In: Philosopher's Imprint, 19(6), 1–18.

18. *Korsgaard, C.M.* (1996). The Sources of Normativity. Cambridge: Cambridge University Press.

19. *Korsgaard, C.M.* (2018). Fellow Creatures: Our Obligations to the Other Animals. Oxford: Oxford University Press.

20. *Langton, R.* (2007). Objective and Unconditioned Value. In: Philosophical Review, 116(2), 157–185.

21. *Lueck, B.* (2009). Kant's Fact of Reason as Source of Normativity. In: Inquiry, 52(6), 596–608.

22. *Merkel, R.* (2002). Rechte für Embryonen. In: Julian Nida-Rümelin (Hrsg.): Ethische Essays. Frankfurt: Suhrkamp, 427–239.

23. *Nida-Rümelin, J.* (2002). Humanismus ist nicht teilbar. In Julian Nida-Rümelin (Hrsg.): Ethische Essays. Frankfurt: Suhrkamp, 463–469.

24. *O'Neill, O.* (2003). Autonomy: The Emperor's New Clothes. In: Aristotelian Society Supplementary 77. pp. 1–21.

25. *Puls, H.* (2016). Sittliches Bewusstsein und kategorischer Imperativ in Kants ›Grundlegung‹: Ein Kommentar zum dritten Abschnitt. Berlin: de Gruyter.

26. *Rawls, J.* (1980). Kantian Constructivism in Moral Theory. In: Journal of Philosophy, 77(9), 515–572.

27. *Regan, D.* (2002). The Value of Rational Nature. In: Ethics 112, 267–291.

28. *Regan, T.* (2004). The Case for Animal Rights. Berkeley: University of California Press.

29. *Reath, A.* (1994). Legislating the Moral Law. In: Noûs 28, 4, 435–464.

30. *Rohlf, M.* (2020). Immanuel Kant, The Stanford Encyclopedia of Philosophy,

Edward N. Zalta (Hrsg.), URL: <https://plato.stanford.edu/archives/spr2020/entries/kant/>.

31. *Russell, S. J., Norvig, P.* (2016). Artificial Intelligence: A Modern Approach. Upper Saddle River: Pearson.

32. *Scanlon, T.M.* (1998). What We Owe to Each Other. Cambridge/MA: Harvard University Press.

33. *Schönecker, D.* (1999). Kant: Grundlegung III Die Deduktion des kategorischen Imperativs. Freiburg: Karl Alber.

34. *Schönecker, D.* (2010). Kant über Menschenliebe als moralische Gemütsanlage. In: A rchiv für Geschichte der Philosophie 92. 133–175.

35. *Schönecker, D.* (2013). Das gefühlte Faktum der Vernunft. Skizze einer Interpretation und Verteidigung. In: Deutsche Zeitschrift für Philosophie, 61(1), 91–107.

36. *Schönecker, D.* (2018). Gemütsanlagen, moralische. In: In Larissa Berger/Elke E. Schmidt: Kleines Kant-Lexikon. Paderborn: Wilhelm Fink/UTB, 156–7.

37. *Schönecker, D.* (2018). Can Practical Reason be Artificial?. In: Journal of Artificial Intelligence Humanities, 2, 67–91.

38. *Schönecker, D., Wood, A.W.* (2015). Immanuel Kant's Groundwork for the Metaphysics of Morals. Cambridge/MA: Harvard University Press.

39. *Sensen, O.* (Hrsg.) (2012). Kant on Moral Autonomy. Cambridge: Cambridge University Press.

40. *Sharkey, A.* (2020). Can we Program or Train Robots to be Good?. In: Ethics and Information Technology, 22. 283–295.

41. *Talbert, M.* (2014). The Significance of Psychopathic Wrongdoing. In: Thomas Schramme (Hrsg.): Being Amoral: Psychopathy and Moral Incapacity. Cambridge, MA: MIT Press. 275–300.

42. *Ulgen, O.* (2017). Kantian Ethics in the Age of Artificial Intelligence and Robotics. In: Questions of International Law, 43. 59–83.

43. *Ware, O.* (2014). Rethinking Kant's Fact of Reason. In: Philosopher's Imprint, 14(32), 1–21.

44. *Warren, Mary Anne* (1997). Moral Status: Obligations to Persons and Other Living Things. Oxford: Oxford University Press.

45. *Williamson, D.M.* (2009). Emotional Intelligence and Moral Theory: A Kantian Approach. Dissertation, Vanderbilt University.

46. *Wood, A.W.* (1998). Kant on Duties Regarding Nonrational Nature. In: Proceedings of the Aristotelian Society Suppl. 72. 189–210.

47. *Wood, A.W.* (1999). Kant's Ethical Thought. Cambridge: Cambridge University Press.

48. *Wu, Katherine J.* (2020). Scientists Assemble Frog Stem Cells into First 'Living Machines'. URL: https://www.smithsonianmag.com/innovation/scientists-assemble-frog-stem-cells-first-living-machines-180973947/, visited on 6.7.2021

Сведения об авторе

Зайкова Алина Сергеевна – научный сотрудник Института философии и права СО РАН (630090, Новосибирск, ул. Николаева, 8),
Zaykova.a.s@gmail.com

Information about the author

Zaykova Alina Sergeevna – Researcher of Institute of Philosophy and Law of Siberian Branch of the Russian Academy of Sciences (8, Nikolaev st., Novosibirsk, 630090, Russia).

Дата поступления 04.11.2024