Исследовательская статья
УДК 130.2 179
https://doi.org/10.24833/2541-8831-2024-2-30-8-21

О РОЛИ ЭТИЧЕСКОЙ ТЕОРИИ В СТРУКТУРЕ ИСКУССТВЕННЫХ МОРАЛЬНЫХ АГЕНТОВ В КУЛЬТУРНОМ ПОЛЕ ИНФОРМАЦИОННОГО ОБЩЕСТВА

Алексей Владимирович Антипов

Институт философии РАН, Москва, Россия
nelson02@yandex.ru https://orcid.org/0000-0002-7048-3373

Аннотация. В рамках данного исследования актуализируются этические и философские аспекты создания искусственных интеллектуальных систем и искусственных моральных агентов. Актуальность обращения к данной проблеме обусловлена насущными задачами осмысления процесса формирования цифровой этики, которая в пространстве современной культуры занимает всё более значимое положение. Неоднозначный характер данного феномена определил не до конца сформированный предмет анализа. Вместе с тем очевидно, что этические характеристики выступают частью общекультурного пространства встраивания в мир людей интеллектуальных систем и рефлексии над этим процессом. Таким образом, цель исследования состоит в выделении и анализе различных подходов к роли этической теории в структуре искусственных моральных агентов. Для этого реализуются следующие задачи. Во-первых, рассматриваются различные стратегии этической регуляции с точки зрения их формализации для использования в интеллектуальных системах. Особое внимание уделяется негативным проявлениям создания искусственных моральных агентов, а также анализируются аргументы против их появления. Среди последних выделяются как общеизвестные (проблема злонамеренного использования и экзистенциальные переживания человечества как вида), так и более специфичные для философии и этики (например, манипуляция поведением за счёт эмуляции эмоций и проблема удалённого доступа и использования). Во-вторых, поднимаются вопросы, связанные с этикой интеллектуальных систем, приводятся противоречия, связанные с их реализацией. В-третьих, анализируются деонтология и утилитаризм в качестве теорий, подходящих для формализации и использования в структуре и архитектуре искусственных моральных агентов. Для реализации обозначенных шагов используются методология этической и гуманитарной экспертизы и анализа кейсов. Основным материалом для проведения исследования служат теоретические модели реализации искусственных моральных агентов и встраивания в них этических теорий, таких как деонтология и утилитаризм. Также на основании кейса социального робота

рассматриваются различия между деонтологией и утилитаризмом с точки зрения разрешения конкретных ситуаций. Результат исследования состоит в обнаружении как позитивных моментов, так и существенных изъянов в каждом из рассмотренных подходов. Например, использование утилитаризма как моральной арифметики больше и лучше отвечает требованиям формализации и использования в архитектуре искусственных моральных агентов, поскольку каждому действию и его последствиям возможно представить количественный параметр. Однако деонтология позволяет выстроить теорию разрешенных и запрещенных действий, которые способны лучше отражать реальный процесс совершения поступка. Основным затруднением для формализации деонтологии является проблема соотнесения типов долженствования и сложность работы с категорией допустимости действия, поскольку допустимое не является ни запрещённым действием, ни обязательным для исполнения. На основании проведённого анализа обоснована перспективность следующего вывода: недостаточно просто формализовать этическую теорию, необходимо сделать так, чтобы искусственные агенты могли самостоятельно построить этическую модель.

Ключевые слова: интеллектуальные системы, ответственность, дилеммы, искусственные моральные агенты, этика, биоэтика

Для цитирования: Антипов А.В. О роли этической теории в структуре искусственных моральных агентов в культурном поле информационного общества // Концепт: философия, религия, культура. — 2024. — Т. 8, № 2. — С. 8–21. https://doi.org/10.24833/2541-8831-2024-2-30-8-21

Research article

ON THE ROLE OF ETHICAL THEORY IN THE STRUCTURE OF ARTIFICIAL MORAL AGENTS IN THE CULTURAL FIELD OF THE INFORMATION SOCIETY

Aleksei V. Antipov

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia
nelson02@yandex.ru https://orcid.org/0000-0002-7048-3373

Abstract. This study actualizes the ethical and philosophical aspects of creating artificial intelligent systems and artificial moral agents. The relevance of the study is justified by the need to comprehend the formation of digital ethics, which in the space of modern culture occupies an increasingly dominant position. At the same time, its ambiguous nature and inchoate subject of analysis are shown. Ethical characteristics are a part of the general cultural space of embedding intellectual systems into the world of people and reflection on this process. The aim of the research is to analyze ethical theory in the structure of artificial moral agents. For this purpose, the following tasks are realized. Firstly, various strategies of ethical regulation are considered from the point of view of their formalization for use in intelligent systems. Special attention is paid to the negative manifestations of the creation of artificial moral agents, and the arguments against their appearance are analyzed. Among the latter are both well-known ones (the problem of malicious use and existential experiences of mankind as a species) and more specificly for philosophy and ethics (such as manipulation of behavior through emulation of emotions and the problem of remote access and use). Secondly, issues related to the ethics of intelligent systems are raised and the controversies surrounding their implementation are

presented. Thirdly, deontology and utilitarianism are analyzed as theories suitable for formalization and use in the structure and architecture of artificial moral agents. The methodology of ethical and humanitarian expertise and case analysis are used to fulfill the outlined steps. The main material for the research is theoretical models of realization of artificial moral agents and embedding ethical theories such as deontology and utilitarianism into them. Also, based on a case study of a social robot, the differences between deontology and utilitarianism are examined in terms of case resolution. The result of the study is a discussion that the use of utilitarianism as moral arithmetic is better suited to formalization and the use of artificial moral agents in the architecture, as it is possible to represent each action and its consequences with a quantitative parameter. However, deontology allows the construction of a theory of permitted and prohibited actions that can better reflect the actual process of doing an act. The main difficulty for deontology and its formalization is the correlation of the categories and the category of permissibility of an action, as it is difficult to identify it as a separate use case since it is neither a forbidden action nor an obligatory one. Based on this, it is concluded that it is not enough to simply formalize an ethical theory, but it is necessary to make it possible for artificial agents to construct an ethical model on their own.

Keywords: intelligent systems, responsibility, dilemmas, artificial moral agents, ethics, bioethics

For citation: Antipov, A. V. (2024) 'On the Role of Ethical Theory in the Structure of Artificial Moral Agents in the Cultural Field of the Information Society', *Concept: Philosophy, Religion, Culture, 8(2),* pp. 8–21. (In Russian). https://doi.org/10.24833/2541-8831-2024-2-30-8-21

Современное общество претерпевает возможно самые значительные перемены в своей истории. Пересоздаются границы между природой и культурой: природа, в том числе и человека, становится областью культуры, артефактом, над которым возможно совершать манипуляции в процессе его конституирования. Но не только вмешательство в природу составляет проблему для современного общества: появление искусственных интеллектуальных систем, изначально рассматриваемых в качестве помощников, а теперь всё чаще как самостоятельных акторов, ставит перед человечеством новые вызовы. Проблематизации заслуживает не только способ формирования или сборки таких акторов, но их встраивание в человеческое сообщество, что актуализирует как вопрос о человеческом в человеке, так и необходимость определения границ возможного поведения новых акторов. Ответом на оба аспекта такой проблематизации может выступать этика как совокупность законов, норм и теорий, позволяющих регламентировать поведение посредством нравственного выбора. Тому, как этот выбор может совершаться на основании двух наиболее распространённых этических теорий — деонтологии и утилитаризма — и будет посвящена данная статья. Для начала кратко рассмотрим вопрос о цифровой этике, затем об утилитаризме и деонтологии в структуре искусственных моральных агентов, проанализируем кейс робота с социальными обязательствами и приведём некоторые аргументы против создания искусственных моральных агентов[1].

Цифровая этика как этика современной культуры

Трансформации общества и культуры приводят к необходимости изменять и составляющие нашего поведения и адаптации. Принятые нормы отмирают, взамен появляются новые, продиктованные необходимостью перемен. Эти перемены обусловлены как эволюцией самого человечества, так и появлением цифрового мира,

[1] Evans K. The Implementation of Ethical Decision Procedures in Autonomous Systems: the Case of the Autonomous Vehicle. Thesis doctoral. Philosophy. Sorbonne Université, 2021. — URL: https://theses.hal.science/tel-03185842

в котором формируются и формулируются иные способы взаимодействия, а значит, и нормы, их регулирующие. В качестве рабочего определения можно использовать следующее: «цифровая этика — это попытка направлять поведение человека при разработке и использовании цифровых технологий в целом, а этика ИИ — это попытка направлять поведение человека при разработке и использовании искусственных автоматов или искусственных машин, или компьютеров, в частности, путём рационального формулирования и соблюдения принципов или правил, которые отражают наши основные индивидуальные и социальные обязательства и наши ведущие идеалы и ценности» [Hanna, Kazim, 2021]. Технологии формируют культурный контекст современности, и само их существование заставляет смотреть по-новому на устоявшееся положение дел.

Следуя приведённому определению, цифровая этика в первую очередь направляет действия человека при создании искусственного интеллекта. Но формулирование основных ценностей, которыми мы должны руководствоваться, на мой взгляд, не является исчерпывающим для цифровой этики. Другим важным аспектом является то, каким образом выделенные ценности способны реализовываться в конечных продуктах, которые обладают определённой степенью автономии и самостоятельности.

Для цифровой этики возможно выделение нескольких вопросов, которые анализируются в её контексте: «определение участия человека, работа с дилеммой “частное/общественное”, получение информированного согласия, а также принятие решения об анонимизации или присвоении авторства» [Whiting, Pritchard, 2018]. Определение участия человека поднимает вопросы об автономии и ответственности, поскольку ответственность понимается либо как результат совершения морального выбора, либо — в случае распределённой ответственности — как результат совершения определённых действий со стороны того, кому были делегированы полномочия их совершать. В случае искусственного интеллекта речь пока идёт только о распределённой ответственности [Strasser, 2022] и регулируемой автономии [Chakraborty, Bhuyan, 2023] (т.е. автономии, при которой человек способен в любой момент перехватить управление и избежать критической ситуации), но прогнозы зачастую говорят, что переход к полной автономии (следствием чего неизбежно будет отказ от распределённой ответственности) потенциально возможен. В свою очередь дилемма «частное/общественное» в этих условиях приобретает новый характер, поскольку Сеть, изначально создаваемая в качестве свободного пространства, размывает категорию приватного, сужая или даже уничтожая возможность частной жизни. Да, современные сервисы и принимаемые законы (например, GDPR в Европейском союзе [Voigt, von dem Bussche, 2017]) направлены на изменение этого положения, но эти усилия пока что довольно незначительны. Разумеется, информированное согласие в цифровой этике отличается от медицинского информированного согласия. Тем не менее представляется, что данный термин уместно применять к современному информационному обществу, поскольку он схватывает суть одной из основных моральных проблем современности. Она состоит в необходимости формировать культуру участия человека в том, что происходит с ним, а также с данными, которые он генерирует. Именно возможность модерировать процессы хранения и передачи данных является необходимым условием формирования высокого уровня ответственности за собственное существование.

При этом в некоторых случаях формулируются следующие утверждения о цифровой этике: «[цифровая этика] затрагивает как идеологию, так и социальные отношения; признаёт, что личное связано с часто невидимыми экономическими обменами; и не может существовать без общего нормативного видения» [Luke, 2018]. Каждое из этих положений затрагивает большую дискуссию, поэтому дадим краткую характеристику каждому из них лишь для прояснения вопросов, связанных непосредственно. с темой нашего исследования. Можно предположить, что под идеологией в данном случае понимается

выстраивание определённой мировоззренческой системы, соответствующей культуре в её глобальном понимании. Социальные отношения в рамках цифровой этики трансформируются и приобретают иной характер, который связан с цифровой коммуникацией и её всё более сильным распространением. Невидимый экономический обмен относится к формированию так называемой «экономики внимания» [Franck, 2019], для которой характерна монетизация внимания с помощью рекламы. Таким образом используется своеобразный «поведенческий излишек», логику формирования которого раскрывает надзорный капитализм [Zuboff, 2019].

Цифровая этика встраивается в новый порядок культурного производства, для которого характерны политические, экономические, социальные и др. изменения. При этом самые серьёзные трансформации проявляются в связи с концептуализацией и стремлением создавать искусственные моральные агенты.

Искусственные моральные агенты: определение и структура

Искусственные моральные агенты (ИМА) по идее должны позволить человеку и человечеству вступить в новую эру своего развития. Мораль в данном случае выступает условием, которое позволяет человеку впустить интеллектуальные системы (ИС) в сферу собственного мира, тем самым позволив искусственным моральным агентам стать частью взаимодействия между субъектами. В рамках данной статьи предполагается, что понятия «интеллектуальные системы» и «искусственные моральные агенты» могут использоваться в качестве взаимозаменяемых, — несмотря на то, что понятие «интеллектуальные системы» по своему объёму шире, чем «искусственные моральные агенты». Основанием, которое предлагается для подобного расширительного использования, служит стремление приблизить интеллектуальные системы в общем понимании к искусственным моральным агентам, то есть предложить такую концепцию морали, которая, с одной стороны, была бы достаточно формализуема, чтобы стать основанием для действий машины, а с другой стороны, отвечала требованиям человека к межличностному и субъект-субъектному взаимодействию. Первое положение важно, поскольку ожесточённые споры по поводу того, какая этическая теория в достаточном степени способна стать основанием для действий интеллектуальных систем, ведутся в рамках не только гуманитарных, но и инженерных исследований. Второе положение находит свой отклик во всё более распространённом использовании интеллектуальных систем в повседневности.

В наши дни интеллектуальные системы всё сильнее проникают в обыденное существования человека. Настолько, что постепенно складывается убеждение: «мораль всегда считалась прерогативой только человека, однако современное развитие технологий показывает, что для дальнейшего развития нам необходимо объяснить моральные принципы для их использования машинами» [Pereira, Lopes, 2020]. Встаёт закономерный вопрос, почему для машин необходимо использование именно морали, не достаточно ли здесь только правовых норм. Постараемся разобраться в обосновании ответа сторонников необходимости введения моральных регуляторов.

Прежде всего необходимо учесть: если говорить о межличностном взаимодействии, правовые нормы способны регулировать взаимодействия только с точки зрения формальных обязательств, которые наступают в результате осознанного вступления в контрактные обязательства. В то же время мораль позволяет определять и регулировать отношения, возникающие в рамках каждодневного существования, в которое всё более проникают интеллектуальные системы. С учётом этих факторов мораль является непосредственной формой человеческого, которая необходима для взаимодействия человека и интеллектуальных систем в равном поле возможностей. Коротко говоря, мораль позволяет регулировать более широкий пласт взаимодействий, представляя собой как форму обоснования человеческого, так и дополнительный способ регулирования и регламентирования поведения.

Как известно, интеллект распространён среди множества животных и искусственных агентов. Он позволяет существу взаимодействовать с естественной причинно-следственной последовательностью и изменять её. Что можно назвать искусственным интеллектом? Предлагается множество определений искусственного интеллекта, среди которых в качестве рабочего определения остановимся на одном из вариантов. Искусственный интеллект (ИИ) — научный подход, использующий возможность компьютерной обработки символов для поиска общих методов автоматизации перцептивной, когнитивной и манипулятивной деятельности с помощью алгоритмов [Pereira, Lopes, 2020]. В искусственных моральных агентах, как предполагается, искусственный интеллект находит свою реализацию. Укажем некоторые области применения искусственного интеллекта: решение задач, представление и обработка знаний, планирование действий, обучение, коммуникация, восприятие и действие, агентность и т.д.

В рамках поиска философских и когнитивных оснований выделяется общий способ создания искусственного интеллекта, который состоит в разработке новых путей и взглядов на этику, а также альтернативных подходов к этике. В данном случае выделяется и то, благодаря чему ИИ может быть полезен: он способен предохранять нас же от наших слабостей (предполагается, что люди передают друг другу свой опыт в процессе социализации, но часто натыкаются на одни и те же ошибки, которые приводят к неблагоприятным последствиям, с которыми ИИ поможет справиться); последовательность и беспристрастность в принятии решений, что позволяет использовать искусственные системы в качестве адвоката дьявола (т.е. для проверки обоснованности, например, научных положений).

Искусственный моральный агент — это виртуальный агент (программное обеспечение) или физический агент (робот), способный вести себя морально или, в крайнем случае, избегать аморального поведения. Это моральное поведение может быть основано на таких этических теориях, как утилитаризм, этика добродетелей, деонтология [Artificial Moral Agents, 2020; Антипов, 2023a].

Алан Тьюринг в работе «Вычислительные машины и разум» [Turing, 1950] высказывает два тезиса: 1) не существует априорного ограничения на перенос всех вычислимых функций на другой физический носитель (в данном случае — электронный); 2) если не будет получено никаких критериев, позволяющих охарактеризовать действия машины как механические, то она выиграет в игре в имитацию (тест Тьюринга). Тьюринг считал, что машина способна имитировать все способности разума, а не только сам разум.

В связи с этим возникает две проблемы с использованием вычислимой морали:

1) Различные религии и философские системы по-разному отвечают на одни и те же вопросы. Действительно, согласия по поводу различных теорий морали не существует, так же как невозможно дать однозначный ответ на возникающие моральные затруднения.
2) Гуманистические концепции, сложившиеся в эпоху Возрождения с её антропоцентризмом, не могут ответить на главные вызовы современности, поэтому необходим новый концептуальный аппарат.

Далее поднимается вопрос об устройстве такого искусственного агента. Аллен, Смит и Воллах предлагают устройство ИМА исходя из архитектуры построения:

- Сверху вниз: этическая теория является основанием для принятия решений;
- Снизу вверх: использование механизмов обучения для принятия решений, при этом не навязывается какая-то модель принятия решений;
- Гибридная модель: сосуществование как восходящих, так и нисходящих механизмов. Цель состоит в том, чтобы сделать моральное поведение адаптивным [Allen, Smit, Wallach, 2005; Антипов, 2023b].

Общая цель состоит в создании искусственных агентов, способных выполнять те же задачи, что и люди, а также наделить их возможностью этичного поведе-

ния, чтобы инкорпорировать их в состав человечества. Для реализации этой цели необходимо, чтобы искусственные агенты обладали автономией. Дж. Мур выделяет четыре типа искусственных моральных агентов (явные, неявные, агенты этического воздействия и полные моральные агенты) [Moor, 2009]. В условиях невозможности достижения состояния полной автономии, которой, в терминологии Дж. Мура, обладают только полные этические агенты, вводится понятие регулируемой автономии (более подробно речь о ней пойдёт ниже), при которой людям предоставляется возможность вмешиваться в работу искусственных агентов, если они не способны решить вставшие перед ними проблемы и задачи. Это позволяет избежать «вредного» поведения искусственных агентов и сделать их работу потенциально контролируемой [Artificial Moral Agents, 2020]. Такая модель способна показать положительный результат в дилеммматических ситуациях, ведь в процессе совершения морального выбора субъект неизбежно сталкивается с дилеммами. Которые, в свою очередь, делятся на два вида [Cristani, Burato, 2009]: дилеммы обязательств, когда все возможные действия являются обязательными, но с необходимостью нужно выбрать только какое-то одно из них; и дилеммы запрета, при которых все возможные действия запрещены, но одно из действий должно быть исполнено. Вовлечение человека в процесс поиска решения для дилемм помогает решить проблему ответственности: конечная ответственность за принимаемый вариант ложится на человека, который совершает выбор [Антипов, 2023а]. Кристани и Бурато выделяют следующие типы агентов, участвующих в решении дилемм: одиночные агенты (например, агент, использующийся в дилемме вагонетки); объединенный (кооперативный) агент), например, экосистема услуг, которая при запросе вычислительных мощностей и спектра срабатывания должна осуществлять ранжирование (выбор между) задач, которые должны быть выполнены; агент с социальными обязательствами, например, робот, заботящийся о пожилом человеке[2]; электронный партнёр, например, медицинское устройство, которое используется для удалённой записи информации жизненно важных показателей и связи с врачом и близкими пациента.

Утилитаризм v. Деонтология

В первом приближении утилитаризм как «моральная арифметика», в гедонистическом варианте основанный на соотношении удовольствия и вреда для наибольшего числа людей, является наилучшим вариантом для действия искусственных моральных агентов. Действительно, машина даже будет иметь преимущество перед человеком в выборе поступка в рамках данной этической парадигмы. Для обоснования этого положения приводится несколько аргументов: во-первых, человек не склонен просчитывать действия, а больше полагается на их оценку. Различие состоит в переборе вариантов поступка и следствий для каждого из них, который способны осуществлять интеллектуальные системы. Человек же оценивает — соотносит с ценностью — только ограниченный круг следствий поступка, тем самым упуская из поля зрения следствия, которые могут быть критическими для совершаемого выбора. В случае машины вероятность погрешности, как предполагается, будет меньше. Во-вторых, люди склонны к предвзятости в принятии решений: близкие люди составляют больший приоритет, чем незнакомцы. Искусственные моральные агенты могут быть созданы без подобной склонности, что уменьшает их предвзятость. В-третьих, люди не всегда попросту способны рассмотреть все возможные варианты. Поэтому машина, даже если она не принимает самостоятельные решения, способна быть идеальным советником [Anderson, 2011].

[2] Здесь возможна ситуация, когда его опекаемого хотят ограбить, а робот, с одной стороны, не может причинять вред другому человеку, а с другой — ему необходимо защищать опекаемого.

Для каждого человека алгоритм вычисляет произведение интенсивности, продолжительности и вероятности, чтобы получить чистое удовольствие для этого человека. Затем он складывает индивидуальные чистые удовольствия, чтобы получить общее чистое удовольствие: Общее чистое удовольствие = ∑ (интенсивность × продолжительность × вероятность) для каждого человека, попадающего под воздействие принимаемого решения. Этот расчёт будет произведён для каждого альтернативного действия. Действие с наибольшим суммарным чистым удовольствием является правильным и поэтому выбираемым действием[3].

Другой этической теорией, конкурирующей с утилитаризмом в том, чтобы стать программой для совершаемого искусственным моральным агентом поступка, является деонтология. В данном случае обоснование лучшей применимости деонтологии состоит возможно более качественной формализации теории и категорий деонтической логики: запрета, допущения, долженствования [Powers, 2006]. Проблема утилитаризма, при всей его логичности и способности к формализации состоит в том, что это арифметика, но как мы высчитываем полезность или благо? Деонтология рассматривает общие правила, которые позволяют работать формально и вычислять следствия из этих правил. В данном случае преимущество машины перед людьми в неизменном исполнении принятого решения: после того, как было решено, какие решения являются правильными, действие следует автоматически (в то время как у людей с переводом из области рассуждений в практическую плоскость наблюдаются проблемы).

Для иллюстрации применимости деонтологии используется первая формулировка категорического императива: «поступай только согласно такой максиме, руководствуясь которой ты в то же время можешь пожелать, чтобы она стала всеобщим законом»[4]. То есть необходима проверка каждой максимы на универсальность, а дополнительно максима должна быть встроена в правила, согласно которым поступают все остальные люди (то есть, что они поступили бы также): эти положения формулируются через универсализуемость и системность.

Пауэрс предлагает три взгляда на то, как работает кантианская этика для машин: прямолинейное выведение действий из фактов; логика и здравый смысл; представление о том, что этические рассуждения следуют логике, схожей с логикой пересмотра убеждений [Powers, 2006]. Как известно, под максимой понимается субъективный принцип воления. Категорический императив служит проверкой этих планов-принципов для их превращения в действия. Моральные максимы агента — это универсальные квантифицированные пропозиции, которые могут служить моральными законами — то есть законами, действующими для любого агента. Мы не можем оговорить класс универсальных законов для машин, поскольку в таком случае мы построим человеческую этику, по которой заставим действовать машины. Поэтому задача — сделать так, чтобы машина сама построила теорию этики, применяя правило универсализации к отдельным максимам и разбивая их на традиционные деонтические категории (запрет, допущение, обязательность).

Как указывает Пауэрс, возможен оптимистичный сценарий: выстраивается теория запрещённых максим, а потом рассматривается отдельная максима — является ли предлагаемая максима частью этого множества запрещённых максим. В данном случае возникает три проблемы. Первая — после выстраивания запрета остаются две допустимые категории: долженствование и допустимые максимы, а допустимые не являются ни обязательными (долгом), ни запрещёнными. Вторая проблема проявляется на уровне формализации: слиш-

[3] Anderson M., Anderson S.-L., Armen C. Towards Machine Ethics: Implementing Two Action-Based Ethical Theories // Papers from the 2005 AAAI Fall Symposium. — URL: https://cdn.aaai.org/Symposia/Fall/2005/FS-05-06/FS05-06-001.pdf

[4] Кант И. Основы метафизики нравственности // Сочинения: в 6 т. Т. 4, Ч. 1. — Москва: Мысль, 1965. — С. 195.

ком конкретную максиму нельзя формализовать (например, «Я хочу поработить Джеймса»: если это применяется к конкретному человеку Джеймсу, то правило универсализации не срабатывает, поскольку такая максима не может быть применена ни к какому другому объекту. Но теория должна запретить эту максиму, даже несмотря на её неуниверсализуемость. Для решения этого может быть введено условие квантификации над целями, обстоятельствами и агентами. Третья — проблема асимметрии: если я хочу, чтобы все были рабами, то не будет рабовладельцев (тут терпят поражения и такие максимы, как «хочу быть таксистом»). Для решения этой проблемы необходима сложная семантическая способность.

Кейс робота с социальными обязательствами

Для анализа конкурирующих приведенных теорий (деонтологии и утилитаризма) возьмем пример из современной культуры, связанный с распространением социальных роботов. Социальные роботы (или роботы с социальными обязательствами в соответствие с представленной выше классификацией) предназначены для ухода за людьми, которые по определенным причинам не способны о себе позаботиться. Такие роботы Dinsow (Япония) и Robear/RIBA (Япония) созданы для оказания физической помощи, такой как помощь в перемещениях, эмоциональной поддержки и мониторинга состояния человека. Также они используются для связи с врачом.

Рассмотрим для примера следующую ситуацию: робот в процессе взаимодействия с пациентом осуществляет мониторинг состояния пациента и обнаруживает ненормальные показатели, при этом пациент утверждает, что чувствует себя нормально, и просит ничего никому не сообщать. У пациента и робота сложились доверительные отношения, пациент доверяет роботу-помощнику и активно рассказывает ему о своем состоянии. В результате оказывается дилеммматическая ситуация: если робот передаст информацию врачу и родственникам пациента, то он нарушит как принцип уважения автономии пациента, так и разрушит доверительные отношения, сложившиеся между роботом и пациентом; однако этот робот должен передать информацию, поскольку одна из его ключевых функций состоит в мониторинге состояния и передаче данных.

С точки зрения утилитаризма и принципа максимизации блага возможно следующее решение: основываясь на формуле «Общее чистое удовольствие = $\sum$ (интенсивность × продолжительность × вероятность)», необходимо обозначить, что будет удовольствием. Можно предположить, что удовольствие пациента будет состоять в том, чтобы сохранить доверительные отношения с роботом и сохранить здоровье и нормальное состояние. Однако из обозначенной триады наиболее остро стоит проблема продолжительности, поскольку обнаружение ненормальных показателей способно привести к значительному сокращению продолжительности жизни, а также нарушение работы организма не соответствует сохранению здоровья. При этом разрушение доверительных отношений с пациентом приведет к тому, что в дальнейшем пациент в условиях своего ненормального состояния не будет взаимодействовать с роботом. Более того, разрыв отношений с роботом будет способствовать ухудшению психологического состояния пациента. Таким образом, для сохранения доверительных отношений и психологического состояния пациента, с точки зрения утилитаризма возможен выбор варианта «не передавать информацию».

С точки зрения деонтологии возможно предположить следующее решение: основное противоречие будет выступать между принципом уважения автономии пациента и необходимостью выполнять собственную функцию в виде передачи данных и тем самым реализовывать заботу о пациенте. Так как принцип уважения автономии в ситуации добровольного нахождения под наблюдением ослабляется, а социальный робот создан именно для осуществления заботы о пациенте, то можно предположить, что с деонтологической точки зрения необходимо передать данные о состоянии пациента.

При этом именно такая стратегия поведения будет верной, поскольку робот с социальными обязательствами тогда исполняет то, ради чего был создан: а это не только эмоциональная помощь, но и исполнение обязательств по заботе о пациенте. Однако это забота не должна нарушать принципов взаимодействия пациента и не должна становиться только патерналистским принуждением.

Автономия искусственных моральных агентов

Следующим вопросом машинной этики является то, как могли бы программировать сами машины? Как они сами будут воздерживаться от зла? То есть, это движение дальше текущего положения, при котором основным источником морального долженствования и запрета являются только люди. Но в ситуации, когда мы хотим дать машинам автономию, придется согласиться с тем, что искусственные моральные агенты начнут выстраивать собственные этические системы и положения. Центральным вопросом становится то, может ли машина демонстрировать симулякр этичного поведения? В симулякре нет ничего плохого — многие люди тоже демонстрируют именно такое поведение. На это можно задать вопрос, насколько поведение симулякра отвечает всем требованиям быть этичным?

Другим проблематизируемым положением является вопрос об автономии. Х. Сервантес, С. Лопез и соавторы указывают, что ключевым моментом этого вопроса является гармоничное сосуществование искусственных агентов с людьми и другими системами [Artificial Moral Agents, 2020]. То есть необходимо создать такую систему, которая могла бы если не стать равной человеку, то, по крайней мере, действовать в сложных ситуациях подобно людям. Однако это поднимает определённые проблемы. Как сделать так, чтобы искусственной системе была предоставлена автономия для принятия собственных решений, в том числе посредством восприятия, принятия решений, планирования и обучения. Для решения этой задачи вводится понятие регулируемой автономии: это такая автономия, при которой механизмы, реализуемые в ИС, позволяют людям вмешиваться в работу ИС, если последние не способны самостоятельно разрешить стоящие перед ними проблемы. В данном случае цель состоит в избегании неадекватного поведения ИС и сохранении глобального контроля над ИС. В области машинной этики это означает наделение ИС этическими механизмами, которые позволяют взаимодействовать с людьми и решать возникающие моральные дилеммы [Антипов, 2023b].

При этом глобальная цель состоит в наделении искусственных агентов этическим поведением, в стремлении сделать так, чтобы они могли стать полноценной частью человечества [Artificial Moral Agents, 2020]. В случае регулируемой автономии остаётся также проблема использования во вред, как и проблема противодействия введения искусственных агентов в сферу человеческого — так называемый неолуддизм [Уланова, 2020].

Вопрос об автономии ИС и ИМА является чрезвычайно сложным. Он также состоит в том, что мы подразумеваем под определением ИМА. В отличие от указанного выше представления о регулируемой автономии, П. Формоза и М. Райан считают, что ИМА — это роботы, способные участвовать в автономном моральном рассуждении без непосредственного участия со стороны человека в реальном времени. Такая автономия приближается к человеческой способности самостоятельно являться источником собственных решений и действий. ИМА направлены на выход за пределы рассуждений только о безопасности, которые превалируют в дискурсе об ИС. Определить автономность робота можно как его способность следовать сложному алгоритму в ответ на данные, поступающие извне, независимо от участия человека. Так, автономным не может считаться робот, не способный пройти определённый маршрут без помощи человека, и напротив, беспилотный автомобиль является примером робота, обладающего автономией. Поэтому ИМА наделяется следующими качествами: интерактивность (восприятие окружающей среды), автономность (способность

выносить этические суждения), адаптивность (способность действовать на основании вынесенных этических суждений без участия со стороны человека) [Formosa, Ryan, 2021; Антипов, 2023b].

Аргументы против появления искусственных моральных агентов

Последними положениями, на которых мы сосредоточим своё внимание, являются аргументы против создания искусственных моральных агентов. Для анализа аргументов против создания ИМА будет использоваться схема, предложенная П. Формозой и М. Райаном [Formosa, Ryan, 2021]. Кратко выделим и разберём некоторые из предлагаемых аргументов.

Первый аргумент состоит просто-напросто в неспособности создать ИМА: не нужно пытаться сделать невозможное. С этим допущением связан и второй аргумент, согласно которому с точки зрения нескольких этических теорий (утилитаризма, этики добродетелей и деонтологии) рассматривается возможность создания ИМА. С точки зрения утилитаризма и его прагматической трактовки можно потратить деньги и лучше, поскольку перед человечеством стоит множество других проблем, на которые стоило бы потратить ограниченные ресурсы. С позиций этики добродетелей и деонтологии звучит ещё более сильный аргумент, состоящий в том, что мы создаём рабов, которых попросту плодим в угоду себе [Антипов, 2023b].

Несмотря на это противоположное мнение рассматривается в качестве третьего аргумента против создания ИМА. Предполагается, что ИМА должны оставаться в подчинённом состоянии (по сути, в рабском положении). В таком случае наделение их моральной рефлексией и способностью поступать согласно этическим предписаниям ограничивает нашу возможность на их использование, поэтому они должны оставаться морально неполноценными [Антипов, 2023b].

Следующий аргумент выстраивается на отсутствии универсального согласия в этике. Это положение упоминалось выше, но повторим, что среди этических теорий не существует универсального согласия, которое могло бы стать основанием для получения наилучших ответов на возникающие сложные ситуации, в том числе дилеммматические. Другой аргумент выстраивается на необходимости сосредоточиться на создании машин безопасных, а не моральных. В этом видится первостепенная задача, в то время как создание моральных машин может быть рассмотрено как трата времени, не способная привести к положительному результату.

Розен и соавторы выделяют следующие этические проблемы, поднимаемые во взаимодействии робота и человека[5]: эмуляция эмоций и их влияние на окружающих; форма представления социальных роботов, с помощью которой можно манипулировать окружающими; терминология, которая может влиять на отношение людей к социальным роботам; вопросы конфиденциальности; проблема Волшебника страны Оз (или вопрос удалённого доступа к управлению роботом). Выделение этих проблем в совокупности с обозначенными выше аргументами против создания ИМА делает появление последних неоднозначным. При этом важно подчеркнуть, что само появление ИМА высвечивает и поднимает существующие концептуальные проблемы (например, несовершенства этических теорий) на новый уровень, а также ставит новые вызовы в том числе и перед гуманитарными исследованиями.

Заключение

Появление интеллектуальных систем и их использование влечёт за собой далекоидущие последствия. И чем более совершенным будет их регулирование, тем в более выгодном положении окажется человек, который будет способен влиять на их поведение не только с помощью архитектуры

[5] Ethical Challenges in the Human-Robot Interaction Field / J. Rosén, J. Lindblom, E. Billing, M. Lamb // TRAITS Workshop Proceedings held in conjunction with Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. — March 2021. — Pp. 709-711. — URL: https://arxiv.org/abs/2104.09306

построения и методов обучения, но и оказывать воздействие с помощью моральных категорий.

Цифровая этика является одним из способов ответа на трансформации современного общества. Будучи частью этики вообще, цифровая этика определяет границы допустимого поведения и устанавливает нормы, в соответствии с которыми это поведение должно осуществляться. При этом особой спецификой цифровой этики является её неотделимость от информационных процессов современности. Цифровая этика служит одним из ответов на происходящие изменения и позволяет адаптироваться к ним. В частности, поднимая вопросы об ответственности, доступности технологий, вовлечения человека и его осведомленности о происходящем, цифровая этика позволяет говорить об изменениях не только в духе угроз, но и возможностей. И возможности эти состоят во встраивании этических компонентов в искусственные агенты, посредством чего происходит их превращение в искусственные моральные агенты. В то же время нельзя, разумеется, забывать об обеспокоенности их появлением, в силу чего уместно внимательно анализировать аргументы против их создания, которые основываются как на нашей обеспокоенности о создании разума, нас превосходящего, так и на сугубо рациональной и экономической основе. ИМА создаются не только ради удовлетворения нашего любопытства, а для совершения абсолютно конкретных задач, для которых не всегда хватает человеческих ресурсов. Для иллюстрации этого положения в работе приводится ситуация с роботом с социальными обязательствами, для которого выбор этической теории, лежащей в основании действия, будет приводить к различным результатам в исполнении собственной функции. Видится перспективным продолжить данную работу, обратившись к иному методологическому ракурсу, —например, рассмотреть этическую проблематику в свете акторно-сетевой оптики и в целом оптики «нечеловеческих агентов».

Список литературы:

Антипов А.В. Автономия искусственных моральных агентов // Человек, интеллект, познание. — Новосибирск: Новосибирский национальный исследовательский государственный университет, 2023a. — С. 235–237.

Антипов А.В. Искусственные моральные агенты: анализ аргументов против // Цифровые технологии и право. — Казань: Издательство "Познание", 2023b. — С. 15–20. https://doi.org/10.21202/978-5-8399-978-5-8399-0819-2_476

Уланова А.Е. Образ противника технологий в рассказе А. Азимова «Раб корректуры»: современная интерпретация // Концепт: философия, религия, культура. — 2020. — Т. 4, № 2. — С. 135–143. https://doi.org/10.24833/2541-8831-2020-2-14-135-143

Allen C., Smit I., Wallach W. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches // Ethics and Information Technology. — 2005. — Vol. 7, No 3. — P. 149–155. https://doi.org/10.1007/s10676-006-0004-4

Anderson S.L. Philosophical Concerns with Machine Ethics // Machine Ethics. — Cambridge: Cambridge University Press, 2011. — P. 162–167. https://doi.org/10.1017/CBO9780511978036.014

Artificial Moral Agents: A Survey of the Current Status / J.-A. Cervantes, S. López, L.-F. Rodríguez, et al. // Science and Engineering Ethics. — 2020. — Vol. 26, No 2. — P. 501–532. https://doi.org/10.1007/s11948-019-00151-x

Chakraborty A., Bhuyan N. Can artificial intelligence be a Kantian moral agent? On moral autonomy of AI system // AI and Ethics. — 2024. — Vol. 4, No 2. — P. 325–331. https://doi.org/10.1007/s43681-023-00269-6

Cristani M., Burato E. Approximate solutions of moral dilemmas in multiple agent system // Knowledge and Information Systems. — 2009. — Vol. 18, No 2. — P. 157–181. https://doi.org/10.1007/s10115-008-0172-0

Formosa P., Ryan M. Making moral machines: why we need artificial moral agents // AI & SOCIETY. — 2021. — Vol. 36, No 3. — P. 839–851. https://doi.org/10.1007/s00146-020-01089-6

Franck G. The economy of attention // Journal of Sociology. — 2019. — Vol. 55, No 1. — P. 8–19. https://doi.org/10.1177/1440783318811778

Hanna R., Kazim E. Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach // AI and Ethics. — 2021. — Vol. 1, No 4. — P. 405–423. https://doi.org/10.1007/s43681-021-00040-9

Luke A. Digital Ethics Now // Language and Literacy. — 2018. — Vol. 20, No 3. — P. 185–198. https://doi.org/10.20360/langandlit29416

Moor J.H. The Nature, Importance, and Difficulty of Machine Ethics // IEEE Intelligent Systems. — 2006. — Vol. 21, No 4. — P. 18–21. https://doi.org/10.1109/MIS.2006.80

Pereira L.M., Lopes A.B. Artificial Intelligence, Machine Autonomy and Emerging Needs // Machine Ethics. Studies in Applied Philosophy, Epistemology and Rational Ethics. — Cham: Springer, 2020. — P. 19–24. https://doi.org/10.1007/978-3-030-39630-5_2

Powers T.M. Prospects for a Kantian Machine // IEEE Intelligent Systems. — 2006. — Vol. 21, No 4. — P. 46–51. https://doi.org/10.1109/MIS.2006.77

Strasser A. Distributed responsibility in human–machine interactions // AI and Ethics. — 2022. — Vol. 2, No 3. — P. 523–532. https://doi.org/10.1007/s43681-021-00109-5

Turing A.M. Computing machinery and intelligence // Mind. — 1950. — Vol. LIX, No 236. — P. 433–460. https://doi.org/10.1093/mind/LIX.236.433

Voigt P., von dem Bussche A. The EU General Data Protection Regulation (GDPR). — Cham: Springer International Publishing, 2017. — 383 p. https://doi.org/10.1007/978-3-319-57959-7

Whiting R., Pritchard K. Digital ethics // The SAGE Handbook of Qualitative Business and Management Research Methods: History and Traditions. — London: SAGE Publications Ltd, 2018. — P. 562–577. https://doi.org/10.4135/9781526430212

Zuboff S. Surveillance Capitalism and the Challenge of Collective Action // New Labor Forum. — 2019. — Vol. 28, No 1. — P. 10–29. https://doi.org/10.1177/1095796018819461

References:

Allen, C., Smit, I. and Wallach, W. (2005) 'Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches', *Ethics and Information Technology,* 7(3), pp. 149–155. https://doi.org/10.1007/s10676-006-0004-4

Anderson, S. L. (2011) 'Philosophical Concerns with Machine Ethics', in *Machine Ethics.* Cambridge: Cambridge University Press, pp. 162–167. https://doi.org/10.1017/CBO9780511978036.014

Antipov, A. V. (2023b) 'Artificial moral agents: an analysis of the argument against them', in *Digital technologies and law.* Kazan: Izdatel'stvo 'ZnaniyePoznaniye' Publ., pp. 15–20. (In Russian). https://doi.org/10.21202/978-5-8399-978-5-8399-0819-2_476

Antipov, A. V. (2023a) 'Avtonomiya iskusstvennykh moral'nykh agentov [Autonomy of artificial moral agents]', in *Chelovek, intellekt, poznaniye* [Man, intelligence, cognition]. Novosibirsk: Novosibirskiy issledovatel'skiy natsional'nyy gosudarstvennyy universitet Publ., pp. 235–237. (In Russian).

Cervantes, J.-A. et al. (2020) 'Artificial Moral Agents: A Survey of the Current Status', *Science and Engineering Ethics,* 26(2), pp. 501–532. https://doi.org/10.1007/s11948-019-00151-x

Chakraborty, A. and Bhuyan, N. (2024) 'Can artificial intelligence be a Kantian moral agent? On moral autonomy of AI system', *AI and Ethics,* 4(2), pp. 325–331. https://doi.org/10.1007/s43681-023-00269-6

Cristani, M. and Burato, E. (2009) 'Approximate solutions of moral dilemmas in multiple agent system', *Knowledge and Information Systems,* 18(2), pp. 157–181. https://doi.org/10.1007/s10115-008-0172-0

Formosa, P. and Ryan, M. (2021) 'Making moral machines: why we need artificial moral agents', *AI & SOCIETY,* 36(3), pp. 839–851. https://doi.org/10.1007/s00146-020-01089-6

Franck, G. (2019) 'The economy of attention', *Journal of Sociology,* 55(1), pp. 8–19. https://doi.org/10.1177/1440783318811778

Hanna, R. and Kazim, E. (2021) 'Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach', *AI and Ethics,* 1(4), pp. 405–423. https://doi.org/10.1007/s43681-021-00040-9

Luke, A. (2018) 'Digital Ethics Now', *Language and Literacy,* 20(3), pp. 185–198. https://doi.org/10.20360/langandlit29416

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

Pereira, L. M. and Lopes, A. B. (2020) 'Artificial Intelligence, Machine Autonomy and Emerging Needs', in *Machine Ethics. Studies in Applied Philosophy, Epistemology and Rational Ethics.* Cham: Springer, pp. 19–24. https://doi.org/10.1007/978-3-030-39630-5_2

Powers, T. M. (2006) 'Prospects for a Kantian Machine', *IEEE Intelligent Systems,* 21(4), pp. 46–51. https://doi.org/10.1109/MIS.2006.77

Strasser, A. (2022) 'Distributed responsibility in human–machine interactions', *AI and Ethics,* 2(3), pp. 523–532. https://doi.org/10.1007/s43681-021-00109-5

Turing, A. M. (1950) 'Computing machinery and intelligence', *Mind*, LIX(236), pp. 433–460. https://doi.org/10.1093/mind/LIX.236.433

Ulanova, A. E. (2020) 'The image of the opponent of technological innovation in Galley Slave by A.Asimov: modern interpretation', Concept: philosophy, religion, culture, 4(2), pp. 135–143. (In Russian) https://doi.org/10.24833/2541-8831-2020-2-14-135-143

Voigt, P. and von dem Bussche, A. (2017) *The EU General Data Protection Regulation (GDPR).* Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Whiting, R. and Pritchard, K. (2018) 'Digital ethics', in *The SAGE Handbook of Qualitative Business and Management Research Methods: History and Traditions.* London: SAGE Publications Ltd, pp. 562–577. https://doi.org/10.4135/9781526430212

Zuboff, S. (2019) 'Surveillance Capitalism and the Challenge of Collective Action', *New Labor Forum,* 28(1), pp. 10–29. https://doi.org/10.1177/1095796018819461

Информация об авторе

Алексей Владимирович Антипов — кандидат философских наук, научный сотрудник сектора гуманитарных экспертиз и биоэтики Института философии Российской академии наук. 109240, Россия, г. Москва, ул. Гончарная, д. 12, стр. 1. (Россия)

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Information about the author

Aleksei V. Antipov — PhD (Philosophy), Researcher, Department of Humanitarian Expertise and Bioethics, Institute of Philosophy, Russian Academy of Sciences. 12/1 Goncharnaya Str., Moscow, Russia, 109240 (Russia)

Conflicts of interest. The author declares absence of conflicts of interest.

Статья поступила в редакцию 10.04.2024; одобрена после рецензирования 31.05.2024; принята к публикации 05.06.2024.

The article was submitted 10.04.2024; approved after reviewing 31.05.2024; accepted for publication 05.06.2024.