ВЫЗОВЫ СОВРЕМЕННОСТИ В ОПТИКЕ ФИЛОСОФИИ И СОЦИОГУМАНИТАРНЫХ НАУК

Научная статья
DOI 10.15826/koinon.2024.04.1.2.001
УДК 179

ИСКУССТВЕННЫЕ МОРАЛЬНЫЕ АГЕНТЫ: ДЕОНТОЛОГИЯ И МОРАЛЬНЫЙ ТЕСТ ТЬЮРИНГА

Алексей Владимирович Антипов

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

Аннотация: Исследования искусственного интеллекта являются одним из основных трендов современности. При этом научным сообществом осознаются значительные этические риски, связанные с его появлением, среди которых непрозрачность принятия решений, невозможность определения субъекта ответственности, проблема предвзятости и справедливости. Для минимизации и элиминации обозначенных рисков в литературе предлагается концепт искусственных моральных агентов, поведение которых выстраивалось бы на человеческих этике и морали. Однако возникают сопутствующие затруднения, наиболее важным из которых является невозможность однозначного выбора этической теории, связанная с их принципиальной несогласованностью. В данной статье предлагается рассмотреть деонтологию как один из вариантов, который может быть использован в структуре искусственных моральных агентов. Автор показывает преимущества деонтологии перед другими этическими системами: более простая формализация самой теории, системность и универсализуемость принципов, — а также потенциальные проблемы, такие как различение долженствования и категории допустимости, а также формализация конкретных максим. Однако помимо построения и выбора подходящей теории возникают затруднения, связанные с тем, каким

образом определить, что созданный искусственный моральный агент в достаточной степени морален. Для этого в статье рассматривается моральный тест Тьюринга, который призван помочь ученым в этом вопросе.

Ключевые слова: искусственный интеллект, искусственный моральный агент, деонтология, тест Тьюринга, моральный тест Тьюринга

Для цитирования: *Антипов А. В.* Искусственные моральные агенты: деонтология и моральный тест Тьюринга // Koinon. 2024. Т. 4. № 1–2. С. 9–17. DOI: 10.15826/koinon.2024.04.1.2.001

Original article

ARTIFICIAL MORAL AGENTS: DEONTOLOGY AND THE MORAL TURING TEST

Aleksei V. Antipov

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

Abstract: Artificial intelligence research is one of the trends of our time. Both it is stated not only about the need to develop artificial intelligence, but also about the significant ethical risks due its emergence. Such risks are understood as non-transparency of decision-making, lack of identifying a subject of responsibility, the problem of bias and fairness. To minimise and eliminate these risks, the science suggests the concept of artificial moral agents, whose behaviour would be based on human ideas about ethics and the morality of action. However, there are additional difficulties, the most important of which is the impossibility of unambiguous choice of ethical theory, i.e. the problem of fundamental incoherence of ethical theories. This article proposes to consider deontology as one of the options that can be used in the structure-building of artificial moral agents. The author shows the advantages of deontology over other ethical systems, which consist in the simpler formalisation of the theory itself, as well as in the systematicity and universalizability of the principles. At the same time, potential problems are also shown, such as the problem of the distinction between oughtness and the category of permissibility, and in the formalisation of specific maxims. However, if we need to construct and embed a suitable theory, there is the problem of how to determine that the artificial moral agent created is sufficiently moral. Finally, the paper discusses the Turing moral test, which is designed to help determine whether an artificial agent is moral.

Key words: artificial intelligence, artificial moral agent, deontology, Turing test, moral Turing test

For citation: Antipov, A. V. (2024), "Iskusstvennye moral'nye agenty: deontologiya i moral'nyi test T'yuringa" [Artificial Moral Agents: Deontology and the Moral Turing Test], *Koinon*, vol. 4, no. 1–2, pp. 9–17 (in Russian). DOI: 10.15826/koinon.2024.04.1.2.001

Исследования искусственного интеллекта являются одними из наиболее перспективных в настоящее время. При этом увеличение количества работ в области компьютерных наук, посвященных возможности создания сильного искусственного интеллекта, порождает большое количество исследований гуманитарного профиля, направленных на осмысление перспектив появления искусственного интеллекта в человеческом сообществе. Одними из наиболее остро стоящих проблем являются этические проблемы, такие как непрозрачность систем искусственного интеллекта, именуемая также проблемой «черного ящика», проблема справедливого распределения, цифрового разрыва и определения субъекта ответственности. Каждой из этих проблем посвящено множество работ. В данной статье деонтология рассматривается как наиболее подходящая теория для использования в искусственных моральных агентах (далее — ИМА), обосновывается пригодность теста Тьюринга для проверки моральности искусственного интеллекта. Эти положения связаны следующим образом: использование этической теории в структуре искусственных моральных агентов необходимо с точки зрения архитектуры, а моральный тест Тьюринга может использоваться в качестве инструмента, который показывает, насколько адекватно встроенная теория отвечает нашим представлениям о морали.

Колин Аллен и соавторы в статье «Зачем машинная этика?» утверждают: нам не важно, чтобы машины были действительно моральными субъектами, нужно сделать так, чтобы они вели себя как моральные субъекты, а их поведение удовлетворяло бы моральным нормам. Зачем же нужна машинная этика? Во-первых, искусственные агенты всё сильнее вторгаются в нашу жизнь и неотрефлексированное внедрение ценностей только разработчиками в создаваемые технологии способно приводить к непредсказуемым последствиям (например, легкое копирование информации заставляет переосмыслить право интеллектуальной собственности) [Allen, Wallach, Smit 2006]. Действительно, разговор о машинной этике должен быть разделен на несколько частей. Во-первых, нам необходимо определить то, как именно возможно приспособить этический аппарат к искусственным агентам. Мало сделать так, чтобы их поведение строго подчинялось каким-либо нормам: для адаптивного поведения необходимо соблюдение условий, при которых агент способен самостоятельно определять траекторию своих действий. В таком случае для ИМА необходима определенная этическая теория, которая лежала бы в основе принятия конкретного решения в определенной ситуации.

Несмотря на то, что искусственные агенты не должны вести себя полностью как моральные субъекты, они должны быть наделены способностью определять то, насколько их поведение удовлетворяет требованиям, существующим в обществе. В рамках морали это особенно сложно, поскольку однозначного морального кодекса не существует, а поэтому каждый раз приходится иметь дело с выбором. Здесь и приходит на помощь этическая теория, даже в редуцированном виде способная дать ИМА ориентиры для выбора наилучшего действия. Во-вторых, поднимается вопрос о ценностях: какие ценности и каким образом должны быть реализованы в ИМА? Речь идет о непредвиденных последствиях, к которым это встраивание может приводить. Однако в данном случае нас будет интересовать именно поведение и инструмент его оценки, а разговор о ценностях не будет затрагиваться.

Необходимо сказать, что деонтология не является единственной теорией, подходящей для реализации в ИМА. Действительно, интуитивно более удобным вариантом является утилитаризм, представляющий собой моральную арифметику, а кто, как не машина, способен наилучшим образом осуществлять арифметику. Однако в данном случае нам представляется важным не останавливаться на утилитаризме, а показать преимущества деонтологии, в том числе и с точки зрения формализации.

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, в 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 можем высчитать полезность или благо? Критерии, выбираемые утилитаризмом, не всегда могут носить общезначимый характер. В то же время 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 определено, 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» [Кант 1965, с. 195]. То есть необходима проверка каждой максимы на универсальность, а дополнительно максима должна быть встроена в правила, согласно которым поступают все остальные люди: эти положения формулируются через универсализуемость и системность.

Т. М. 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 Т. М. Пауэрс, возможен оптимистичный сценарий: определяется множество запрещенных максим, а потом каждая отдельная максима рассматривается на предмет ее отнесенности к запрещенным. В данном случае возникает три проблемы. 1. После выстраивания запрета остаются две допустимые категории: долженствование и допустимые максимы, а допустимые не являются ни обязательными (долгом), ни запрещенными. 2. Проблема на уровне формализации. Слишком 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 ее неуниверсализуемость, поскольку в данном случае нарушается принцип, согласно которому рабство в самой своей сути является безнравственным. Для решения этого может быть введено условие квантификации над целями, обстоятельствами и агентами. 3. Проблема асимметрии. Если я хочу, чтобы все были рабами, то не будет рабовладельцев (тут терпят поражение и такие максимы, как «хочу быть таксистом»). Для решения этой проблемы необходима сложная семантическая способность.

Однако не всегда можно сказать, что указанная выше краткая схема реализации деонтологии в структуре искусственных моральных агентов лишена сопутствующих недостатков. Укажем некоторые этические риски, выделяемые в связи с появлением ИМА. Стоит отметить, что эти риски характерны для всех этических теорий, которые могут быть использованы в ИМА. В данном случае предлагается опираться на риски, выделяемые С. Кэйвом и соавторами [Cave, Nyrup, Vold, Weller 2018]. Они полагают, что может быть выделено четыре типа рисков: разногласия между моральным рассуждением и моральным поведением; объяснимости и прозрачности; моральных обязанностей; моральной ответственности. Остановимся подробнее на каждом из указанных типов.

Требование объяснимого и прозрачного искусственного интеллекта необходимо, поскольку некоторые решения требуют контекстуальных объяснений. Особенно это заметно в медицине, где от каждого решения зависит жизнь человека. Следующим риском является создание новых моральных обязанностей у человека, поскольку если мы рассматриваем искусственных моральных агентов именно в этическом статусе, то и относиться к ним мы должны соответствующе. Это также означает серьезное отношение к интересам ИМА. Предполагается, что к ИМА, если они действительно поступают морально, необходимо будет относиться соответственно. Это означает не только включение в спектр моральных обязательств нечеловеческих агентов, но и размывание самого понятия морали и человеческого. Мораль является сферой исключительно человеческой, и включение в нее акторов, кардинально отличающихся от нас, способно заставлять нас переосмыслить сами основания собственного существования. Еще одним риском выступает парадокс автоматизации, который возникает в том числе во всех случаях использования здоровьесберегающих технологий: автоматизированные системы приспосабливаются к некомпетентности и автоматически исправляют ошибки, что приводит к эрозии навыков, но, в свою очередь, автоматизированные системы могут давать сбой, к которому человек окажется не готов. Эрозия навыков проявляется в потере способности совершать действия, которые предыдущие поколения делать могли. Несмотря на то что эрозия навыков сопровождает человечество на протяжении всей своей истории и мы всё больше полагаемся на технику, зачастую не понимая, как она работает, в случае с ИМА эрозии может быть подвергнуто само моральное чувство. Перекладывание ответственности на искусственный интеллект способно привести к потере чувствительности к ответственности и способности возлагать ее на себя, что является важным для морального субъекта.

Наконец, подвергается сомнению сама традиция интеллектуализма в этике, поскольку возможность морального рассуждения не подразумевает, что поведение будет моральным. Так, человек (или любое иное существо, в данном случае ИМА) способен строить моральные рассуждения, допустим, о разрешении дилеммы вагонетки. Но то, как поступит этот человек в какой-то конкретной ситуации, может кардинально отличаться от его умозрительных построений. Моральность человека определяется не способностью к рассуждениям, но действиями и поступками. Таким образом, необходим инструмент, благодаря которому возможно проверить, является ли само поведение моральным, а не только возможность рассуждения, реализованная в ИМА. Таким инструментом выступает моральный тест Тьюринга.

На каких основаниях реализуется моральный тест Тьюринга? В данном случае необходимо прояснить, что собой представляет классический тест Тьюринга. 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 победит в игре в имитацию (тест Тьюринга). Тьюринг считал, что машина может имитировать все способности разума, а не только сам разум. Так, тест Тьюринга применительно к исследованиям искусственного интеллекта представляет собой игру в имитацию, согласно правилам которой наблюдателю необходимо определить, является ли разговаривающий с ним человеком или компьютером. В случае, если наблюдатель не может однозначно классифицировать актора как машину, компьютер проходит тест. Под компьютером или машиной может пониматься и искусственный интеллект.

Как доказать, что машина является моральным агентом? Здесь ответом может быть моральный тест Тьюринга: в нем происходит смещение акцента с разговорной части на действие, то есть тест выстраивается таким образом, что даются пары описаний морально значимых действий реального человека и искусственного агента. Если человек сможет различить, где ответ дала машина, а где человек, то машина не прошла тест [Kim, Byun 2021]. Но возможно следующее критическое замечание: неотличимость решений машины от решений человека может устанавливать слишком низкий стандарт для ИМА. Имеется в виду то, что в своей повседневной жизни человек не всегда поступает в соответствии с моральными принципами, нормами и правилами. Поэтому действие отдельного человека не всегда может служить ориентиром.

Моральный тест Тьюринга выстраивается на предпосылках бихевиоризма, согласно которым не обязательно рассматривать внутренний мир и психику для того, чтобы говорить о мышлении. В таком случае о морали мы также можем говорить опираясь на то, как человек реагирует на стимулы и инициатором какого поведения он выступает. Мораль реализуется как оценка определенных действий, в основании которой может лежать концепция эмотивизма. Поэтому именно основание, на котором выстраивается возможность как оригинального теста Тьюринга, так и морального теста Тьюринга, концептуально важно. Для теста Тьюринга таким основанием является предположение о том, что если нам кажется, что машина мыслит, то мы должны считать это утверждение верным. Здесь не происходит проблематизация того, что значит «мыслить», а вывод основывается только на впечатлении того, кто проводит тест. Для морального теста Тьюринга используется видоизмененное положение, которое звучит так: если нам кажется, что машина действует морально, то мы должны считать, что ИМА обладают способностью быть моральными. Так становится возможным определить, является ли искусственный интеллект искусственным моральным агентом, только посредством внешней оценки.

Моральный тест Тьюринга выступает средством верификации поступков, продиктованных встраиванием деонтологии в структуру ИМА. Имеется в виду то, что действие, продиктованное универсализуемым и встроенным в общую систему правилом, может рассматриваться как моральное, даже если оно отличается от того, как люди поступают в среднем. В таком случае опровергается возражение, указанное выше и состоящее в том, что неотличимость действий ИМА от действий человека задает слишком низкий стандарт для ИМА. Возьмем для примера патерналистскую модель медицины: врач не всегда мог говорить о смертельном диагнозе пациенту, хотя это нарушает как принцип запрета на ложь, так и принцип уважения автономии пациента. В то же время ИМА, для которого приоритетным выступает правило уважения автономии, должен сказать пациенту правду об ожидающей его участи.

Моральный тест Тьюринга используется для проверки моральности ИМА только на основании впечатления того, кто проводит тест. Основаниями такого подхода выступают бихевиоризм в психологии и эмотивизм в этике. При этом деонтология способна показывать лучшие результаты в сравнении с утилитаризмом, поскольку она лучше формализуема и служит лучшему представлению о том, что значит быть моральным, то есть выстраивать отношения между субъектами с точки зрения идеального. Однако эти подходы не могут обойтись без критических замечаний. Значит ли это, что ИМА будут подлинно моральными, т. е. будут руководствоваться не только общими правилами, выработанными людьми в процессе развития, но и принципами заботы о себе и других. Еще одним важным вопросом будет определение эвристического потенциала технологии: человечество не останавливается на однажды установленных нормах и правилах, а разрабатывает и создает новые. Смогут ли ИМА создать новые нормы, которые впишутся в наши представления о том, что такое мораль?

Список источников

Кант 1965 — *Кант И.* Основы метафизики нравственности // Соч. : в 6 т. Т. 4. Ч. 1. М. : Мысль, 1965. 544 с.

Cave, Nyrup, Vold, Weller 2018 — *Cave S., Nyrup R., Vold K., Weller A.* Motivations and risks of machine ethics // Proceedings of the IEEE. 2018. Vol. 107. № 3. P. 562–574.

Allen, Wallach, Smit 2006 — *Allen C., Wallach W., Smit I.* Why machine ethics? // IEEE Intelligent Systems. 2006. Vol. 21. № 4. P. 12–17.

Kim, Byun 2021 — *Kim H., Byun S.* Designing and applying a moral Turing test // Advances in Science, Technology and Engineering Systems Journal. 2021. Vol. 6. № 2. P. 93–98.

Powers 2006 — *Powers T. M.* Prospects for a Kantian machine // IEEE Intelligent Systems. 2006. Vol. 21. № 4. P. 46–51.

Turing 1950 — *Turing A.* Computing machinery and intelligence // Mind. 1950. Vol. 59, iss. 236. P. 433–460.

References

Allen, C., Wallach, W., Smit, I. (2006), "Why Machine Ethics?", *IEEE Intelligent Systems,* vol. 21, no. 4, p. 12–17.

Cave, S., Nyrup, R., Vold, K., Weller, A. (2018), "Motivations and Risks of Machine Ethics", *Proceedings of the IEEE*, vol. 107, no. 3, p. 562–574 (in Russian).

Kant, I. (1965), "Osnovy metafiziki nravstvennosti" [Groundwork of the Metaphysics of Morals] in *Works*, vol. 4, part 1, Mysl', Moscow, 544 p. (in Russian).

Kim, H., Byun, S. (2021), "Designing and Applying a Moral Turing Test", *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 2, p. 93–98.

Powers, T. M. (2006), "Prospects for a Kantian Machine", *IEEE Intelligent Systems,* vol. 21, no. 4, p. 46–51.

Turing, A. (1950), "Computing Machinery and Intelligence", *Mind*, vol. 59, iss. 236, p. 433–460.

Статья поступила в редакцию 01.04.2024;
одобрена после рецензирования 15.04.2024;
принята к публикации 15.04.2024

The article was submitted 01.04.2024;
approved after reviewing 15.04.2024;
accepted for publication 15.04.2024

Информация об авторе

Антипов Алексей Владимирович — кандидат философских наук, научный сотрудник
Институт философии Российской академии наук, Москва, Россия

Information about author

Aleksei V. Antipov — Cand. Sci. (Philosophy), researcher
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia