9. *Nesterov A.* Working Knowledge: What Do I Know About How I Do Something? // Lecture Notes in Networks and Systems. 2023. Vol. 636 LNNS. P. 36-46. DOI: https://doi.org/10.1007/978-3-031-26783-3_5.

ИСКУССТВЕННЫЕ МОРАЛЬНЫЕ АГЕНТЫ В СФЕРЕ ИИ. ДВА ВЗГЛЯДА

Головков В.В.
Аспирант кафедры этики,
Санкт-Петербургский государственный университет
E-mail: golovkov.spb@mail.ru

Аннотация. В статье рассматривается понятие искусственного морального агента в контексте технологий искусственного интеллекта. В рамках исследования обозреваются два основных взгляда на вопрос о том, что из себя представляет искусственный интеллект как искусственный моральный агент. Первая позиция предполагает преемственность ИИ, как искусственного морального агента от естественного – человека. Человек понимается за идеального морального агента и, в зависимости от степени соответствия этому идеалу, можно говорить о степени моральной агентности искусственного интеллекта. В таком случае, задача заключается в обучении ИИ быть моральным. Вторая позиция заключается в том, что ИИ не может быть признан моральным агентом, однако, это не означает, что технология не имеет морального значения. ИИ воспринимается как инструмент, как посредник в моральных отношениях между людьми. Закономерно, задачей этики является этическое образование профессионалов, создающих ИИ. От выбора точки зрения зависит направление развития этики в сфере искусственного интеллекта.

Ключевые слова: этика, мораль, этика искусственного интеллекта, искусственный моральный агент, моральный агент, автономность.

Исследование выполнено при финансовой поддержке РНФ № 24-28-00562 «Философские основания этических рисков в сфере систем искусственного интеллекта».

ARTIFICIAL MORAL AGENTS IN AI. TWO VIEWS

Golovkov V. V.
Postgraduate student of Department of Ethics
Saint Petersburg State University
E-mail: golovkov.spb@mail.ru

Abstract. This article reviews two main views on the question of what constitutes artificial intelligence as an artificial moral agent. The first position assumes the continuity of AI as an artificial moral agent from the natural one - a human being. The human is understood as an ideal moral agent and, depending on the degree of compliance with this ideal, we can talk about the degree of moral agency of artificial intelligence. The ethical task is to train the AI to be moral. The second position is that AI cannot be recognized as a moral agent, however, this does not mean that the technology has no moral agency. AI is seen as an intermediary for humans in moral relationships. The task of ethics is the ethical education of professionals who create AI. The choice of point of view determines the direction of ethics development in the field of artificial intelligence.

Key words: ethics, morality, artificial intelligence ethics, artificial moral agent, moral agent, autonomy.

В современном этическом дискурсе одним из актуальных вопросов стало понятие морального агента. Помимо того, что исследователи спорят о самом феномене морального агентства, так же поднимается тема искусственных моральных агентов. В настоящий момент времени существует пример нечеловеческого, искусственного морального агента – корпорация. Первые дискуссии о моральном статусе корпораций были порождены концепцией корпоративной социальной ответственности (КСО), суть которой заключается в том, что корпорация имеет ответственность

не только юридическую – перед государством, но и моральную – перед обществом. Таким образом, исходя из этой концепции, коммерческие компании являются не только юридическими агентами, но и моральными. Несмотря на всю свою неоднозначность, КСО стала важной концепцией в мире бизнеса. Многие компании сегодня имеют специальные программы КСО и отчитываются о своей социальной и экологической деятельности. Однако, в последнее время не только корпорации стали претендовать на роль нечеловеческих моральных агентов. Многие исследователи считают, что искусственный интеллект так же мог бы быть признан моральным агентом. «Настаивать на неизбежной человеческой природе агента означает подорвать возможность понимания другой крупной трансформации в этической сфере, а именно появление искусственных агентов (ИА), достаточно информированных, "умных", автономных и способных совершать морально значимые действия независимо от создавших их инженеров-людей, вызывая "искусственное добро" и "искусственное зло"» [3, p. 249]. «Основной защищаемый тезис заключается в том, что искусственные агенты (ИА) являются легитимными источниками а/моральных действий, следовательно, понятие агента должно быть расширено, чтобы включить ИА, этический дискурс должен включать анализ их морали и, наконец, этот анализ необходим для понимания ряда новых моральных проблем не только в компьютерной этике, но и в этике в целом, особенно в случае распределенной морали» [1, p. 351]. Мы не будем рассматривать позиции, полностью отрицающие уникальную моральную роль ИИ, ведь они тривиальны в том плане, что сводятся к классическому пониманию концепта морального агента.

В общем и целом, можно выделить две позиции в отношении искусственного морального агента (ИМА). Первая предполагает, что ИМА не отделим от естественного морального агента – человека. Соответственно, вопрос о его статусе решается на основании приближённости искусственного агента к естественному, человек понимается за идеального морального агента и, в зависимости от степени соответствия этому идеалу, можно говорить о степени моральной агентности искусственного интеллекта. «Построение ИМА подчеркивает необходимость создания всеобъемлющей модели того, как люди приходят к удовлетворительным моральным суждениям» [7, p. 243]. Например, по такому принципу построена классификация, предложенная Дж. Муром в статье «Природа, значение и сложность машинной этики». В своей работе автор делит моральных агентов на 4 категории в зависимости от их автономности и сложности: агенты, оказывающие моральное воздействие; неявные моральные агенты; явные моральные агенты и полноценные моральные агенты.

К первой категории относятся машины, которые не являются автономными и не обладают интерактивностью. Это обычные, повседневные устройства. Однако, оказывающие положительное этическое влияние на общество. Например, роботизированные жокеи в Катаре, косвенно спасающие африканских детей от рабства.

Вторая категория — это машины, которые могут работать с некоторой степенью автономности в контролируемых ситуациях, но они не обладают подлинной моральной ответственностью. Их действия могут выглядеть этично, но не будут являться таковыми, потому что они основаны не на неких внутренних принципах, заложенных в систему, а на программе, составленной разработчиками.

К третьей категории относятся искусственные агенты, которые обладают определенным уровнем автономности и интерактивности. «Каким представляется подобный агент? Предположительно, он будет способен выносить правдоподобные этические суждения и обосновывать их. Самым впечатляющим может быть этический агент, автономный в том смысле, что он мог бы справляться с реальными ситуациями, включающими непредсказуемую последовательность событий» [5, p. 20].

Четвертая категория включает в себя человеческих агентов, которые обладают полной моральной агентностью. «Средний взрослый человек является полноценным этическим агентом. Обычно, мы наделяем людей сознанием, преднамеренностью и свободой воли. Может ли машина быть полноценным этическим агентом? Именно здесь дебаты о машинной этике становятся наиболее жаркими» [5, p. 20].

Статья Мура подчеркивает сложность разработки машинной этики, которая способствует адекватному рассуждению о моральных ценностях и принципах. Искусственные агенты имеют потенциал для развития в сложных моральных агентов, но они требуют обширных исследований и разработок для понимания различных этических соображений при принятии решений.

С точки зрения данного подхода, разница между ИМА и человеком как моральным агентом должна быть не качественная, а количественная. Основная задача заключается в том, чтобы сымитировать естественного морального агента на искусственной базе по аналогии с имитацией человеческого интеллекта у ИИ. Из данного подхода следует два вывода: либо рано или поздно ИИ станет моральным агентом в лучшей степени, чем человек, либо данная концепция окажется ошибочной и будет применима исключительно как «костыли» для создания внешне этичных ИИ.

«Если подход непрерывности воспринимается как утверждение о том, что моральное агентство является абсолютным, то будущие ИМА обязательно преодолеют наши моральные представления и станут гораздо лучшими моральными агентами, чем мы есть и можем когда-либо стать. <…> Напротив, если существенное различие между ИМА и человеческими моральными агентами может быть определено и правильно описано, тогда подход непрерывности окажется в целом ошибочным, хотя и сохранит свою силу в качестве постулата в области машинной этики» [2, p. 8].

Вторая позиция рассматривает ИИ как технологию. В рамках данного подхода ИИ не признаётся моральным агентом. Тем не менее, «отрицать, что компьютерные системы являются моральными агентами, не значит отрицать, что компьютеры имеют моральное значение или моральный характер; а утверждать, что компьютерные системы моральны, не обязательно значит утверждать, что они являются моральными агентами» [4, p. 195]. Человек и ИИ, в отличие от предыдущей точки зрения, признаются субъектами разного порядка. Искусственный интеллект – это инструмент, а он может существовать только в рамках человеческого целеполагания. Соответственно, «компьютерные системы имеют смысл и значение только по отношению к человеку» [4, p. 196]. Однако, искусственный интеллект всё равно оказывает моральное влияния. «Системы и устройства будут воплощать ценности независимо от того, хотим мы этого или нет» [6, p. 120]. Выступает же ИИ здесь не как моральный агент, но как посредник между людьми: между разработчиком и заказчиком/пользователем, между пользователем и пользователем, между пользователем и не-пользователем. «Выполняя свои функции, технологические продукты передают ценности, сознательно или бессознательно заложенные в них разработчиками, и при этом морально взаимодействуют с пользователями и другими вовлеченными людьми» [2, p. 17]. При этом, ИИ может действовать достаточно автономно, но всегда это делать сообразно тем ценностям, которые закладывают в него люди. Исходя из сказанного выше, данная точка зрения подразумевает ИИ как морального субъекта, так как, получается, он способен к моральному суждению, однако, в тех рамках и координатах, которые ему задают его создатели. А вот деятельность систем ИИ может лишь выглядеть как моральная, но она становится таковой только в контексте человека. Конечно, отвергая ИИ как морального агента, не стоит отвергать его значимую роль в моральном климате общества.

Рассмотрение данных позиций важно в контексте практического применения теоретического знания. При использования первого подхода своей задачей нам необходимо ставить обучение самого ИИ быть моральным. В случае второй позиции необходимо сосредоточиться, в первую очередь, на самих людях, которые создают ИИ.

В заключение следует сказать, что в современном этическом дискурсе, понятию морального агента уделяется большое количество внимания. В сфере технологий искусственного этот вопрос трансформируется в проблему искусственного морального агента. Несмотря на кажущуюся новизну данного понятия, ещё в прошлом веке появляется теория о корпорации как искусственном моральном агенте. Из этого следует, что, теоретически, появление новых нечеловеческих моральных агентов вполне реально. В сфере этики ИИ существует две точки зрения по поводу искусственных моральных агентов: первая говорит о преемственности между ИИ и человеком, а вторая утверждает ИИ как морального субъекта. В зависимости от приверженности той или иной позиции меняется вектор применения этики в сфере искусственного интеллекта: либо мы обучаем морали ИИ, либо людей, которые их создают.

Список литературы

1. *Floridi L., Sanders J.W.* On the Morality of Artificial Agents // Minds and Machine. 2004. № 14. P. 349-379 [Электронный ресурс]. URL: https://www.researchgate.net/publication/227190404_On_the_Morality_of_Artificial_Agents (дата обращения 24.04.2023).
2. *Fossa F.* Artificial moral agents: moral mentors or sensible tools? // Ethics and Information Technology. 2018. 20(1). URL: https://www.researchgate.net/publication/323819683_Artificial_moral_agents_moral_mentors_or_sensible_tools (дата обращения 24.04.2023).
3. *Gips J.* Towards the ethical robot // Android Epistemology. MIT Press, Cambridge MA, 1995. [Электронный ресурс]. URL: https://www.academia.edu/3089920/Towards_the_Ethical_Robot (дата обращения 24.04.2023).
4. *Johnson D.G.* Computer Systems. Moral Entities, but Not Moral Agents // Ethics and Information Technology 2011. № 8(4). P. 195-204, [Электронный ресурс]. URL: https://www.researchgate.net/publication/225564436_ Computer_systems_Moral_entities_but_not_moral_agents (дата обращения 24.04.2023).

5. *Moor J. H.* The Nature, Importance, and Difficulty of Machine Ethics // Intelligent Systems, IEEE. 2006. № 21(4). P. 18-21, [Электронный ресурс]. URL: https://www.researchgate.net/publication/220629129_The_Nature_Importance_and_Difficulty_of_Machine_Ethics (дата обращения 24.04.2023).

6. *Nissenbaum H.* How Computer Systems Embody Values // Computer. 2001. № 34(3). P. 118-120. [Электронный ресурс]. URL: https://www.researchgate.net/publication/2955398_How_Computer_Systems_Embody_Values (дата обращения 24.04.2023).

7. *Wallah W.* Robot minds and human ethics: The need for a comprehensive model of moral decision making // Ethics and Information Technology. 2010. № 12(3). P. 243-250. [Электронный ресурс]. URL: https://www.researchgate.net/publication/226274322_Robot_minds_and_human_ethics_The_need_for_a_comprehensive_model_of_moral_decision_making (дата обращения 24.04.2023).

ТРАНСФОРМАЦИЯ КАТЕГОРИИ СУБЪЕКТА ПОД ВЛИЯНИЕМ СОВРЕМЕННЫХ КОГНИТИВНЫХ ТЕХНОЛОГИЙ

Даниелян Н. В.
Доктор философских наук, профессор
Национальный исследовательский университет
«Московский институт электронной техники»
E-mail: vend22@yandex.ru

Аннотация. В статье рассматривается трансформация эпистемологических подходов к пониманию познавательных возможностей субъекта относительно объекта с позиции конструктивизма, согласно которой любые представления субъекта об окружающем мире представляют собой некий конструкт вероятностной модели. Автор полагает, что целью познания все больше становится не описание объективной реальности, а определенная организация ее субъективного восприятия, то есть совершенствование познавательных моделей, или трафаретов, через которые субъект «схватывает» реальность. Данное предположение полностью подтверждают концепции аутопоэза У. Матураны и Ф. Варелы, а также нейробиологический конструктивизм Г. Рота. В статье анализируется эпистемологический сдвиг под воздействием когнитивных технологий, являющихся неотъемлемой частью НБИКС-конвергенции. Автор делает вывод, что с созданием систем искусственного интеллекта, конструирующих определенным образом поле возможных состояний для построения представлений об объективном, происходит «размывание» границ понимания субъективного.

Ключевые слова: субъект, познание, опыт, конструктивизм, аутопоэз, когнитивные технологии, искусственный интеллект.

TRANSFORMATION OF CATEGORY OF SUBJECT UNDER THE INFLUENCE OF MODERN COGNITIVE TECHNOLOGIES

Danielyan N.V.
DSc in Philosophy, Professor
National Research University of Electronic Technology
E-mail: vend22@yandex.ru

Abstract. The article considers the transformation of epistemological approaches to understanding of cognitive opportunities of subject relative to object from the constructivism point of view. According to it, any human representation of the reality is just a construct of a probabilistic model. The author assumes the purpose of cognition in modern science is not to describe an objective reality, but to organize its subjective perception by some definite way. So, the purpose of cognition is the perfection of cognition models permitting a subject 'to capture' the reality. This idea is completely confirmed by Maturana and Varela's autopoiesis theory and Roth's neurobiological constructivism. The article analyses an epistemological shift under the influence of cognitive technologies, which are an integral part of NBICS-convergence. The author concludes the development of artificial intelligence technologies results in 'blurring' borders of