УДК 167.7
DOI: 10.15372/PS20240507
EDN: FNYLKK

А.С. Зайкова

АГЕНТНОСТЬ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ИНТЕЛЛЕКТУАЛЬНАЯ, СОЦИАЛЬНАЯ, МОРАЛЬНАЯ

Статья представляет собой комментарий к работе Лизы Беносси и Свена Бернекера «Кантианский взгляд на робоэтику». Развитие искусственных интеллектуальных систем привело к расширению или даже размытию понятия агентности, акцент в котором сместился с автономности намерения и действия на техническую автономность и социальный эффект человеко-машинного взаимодействия. Статья Беносси и Бернекера, рассматривающая кантианский взгляд на моральную агентность, возвращает нас ближе к классическому пониманию агентности, позволяя поставить на первое место не иллюзию автономности, но свободу выбора и моральную ответственность.

Ключевые слова: агентность, моральная агентность, искусственный интеллект, агент ИИ, человеко-машинное взаимодействие, социальный агент, кантианская позиция, автономность намерения

A.S. Zaykova

AGENCY OF ARTIFICIAL INTELLIGENCE SYSTEMS: INTELLIGENT, SOCIAL, MORAL

This article is a commentary on Lisa Benossi and Sven Bernecker's paper “A Kantian Perspective on Robot Ethics”. The development of artificial intelligent systems has led to an expansion or even blurring of the concept of agency, with the emphasis shifting from the autonomy of intention and action to the social effect of human-machine interaction. Benossi and Bernecker's article, which examines the Kantian view of moral agency, brings us back closer to the classical understanding of agency, allowing us to prioritize not the illusion of autonomy, but freedom of choice and moral responsibility.

Keywords: agency, moral agency, AI, AI agent, HRI, social agent, Kantian perspective, autonomy of intention

Вступление

С развитием искусственных интеллектуальных систем и робототехники понятие агентности подверглось серьезному пересмотру. Хотя классическая теория агентности по-прежнему имеет своих сторонников, предлагающих аргументы в пользу понимания агентности в первую очередь через автономность намерения и действия, значительную популярность приобрело направление исследований, сконцентрированное на агентности как на возможности машин, чат-ботов и иных искусственных интеллектуальных систем к самостоятельным выводам и действиям. Кроме того, всё более популярным становится подход к рассмотрению систем ИИ как уже включенных в нашу жизнь социальных агентов. Тем важнее становятся исследования, выбирающие своим предметом моральную агентность, включая потенциальную моральную агентность ИИ. Статья Беносси и Бернекера «Кантианский взгляд на робоэтику» [3] рассматривает проблему агентности именно с такой точки зрения, не только задаваясь вопросом о возможности моральной агентности для роботов, но и о её целесообразности для нас и для морального закона. В данном случае для нас важна не только сама постановка таких вопросов и предложение их рассмотреть с кантианской и стандартной позиции, но и различия в понимании агентности в разных направлениях исследований ИИ.

I. Философия действия

В философии действия наиболее влиятельными направлениями можно назвать стандартную концепцию действия и стандартную теорию действия. Стандартная концепция действия, во-первых, считает понятие намеренного действия более фундаментальным, чем понятие действия, и, во-вторых, утверждает о существовании тесной связи между намеренным действием и действием по причине. Как правило, имеется в виду, что намеренные действия обычно совершаются по причинам, то есть такие действия могут быть рационализированы «предпосылками обоснованного практического силлогизма» [12]. В отличие от стандартной концепции, в рамках стандартной теории обязательны описание намеренного действия и объяснение причины, чаще всего через правильные события и ментальные состояния. Тем не менее, и стандартная концепция действия, и стандартная теория действия во многом опираются на понятие намеренности.

Несмотря на широкую поддержку обоих направлений, требование объяснять агентность с точки зрения желаний, убеждений и намерений агента наталкивается на проблемы, возникшие из неопределенности

ментальных состояний, невозможности их точно определить, и возможности существования агентов без репрезентативных ментальных состояний, в результате чего ряд исследователей настаивают на отказе от обращения к ментальным состояниям при определении агентности. К примеру, Барандиаран с соавторами предлагает сосредоточиться на понятии минимальной агентности, предлагая следующее определение агента как «единой сущности, которая отличается от своей среды и которая делает что-то сама по себе в соответствии с определенной целью (или нормой)» [2]. Такой подход в значительной степени облегчает попытки применения понятия агентности не только к живым существам (включая бактерий), но и к искусственным системам. Конечно, сохраняется и другая позиция: так, согласно позиции Деннета, мы вполне можем приписывать ментальные состояния и животным, и искусственным системам [6]. В целом, разговоры об искусственной агентности уже не кажутся дурным тоном; напротив, они выступают как источник дополнительных аргументов для дискуссий, однако, как правило, предполагает, что мы можем встретиться с полноценной искусственной агентностью только в далеком будущем [2, 382].

Статья Беносси и Бернекера «Кантианский взгляд на робоэтику» также рассматривает возможность автономии и моральной агентности роботов, определяя моральную агентность как способность разумного существа сознательно выбирать свои жизненные цели [3]. Они показывают, что, согласно кантианскому взгляду, для достижения моральной агентности робот не только должен быть способен свободно принимать решения и действовать так, чтобы это выглядело морально, но и его практический разум должен быть автономным. Беносси и Бернекер приходят к заключению, что достижение моральной агентности для роботов невозможно или близко к невозможности, поскольку они не являются источником обаятельной силы морального закона. Даже если бы у роботов была бы внутренняя жизнь с заданными моральными нормами, это бы не сделало их моральными агентами.

II. Агенты ИИ

Несмотря на то, что с точки зрения философии действии полноценная агентность искусственных систем на текущий момент недостижима, в исследованиях, посвященных искусственному интеллекту, термин «агент» является одним из основных. В качестве примера можно привести книгу Рассела и Норвига «Искусственный интеллект: современный подход», где само понятие искусственного интеллекта определяется как наука об агентах, которые воспринимают (получают информацию из окружающей среды) и действуют (выполняют какие-

либо действия). Каждый такой агент реализует функцию, которая сопоставляет последовательности актов восприятия с действиями. Способами представления этих функций могут быть продукционные системы, реактивные агенты, нейронные сети, системы на базе теории решений и др. [11]. Рассел и Норвиг используют термин «агент» как сокращение выражения «интеллектуальный агент», кроме того, они определяют «рационального агента» как агента, выполняющего правильные действия. Для того, чтобы определить интеллектуального агента как рационального, необходимо оценивать показатели его производительности, при этом эта оценка должна быть объективной, а не опираться на субъективное мнение агента, поскольку, по их утверждению, не все агенты способны предоставить свое мнение, а некоторые склонны к самообману.

Как мы видим, такое представление об агентах никак не зависит от понятия намерения и от концепции намеренных действий. Намного важнее здесь то, что агент способен воспринимать и способен для действий использовать свое восприятие, а не только априорные данные, тогда он будет достаточно автономен, чтобы действовать рационально, и, следовательно, считаться рациональным агентом [11].

Действительно, в сфере исследований агентности агентов ИИ большое значение имеет вопрос автономии. Однако, как отмечают Беносси и Бернекер, «понятие автономии, используемое Кантом, существенно отличается от понятия автономии, используемого в ИИ (…). В ИИ робот считается автономным, если он не полностью зависит от своих предыдущих знаний, но способен интегрировать информацию, полученную из собственных представлений», в то время как в философии Канта автономия играет центральную роль, объясняя не только свободу, но и внутреннее достоинство человека [3]. Кантианский взгляд предполагает, что без автономии воли мы были бы подобны роботам [10]. Но если именно автономия воли отличает человека от робота, можем ли мы представить себе, что робот – наиболее автономный (физически) из агентов ИИ – может обладать какой-либо автономией так, как это понимается Кантом? Кажется, что ответ на этот вопрос очевиден: агенты ИИ не обладают ни агентностью, ни автономностью воли, а то, что мы приписываем им автономию действий, является лишь особенностью технического языка исследователей и разработчиков ИИ, подчеркивающего важность «обучения», а также, возможно, социального статуса агентов ИИ.

III. Агенты ИИ как социальные агенты

Выше мы уже отметили, что словосочетание «интеллектуальный агент» или «агент ИИ» является техническим термином, имеющим мало общего с классическим пониманием агентности в рамках философии действий. В свою очередь, агенты ИИ в виде социальных роботов или социальных ботов могут выступать как социальные агенты. В первую очередь при использовании этого термина обращается внимание на автономность социального поведения, хотя некоторые исследователи полагают, что для того, чтобы робот или иная интеллектуальная система рассматривалась как социальный агент, им достаточно обладать социальным интерфейсом [8].

Историю социальной робототехники часто начинают с роботизированных черепах, спроектированных Уолтером в 1940-х годах [7]. Эти черепахи не обладали способностями к социальному взаимодействию, но благодаря «фарам» и фотодатчикам, установленным на каждой из них, казалось, что они общаются («танцуют») друг с другом. В дальнейшем исследователи в большей степени сосредоточились на социальном взаимодействии роботов с людьми, хотя многие исследователи анализировали и взаимодействие роботов друг с другом. На настоящий момент под социальными роботами понимаются автономные роботы, которые взаимодействуют и общаются с людьми или другими автономными физическими агентами, следуя своей социальной роли и общепринятым социальным нормам. Как правило, они имеют мимику, имитируемую механическими (детали, имитирующие брови, глаза, рот; положение рук, головы и туловища) или электронными (экраны) способами. Чаще всего социальные роботы проектируются как гуманоидные (антропоморфные), однако существуют и социальные негуманоидные роботы, одним из самых известных примеров которых является Paro – робот в виде белька, детёныша гренландского тюленя. Paro продаётся с 2004 года и используется в качестве медицинского устройства для облегчения ухода за пациентами с деменцией как немедикаментозный антидепрессант [4]. В частности, эти роботы применяются в тех случаях, когда пациентам рекомендуют анималотерапию (зоотерапию), но использование настоящих животных сложно или невозможно, к примеру, в случае аллергии или неспособности пациента о ком-либо заботиться.

В целом, можно отметить, что социальные роботы уже играют значительную роль в нашей жизни, и эта роль со временем будет только увеличиваться [7]. Исследователи полагают, что во многом успех социальных роботов обосновывается их антропоморфизмом, благодаря которому люди склонны присваивать роботам социальные ха-

рактеристики [1]. Однако если ранее при разработке социальных роботов акцент был на их способностях, то на настоящий момент значительная часть исследователей сосредоточена на том, как робот выглядит в глазах пользователей. Хорстман и Крамер пишут: «границы между естественными и искусственными сущностями размываются, и, возможно, важнее сосредоточиться на том, воспринимается ли искусственный агент как действующий автономно, чем на том, способен ли он обладать собственным сознанием» [9]. Существует широко распространенное мнение, что намного важнее для исследования человеко-машинного взаимодействия (Human-Robot Interaction, HRI) то, какую социальную роль выполняют роботы, а не их фактические способности к автономным действиям (например, [8], [5], [9]).

Действительно, между автономным действием и восприятием действия как автономного существует значительная разница. К примеру, робот с удалённым управлением может восприниматься как обладающий внутренней жизнью с большой долей автономности, хотя он может быть не способен к самостоятельному принятию решений. В этом случае робот является не полноценным автономным агентом, но инструментом, расширяющим возможности агентов, которые им управляют. Однако пользователи, не зная об удаленном управлении, могут оценивать действия робота как последствия принятия решений искусственной интеллектуальной системой, воспринимая робота как социального агента, воплощающего некоторую социальную роль. В некоторых случаях приписывание роботам и иным агентам ИИ социальной агентности может скрывать под собой попытку перенести на СИИ ответственность за принимаемые решения и совершаемые действия, переводя внимание с разработчиков и владельцев систем искусственного интеллекта на несовершенство машинного разума и невозможности машинной морали.

Заключение

Выше мы рассмотрели агентность так, как ее понимает философия действия, а также понятия интеллектуальной и социальной агентности. Представление искусственных интеллектуальных систем как интеллектуальных агентов является во многом техническим вопросом, необходимым для более эффективного развития сферы ИИ. Автономность интеллектуального агента рассматривается как возможность действий, выходящих за рамки запрограммированных. Подход к интеллектуальным агентам с социальным интерфейсом как к социальным агентам является результатом исследований в области человеко-машинного взаимодействия, и важен в первую очередь именно потому,

что с точки зрения общества иллюзия автономности может быть важнее, чем реальная автономность. Оба этих подхода принципиально отличны от классического понимания агентности в рамках философии действия. Однако, хотя понятие моральной агентности также в значительной мере отходит от понятия агентности в рамках стандартной концепции и стандартной теории, они остаются тесно связанным, в частности, через представления о намерениях и свободе воли. Статья Беносси и Бернекера, рассматривающая кантианский взгляд на моральную агентность, возвращает нас ближе к классическому пониманию агентности, позволяя сосредоточиться не на иллюзии автономности, но на свободе выбора и моральной ответственности.

Литература

1. *Гаврилина Е.А*. Агентность не-человеков: взаимодействие людей и социальных роботов // Мониторинг общественного мнения: экономические и социальные перемены. 2023. № 3. DOI: 10.14515/monitoring.2023.3.2318

2. *Barandiaran X.E., Paolo E. Di, and Rohde M.* Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action // Adaptive Behavior. 2009. 17(5). P. 367–386.

3. *Benossi L. and Bernecker S.* A Kantian Perspective on Robot Ethics // Kant and Artificial Intelligence / Ed. by Hyeongjoo Kim and Dieter Schönecker. Berlin, Boston: De Gruyter. 2022. P. 145–168. DOI: 10.1515/9783110706611-005.

4. *Calo, Christopher J., Hunt-Bull N., Lewis, L., Metzler, L.* Ethical implications of using the paro robot with a focus on dementia patient care // Proceedings of the 12th AAAI Conference on Human-Robot Interaction in Elder Care (AAAIWS'11-12). AAAI Press. 2021. P. 20–24.

5. *Chen, N., Zhai, Y. & Liu, X.* The Effects of Robots' Altruistic Behaviours and Reciprocity on Human-robot Trust // Int J of Soc Robotics. 2022. 14, P. 1913–1931. DOI: 10.1007/s12369-022-00899-6.

6. *Dennett, D.C.* The Intentional Stance. Cambridge, MA: MIT Press, 1987.

7. *Fong, T. & Nourbakhsh, I. & Dautenhahn, K.* (2003). A Survey of Socially Interactive Robots // Robotics and Autonomous Systems. 2003. Vol. 42. 143–166. DOI: 10.1016/S0921-8890(02)00372-X.

8. *Fox J., Gambino A.* Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human-Robot Interaction // Cyberpsychol Behav Soc Netw. 2021 May; 24(5), 294–299. DOI: 10.1089/cyber.2020.0181. Epub 2021 Jan 11. PMID: 33434097.

9. *Horstmann A.C., Krämer N.C.* The Fundamental Attribution Error in Human-Robot Interaction: An Experimental Investigation on Attributing Responsibility to a Social Robot for Its Pre-Programmed Behavior // Int J of Soc Robotics. 2022. 14, 1137–1153. DOI: 10.1007/s12369-021-00856-9.

10. *Kant Immanuel*. Practical Philosophy / Trans. and ed. Mary J. Gregor. Cambridge: Cambridge University Press, 1999.

11. *Russel S.J. Norvig P.* Artificial Intelligence: A Modern Approach (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall, 2003.

12. *Schlosser M.* Agency // The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), URL= https://plato.stanford.edu/archives/win2019/ entries/agency/ (10.07.2024).

References

1. *Gavrilina, E.A.* (2023). Agentnost ne-chelovekov: vzaimodeistvie lyudei i sotsialnykh robotov [The agency of non-humans: interaction between humans and social robots]. Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsialnye peremeny [Public Opinion Monitoring: Economic and Social Change], 3. https://doi.org/10.14515/monitoring. 2023.3.2318.

2. *Barandiaran, X.E., E. Di Paolo, and M. Rohde*. (2009). Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action. *Adaptive Behavior*, 17(5), 367–386.

3. *Benossi, L. and Bernecker, S*. (2022). 5 A Kantian Perspective on Robot Ethics. In: H. Kim and D. Schönecker (eds). *Kant and Artificial Intelligence*. Berlin, Boston: De Gruyter, 2022, pp. 145–168. 10.1515/9783110706611-005.

4. *Calo, C.J., Hunt-Bull, N., Lewis, L., Metzler, L.* (2011). Ethical implications of using the Paro robot with a focus on dementia patient care. In *Proceedings of the 12th AAAI Conference on Human-Robot Interaction in Elder Care* (AAAIWS'11-12). AAAI Press, 20–24.

5. *Chen, N., Zhai, Y. & Liu, X*. (2022). The Effects of Robots' Altruistic Behaviours and Reciprocity on Human-robot Trust. *Int J of Soc Robotics*, 14, 1913–1931. 10.1007/s12369-022-00899-6.

6. *Dennett, D.C.* (1987). *The Intentional Stance.* Cambridge, MA: MIT Press.

7. *Fonq, T.* & *Nourbakhsh, I.* & *Dautenhahn, K*. (2003). A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems*, 42, 143–166. 10.1016/S0921-8890(02)00372-X.

8. *Fox, J., Gambino, A.* (2021). Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human-Robot Interaction. *Cyberpsychol Behav Soc Netw*. 2021 May. 24(5), 294–299. 10.1089/cyber.2020.0181. Epub 2021 Jan 11. PMID: 33434097.

9. *Horstmann, A.C., Krämer, N.C*. (2022). The Fundamental Attribution Error in Human-Robot Interaction: An Experimental Investigation on Attributing Responsibility to a Social Robot for Its Pre-Programmed Behavior. *Int J of Soc Robotics*, 14, 1137–1153. 10.1007/s12369-021-00856-9.

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

11. *Russell, S.J.; Norvig, P.* (2003). *Artificial Intelligence: A Modern Approach (2nd ed.).* Upper Saddle River, New Jersey: Prentice Hall.

12. *Schlosser, M.*, (2019). Agency. *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2019/entries/agency/ (10.07.2024).

Информация 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, Siberian Branch of the Russian Academy of Sciences (8, Nikolaev st., Novosibirsk, 630090, Russia).

Дата поступления: 04.11.2024