proceedings

MDPI

Proceeding Paper

Point and Network Notions of Artificial Intelligence Agency †

Marcin Rabiza

Institute of Philosophy and Sociology, Polish Academy of Sciences, 00-330 Warsaw, Poland; marcin.rabiza@gmail.com

† Presented at Philosophy and Computing Conference, IS4SI Summit 2021, Online, 12–19 September 2021.

Abstract: As intelligent machines are more and more present in our environment, the interest of researchers in the problem of AI (artificial intelligence) agency is growing significantly. In light of this, this paper aims to examine the dominant trends in AI agency research regarding their philosophical implications, as well as to provide a research commentary. Recurring themes of point and network notions of agency are identified, based on which the argument for the dual-process nature of agency perception is presented. Emphasis is placed on the agency attribution phenomenon. The novel hypothesis of a negative correlation between the perceived agency of AI and its interpretability outlines the future research direction.

Keywords: artificial intelligence; perceived agency; agency attribution; agential stance

1. Introduction

The notion of agency has a long history of intellectual inquiry, occupying a special place in both philosophy of mind and action. While the traditional theory of agency captures the phenomenon as a distinctively human action, explainable in terms of the agent's desires, beliefs, and intentions, in more recent debates researchers shift their focus to non-human agents, such as animals or machines, exercising a similar capability to act, yet without having causally efficacious mental states [1]. Surely machine agency differs from that of humans. Yet we need an inclusive non-human agency conceptual framework, as without one we can "miss and misunderstand the massive changes in intelligent machine design and interactive media use that open up Pandora's box filled with thousands of agents" [2] (p. 62).

In recent years, research on machine agency has been particularly fruitful, as it has been stimulated by the rapid and widespread development of artificial intelligence (AI) technology. The reason for putting AI in the spotlight is its perceived similarity to human agents in terms of the capability to act in our environment. Thanks to recent advances in machine learning, the performance of AI systems (e.g., social robots, voice assistants, chatbots, or video game bots) is reaching the level where we no longer notice their instrumentality, but we begin to perceive their interactivity (cf. [2]). A quest for human-like "thinking machines" implies non-trivial reasons for perceiving AI systems not only as objects but also as *sui generis* subjects of action. Currently we are experiencing AI's shift from technical artifacts to artificial agents.

This issue is valid, not only for philosophy of action and psychology of agency attribution, but also for a broader debate on the social and ethical impact that technology has on humanity. AI has not only great industrial potential but also unprecedented transformative power. Mariarosaria Taddeo and Luciano Floridi call it "a powerful force that is reshaping daily practices, personal and professional interactions, and environments" [3] (p. 751). As we have reached a point where AI performs human-like actions that impact others in the world, research on AI agency has been more vivid than ever before.

check for updates

Citation: Rabiza, M. Point and Network Notions of Artificial Intelligence Agency. *Proceedings* **2022**, *81*, 18. https://doi.org/10.3390/proceedings2022081018

Academic Editor: Peter Boltuc

Published: 10 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Taking this into account, the paper aims to discuss the latest literature on artificial intelligence agency, and to propose two types of notions of AI agency along with the dual-process model of agency perception as a further research direction.

2. Point and Network Approaches

While many agency types can be distinguished in recent studies, they usually follow one of two general concepts: the point or the network notion of agency.

Point notions of agency are those capturing agency as a trait of an individual actor, explaining the phenomenon in terms of the agent's internal (functional) organization. In those accounts, an agency can be defined according to various attributes likely to be internal features of an agency-examined entity. Such definitions are many—too many to mention here. However, a few examples can help to paint the right picture.

In one of the most recent works, Danielle Swanepoel examines various accounts of agency to establish minimal criteria which an AI system must meet in order to be considered a genuine agent. Swanepoel suggests that there are four of these: deliberative self-reflection, awareness of self in time, critical awareness of environment, and norm violation [4]. Other exemplary studies identify criteria useful for machine agency conceptualizations such as individuality, interactional asymmetry (being a source of activity) and normativity [5], goal-oriented activity contributing to the agent's own endurance or maintenance [6], intentionality and forethought [7], adaptive regulation, self-reactiveness and self-reflectiveness [6–8], and first-person sense of agency [9,10].

One can observe the emergence of an increasing trend toward entirely separate accounts of agency in ongoing research, which I propose to call the network ones. Network notions capture agency not as a fixed essence or a property that something or someone possesses, but as an attribute of many actors' relationships. Such theories are focused on the flow of agency between machine and human actors, giving them equal weight.

Referring to the agency of the network instead of independent actors is typical for actor-network theory and new materialist philosophical traditions. Such an approach is especially relevant in theorizing AI agency as it "Helps to explain how agents interact with each other and allows an analysis of both artificial and non-artificial agents in the same context, avoiding the need to think in human/non-human barriers and ignoring the hierarchical distribution of actors," [11] (p. 8, cf. [12]). In a network agency-status analysis, a human-machine network (HMN) can be perceived as a set of decentralized interconnected nodes that share multiple relations with each other, where AI entities not only take part as passive tools but also actively change them, influencing each other in ways mostly yet unknown to us. However, as the boundaries between the physical, digital, and human factors of an HMN are blurred, one cannot observe the agency of a singular actor—it is displayed only as an emergent product of a process of human-machine interaction with many intertwined agential factors. With the ongoing progress in AI, interactions between humans and intelligent machines become increasingly indistinguishable, which supports the argument (cf. e.g., [13–15]).

A new materialist input for a network agency-status analysis is discovered within a Baradian agential realism framework, in which agency emerges in an intra-actional way [2,16]. An intra-action understands agency, not as an inherent property of an individual actor to be exercised, but as a dynamism of forces in which all observable objects are constantly exchanging and diffracting, influencing and working inseparably. Agency is primarily observed for the network, as "Relations do not follow relata, but the other way around," [16] (p. 136). This account acknowledges the impossibility of an absolute separation of an agential system's network—e.g., separation of a human user from an AI tool, as an AI agency cannot emerge without both sides being involved. There are emergent outcomes stemming from the process of humans and AI interacting, unraveling the intertwined nature of network agency (cf. e.g., [14,15,17,18]).

The network approach can be exploited even further when applied to an individual AI entity investigation. Speaking in Latourian terms, a stable alliance of actors so firmly

established that we can take its interior for granted is called a black box [19]. The internal properties of a black box do not count as long as we are concerned only with its input and output. One can similarly think of an AI agent as being a stable network of hidden-yet-entangled actors itself (as we even use the term 'black boxes' for opaque machine learning algorithms). An AI agent 'box' may include human designers, hardware and software architectures, machine learning algorithms, training and test data, and even its end-users, which are made invisible in AI's everyday use.

The network approaches for agency-status analyses seem more than promising, as they explore and transgress the boundaries of different categories of natural and artificial actors entangled in socio-material relations. This can be especially relevant in the perspective of AI advancements towards general intelligence (AGI) and the intelligence explosion (cf. e.g., [8]).

3. Dual-Process of Agency Perception

Many definitions of machine agency are bound to the ontological context, in which certain conditions are to be met in order to grant AI agential status. Recently, researchers have been more eager to admit that machine agency exists while differing from human agency, and is "Probably the result of human interaction and perception," [18] (p. 4). Taking this into account, a strong focus should be placed on an epistemological perspective of the AI agency-status problem. This is particularly promising for human-computer interaction (HCI) as well as usability and user experience (UX) studies, as AI systems exert an increasing level of influence on both human and machine actors.

Agency in 'thinking machines' is then mostly perceived [14,15,18,20–26] and attributed [27–32] during HCI. The latter means ascribing agential status (also called "agency judgment" [28]) to artificial actors based on the perception of AI action-outcome contiguity and causality that triggers a sense of external agency. This claim can be further supported by empirical pieces of evidence pointing to a human tendency towards both mindful and mindless anthropomorphism in artificial entities [15,28,33–35].

Agency attribution is triggered by a mental mechanism I propose to call the agential stance, which extends the notion of intentional binding (e.g., [36,37]). Daniel Dennett identifies three basic mental strategies for predicting and explaining the behavior of external-world objects he calls 'stances': the physical stance, the design stance, and the intentional stance [38]. Similar to the intentional stance, the agential stance can be regarded as instrumentally rational heuristics for predicting, explaining, and creating stable interpretations of external world phenomena by a mindful—as well as a mindless—attribution of agency (cf. cognitive miser theory, e.g., [39], pp. 65–71). The agential stance does not assume intentionality (or any other perceived mental properties), but rather action-outcome contiguity and intentional binding that trigger a sense of external action ownership in one who is observing actions similar to one's own.

Philosophical research can provide us with many conceptual tools that are useful for AI agency-status examination and that fall into either the point or network category. What I further propose is that the point-network theoretical distinction follows the dual-process nature of AI agency perception in humans. The two types of thinking involved in human cognition and reasoning of machine agency can be explained in a similar way to Daniel Kahneman's System 1 and System 2 processing [40]. An argument can be raised that it is the AI system interpretability which has a major (but not exclusive) impact on HCI and triggers point-type (Type 1) or network-type (Type 2) AI agency-status perceptions.

The final hypothesis I would like to pose regards the existence of a negative correlation between the perceived agency of AI systems and their interpretability. The more the AI system is opaque and hides its inner workings as a nontransparent and poorly explainable black box, the more likely it is that a human user will adopt Type 1 point thinking along with the agential stance. On the other hand, the more AI systems implement explainability, transparency, and interpretability showing their inner working, the more likely it is that a

human user will turn to Type 2 network analysis by trying to figure out the role of the AI system in a bigger HMN and the role of its parts.

The proposed distinction fosters research in philosophy of mind and folk psychology of attribution, but its implications may also turn out to be relevant for HCI and UX domains. The two types of agency perceptions may yield differing and sometimes conflicting results. Spontaneous agency attribution associated with the agential stance may improve overall user experience and trigger social reactions when interacting with AI [14,15,23,26]. The network thinking may impede agency attribution and result in "opening the black box". These claims are, however, outside the scope of this paper.

4. Conclusions

The aim of this paper was to discuss the latest literature on artificial intelligence agency, as well as to identify its dominant trends and provide a research commentary. The review showed that theories of AI agency often follow one of two general concepts: the point or the network notion of agency. The former captures agency as a capacity exercised by individual actors, while the latter focuses on various actors' (human and machine) relations as primarily agential and observable structures. Many similarities to accounts of actor-network theory and new materialist philosophical traditions are shown.

AI agency is explicated as a property perceived and attributed by humans. Agency attribution can be viewed as a mental strategy for interpreting, predicting, and explaining the behavior of AI agents that I propose to call the agential stance. Since we exercise a similar stance towards human actors, agency attribution may involve mindless anthropomorphism and trigger point-type perceptions of the AI agency-status.

An argument can be raised that the point-network theoretical distinction follows the actual dual-process nature of AI agency perception in humans. AI interpretability may affect the attribution process and determine the type of perception. A working hypothesis on the negative correlation between the perceived agency of AI systems and the level of their interpretability should be examined in a further empirical investigation.

While the presented account is yet to be fully explicated, the preliminary research suggests that the point and network dual system of the AI agency-status perception conceptual framework may provide a significant input for further research in not only the philosophy of action but also folk psychology of attribution during human-computer interactions, as well as for AI user experience studies. Even if a strong ontological rationale behind granting AI agential status cannot be reconstructed, there may be some instrumentally rational reasons for attributing agency. This makes perceived agency epistemically relevant for successful and satisfying HCI.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I would like to thank Peter (Piotr) Boltuc for his words of encouragement.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Brooks, R.A. Intelligence without representation. *Artif. Intell.* **1991**, *47*, 139–159. [CrossRef]
2. Rammert, W. Where the Action is: Distributed Agency between Humans, Machines, and Programs. In *Paradoxes of Interactivity: Perspectives for Media Theory, Human-Computer Interaction, and Artistic Investigations*; Seifert, U., Kim, J.H., Moore, A., Eds.; Transcript Verlag: Bielefeld, Germany, 2015; pp. 62–91.
3. Taddeo, M.; Floridi, L. How AI Can Be a Force for Good. *Science* **2018**, *361*, 751–752. [CrossRef]
4. Swanepoel, D. Does Artificial Intelligence Have Agency. In *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts*; Studies in Mind and, Brain; Clowes, R., Gartner, K., Hipólito, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; pp. 83–104.

5. Barandiaran, X.; Di Paolo, E.; Rohde, M. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality inAction. *Adapt. Behav.* **2009**, *17*, 367–386. [CrossRef]
6. Moreno, A.; Etxeberria, A. Agency in Natural and Artificial Systems. *Artif. Life* **2005**, *11*, 161–175. [CrossRef]
7. Bandura, A. Social Cognitive Theory: An Agentic Perspective. *Annu. Rev. Psychol.* **2001**, *52*, 1–26. [CrossRef]
8. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*, 1st ed.; Oxford University Press, Inc.: New York, NY, USA, 2014.
9. Chambon, V.; Sidarus, N.; Haggard, P. From action intentions to action effects: How does the sense of agency come about? *Front. Hum. Neurosci.* **2014**, *8*, 320. [CrossRef]
10. Legaspi, R.; He, Z.; Toyoizumi, T. Synthetic agency: Sense of agency in artificial intelligence. *Curr. Opin. Behav. Sci.* **2019**, *29*, 84–90. [CrossRef]
11. Van Rijmenam, M.; Logue, D. Revising the 'science of the organisation': Theorizing AI agency and actorhood. *Innov. Organ. Manag.* **2020**, *23*, 127–144. [CrossRef]
12. Latour, B. *Reassembling the Social: An Introduction to the Actor-Network Theory*; Oxford University Press: Oxford, UK, 2005.
13. Nass, C.; Steuer, J.; Tauber, E.R. Computers Are Social Actors. In Proceedings of the CHI '94: SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 24–28 April 1994; Association for Computing Machinery: New York, NY, USA, 1994; pp. 72–78.
14. Appel, J.; von der Pütten, A.; Krämer, N.C.; Gratch, J. Does Humanity Matter? Analyzing the Importance of Social Cues and Perceived Agency of a Computer System for the Emergence of Social Reactions during Human-Computer Interaction. *Adv. Hum. -Comput. Interact.* **2012**, *2012*, 13. [CrossRef]
15. Araujo, T.B. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Comput. Hum. Behav.* **2018**, *85*, 183–189. [CrossRef]
16. Barad, K.M. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, 2nd ed.; Duke University Press: Durham, UK; London, UK, 2007.
17. Rose, J.; Jones, M. The Double Dance of Agency: A Socio-Theoretic Account of How Machines and Humans Interact. *Syst. Signs Act.* **2005**, *1*, 19–37.
18. Engen, V.; Pickering, J.B.; Walland, P. Machine Agency in Human-Machine Networks; Impacts and Trust Implications. In *Human-Computer Interaction. Novel User Experiences, Proceedings of the 18th International Conference, HCI International 2016, Toronto, ON, Canada, 17–22 July 2016*; Kurosu, M., Ed.; Lecture Notes in Computer Science ð 9733; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 96–106.
19. Harman, G. *Prince of Networks: Bruno Latour and Metaphysics*; re.press: Melbourne, Australia, 2009.
20. Rose, J.; Truex, D.P. Machine Agency as Perceived Autonomy: An Action Perspective. In Proceedings of the IFIP TC9 WG9.3 International Conference on Home Oriented Informatics and Telematics: Information, Technology and Society, Aalborg, Denmark, 9–11 June 2000; pp. 371–390.
21. Araujo, T.; Helberger, N.; Kruikemeier, S.; de Vreese, C. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI Soc.* **2020**, *35*, 611–623. [CrossRef]
22. Lucas, G.M.; Krämer, N.; Peters, C.; Taesch, L.S.; Mell, J.; Gratch, J. Effects of Perceived Agency and Message Tone in Responding to a Virtual Personal Trainer. In Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA '18, Sydney, Australia, 5–8 November 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 247–254.
23. Banks, J. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Comput. Hum. Behav.* **2019**, *90*, 363–371. [CrossRef]
24. Silva, J. Increasing Perceived Agency in Human-AI Interactions: Learnings from Piloting a Voice User Interface with Drivers on Uber. *Ethnogr. Prax. Ind. Conf. Proc.* **2019**, *2019*, 441–456. [CrossRef]
25. Jackson, R.; Williams, T. On Perceived Social and Moral Agency in Natural Language Capable Robots. In *2019 HRI Workshop on the Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI*; HRI Workshop: Daegu, Korea, 2020.
26. Cowley, S.; Gahrn-Andersen, R. Drones, robots and perceived autonomy: Implications for living human beings. *AI Soc.* **2021**, 1–4. [CrossRef]
27. McEneaney, J.E. Agency Attribution in Human-Computer Interaction. In *Engineering Psychology and Cognitive Ergonomics*; Harris, D., Ed.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 81–90.
28. Nomura, O.; Ogata, T.; Miyake, Y. Illusory agency attribution to others performing actions similar to one's own. *Sci. Rep.* **2019**, *9*, 10754. [CrossRef]
29. Zafari, S.; Koeszegi, S.T. Attitudes Toward Attributed Agency: Role of Perceived Control. *Int. J. Soc. Robot.* **2020**, *13*, 2071–2080. [CrossRef]
30. Ciardo, F.; Beyer, F.; De Tommaso, D.; Wykowska, A. Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition* **2020**, *194*, 104109. [CrossRef]
31. Morewedge, C. Negativity Bias in Attribution of External Agency. *J. Exp. Psychol. Gen.* **2009**, *138*, 535–545. [CrossRef]
32. Farrer, C.; Frith, C. Experiencing Oneself vs. Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency. *NeuroImage* **2002**, *15*, 596–603. [CrossRef]
33. Nass, C.I.; Lombard, M.; Henriksen, L.; Steuer, J. Anthropocentrism and computers. *Behav. Inf. Technol.* **1995**, *14*, 229–238. [CrossRef]

34. Nowak, K.L.; Biocca, F. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and S cial Presence in Virtual Environments. *Presence* **2003**, *12*, 481–494. [CrossRef]
35. Kim, Y.; Sundar, S.S. Anthropomorphism of computers: Is it mindful or mindless? *Comput. Hum. Behav.* **2012**, *28*, 241–250. [CrossRef]
36. Obhi, S.S.; Hall, P. Sense of agency in joint action: Influence of human and computer co-actors. *Exp. Brain Res.* **2011**, *211*, 663–670. [CrossRef]
37. Moore, J.W.; Obhi, S.S. Intentional binding and the sense of agency: A review. *Conscious Cogn.* **2012**, *21*, 546–561. [CrossRef]
38. Dennett, D.C. *The Intentional Stance*, 1st ed.; MIT Press: Cambridge, MA, USA, 1987.
39. Stanovich, K.E. The cognitive miser and focal bias. In *Rationality and the Reflective Mind*; Oxford University Press: New York, NY, USA, 2011; pp. 65–71.
40. Kahneman, D. *Thinking, Fast and Slow*, 1st ed.; Farrar, Straus and Giroux: New York, NY, USA, 2011.