FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Marcin Rabiza

DUAL-PROCESS APPROACH TO THE PROBLEM OF ARTIFICIAL INTELLIGENCE AGENCY PERCEPTION

doi: 10.37240/FiN.2022.10.zs.15

ABSTRACT

Thanks to advances in machine learning in recent years the ability of AI agents to act independently of human oversight, respond to their environment, and interact with other machines has significantly increased, and is one step closer to human-like performance. For this reason, we can observe contemporary researchers' efforts towards modeling agency in artificial systems. In this light, the aim of this paper is to develop a dual-process approach to the problem of AI agency perception, and to discuss possible triggers of various agency perceptions. The article discusses the agency attribution phenomenon, based on which the argument for the dual-process nature of agency perception is developed. Two distinct types of thinking (processing) involved in human reasoning on AI agency are suggested: Type 1 and Type 2. The first one is fast, automatic, routine, and often unconscious; the second is a slower, controlled, more conscious one. These two distinct types of processing can yield differing and sometimes conflicting results for human cognition and interaction. The preliminary philosophical findings may contribute to further investigations in philosophy of mind or cognitive psychology and could also be empirically tested in HCI and UX studies.

Keywords: artificial intelligence; perceived agency; agency attribution.

1. INTRODUCTION

Intentional human action has always been distinguished from machine operation, which is usually described as a repetitive and pre-programmed activity. However, if we continue to define action by the demanding features of intentions, desires, beliefs, and mental capabilities that are typical for humans (cf. Brooks, 1991), we could "miss and misunderstand the massive changes in the intelligent machine design and interactive media use that open up Pandora's box filled with thousands of agents" (Rammert, 2015, p. 62). Artificial intelligence (AI) agents differ from human ones, but they also differ from classical machines. Thanks to recent advances in machine learning and data science, the ability of AI agents to act independently of human oversight, respond to their environment, and interact with other machines has significantly increased, and is now one step closer to human-like performance. For this reason, we can observe contemporary researchers' efforts in defining and modeling agency in artificial systems.

The rapid development of artificial intelligence technology that started at the end of the twentieth century and continues until now, has led to an explosion of various definitions of agency. According to various authors, an agent is "A system that can act on its own behalf in an environment" (Kauffman, 2000, p. 8), "A system that tries to fulfill a set of goals in a complex, dynamic environment" (Maes, 1994, p. 136), "A system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (Franklin, Graesser, 1996, p. 25), "Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" (Russell, Norvig, 1995, p. 33), "[An] embodied system [that pursues] internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated" (Beer, 1995, p. 173), "[A] system that can initiate, sustain, and maintain an ongoing and continuous interaction with their environment as an essential part of their normal functioning" (Smithers, 1995, p. 97), etc.

One can claim, however, that such definitions are incomplete, as they strongly rely on intuitive or commonsense notions such as "acting on one's own," "pursuing one's agenda," or "being in continuous long-term interaction." As such, they leave a significant room for subjective, individual interpretation. A good definition should be able to capture the meaning of the term as used intuitively in science and everyday life, but at the same time, it should be followed by an operational and precise conceptualization of the term. In some cases, we cannot establish whether a system is a genuine agent judging solely by its behavior because others can have a different interpretation of it. Researchers try to mitigate that effect to some extent by enlisting and discussing necessary and sufficient conditions that are to be met to classify a system as a genuine agent. Conditions, that are introduced with precise, yet understandable terms.

For instance, Barandiaran, Di Paolo, and Rohde attempt to start with a simple and non-controversial description of an agent. According to them, our common, minimal understanding of agenthood relates to "A system doing something by itself according to certain goals or norms within a specific environment" (Barandiaran et al., 2009, p. 369). Then, the authors discuss three necessary conditions that are hidden within their definition: individuality, interactional asymmetry, and normativity. *Individuality* means that a system should be distinguishable from its environment. *Interactional asymmetry* refers to agents' ability to be the source of the activity (or "modulations") towards the environment. The third condition, *normativity*, points out that the agent's interactions with the environment are not random or arbitrary, but are done following goals or norms, which provide some sort of reference conditions for this activity.

It seems there is still no consensus on what such a minimal set of conditions should comprise, and new attempts are presented frequently. Thinking about the Barandiaran et al. approach, one may argue it is impossible, or at least is arbitrary, to decouple an agent from its environment, as well as to point out a single source of asymmetrical interaction with no doubt. This reasoning is discovered, for example, in (Rammert, 2015), where the author conceptualizes machine agency as fragmented in many pieces and delegated to myriads of pro-active and cooperative sub-systems showing low-level agency themselves, that are invisible in our everyday interaction, as they are strongly coupled in linear, sequential, or aggregated ways, into opaque yet functional black-boxes. In (Rose, Jones, 2005), the authors introduce the notion of the double dance of agency, claiming that human and machine operation outcomes are not determined by either, but emergent from the process of their interaction. As interactions between humans and intelligent machines and other systems become nowadays increasingly indistinguishable, it is difficult to establish whether a system is an individual, homogenous source of activity in its environment in order to be considered a genuine agent (cf. e.g. Nass et al., 1994; Appel et al., 2012; Araujo, 2018). In human-machine networks (HMNs), agency of the individual parts may be not *prima facie* accessible, and significant cognitive work must be done to decouple relational structure, looking for individual agents in a kind of *ex-post* rationalization attempt.

Considering a range of approaches to defining AI agency, I have recently developed two meta-concepts aggregating other accounts, which are *point* and *network* notions of agency (Rabiza, 2022). By the former, I mean notions that describe conditions for agenthood related to the agent's internal, functional organization. Point notions define AI agency according to various attributes, such as those mentioned by Barandiaran et al. The latter capture agency "not as a fixed essence or a property that something or someone possesses, but as an attribute of many actors' relationships" (Rabiza, 2022, p. 3). Here, we can classify all models attempting to explain machine agency as an emergent product of a process of human-machine interaction with many intertwined and mediated "agentive participants" involved. Furthermore, I suggested that the point-network theoretical distinction follows the dual-process nature of AI agency perception in humans (Rabiza, 2022, 3). In this paper, my aim is to further develop a dual-process approach to the problem of AI agency perception, and to discuss possible triggers of various agency perceptions.

2. DUAL-PROCESS AGENCY PERCEPTION

Researchers tend to admit that machine agency exists, although it differs from that of humans, and is "probably the result of human interaction and perception" (Engen et al. 2016, p. 4). Taking this into account, my focus is on an epistemological perspective of the AI agency perception problem. According to conducted literature research, artificial intelligence is foremostly *perceived* (e.g. Appel et al., 2012; Araujo, 2018; Araujo et al., 2020; Banks 2019; Engen et al., 2016; Jackson, Williams, 2020; Lucas et al., 2018; Rose, Truex, 2000; Silva, 2019) and *attributed* (e.g. Ciardo et al., 2020; Förster, Althoefer, 2021; McEneaney, 2009; Moon, Nass, 1998; Rose, Jones, 2005; Zafari, Koeszegi, 2020) during a human-computer interaction (HCI).[1] This brings us to the idea that agency attribution may be related dual-process nature of AI agency perception in humans.

Various dual-process and dual-system theories have become popular in psychology and cognitive science research (cf. an overview of such theories in Frankish, 2010). Arguably, the most well-known categories in the field are Daniel Kahneman's two systems of the mind, labeled "System 1" ("automatic system") and "System 2" ("effortful system"):

> "System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration." (Kahneman, 2011, p. 22)

A form of somewhat similar dual-process theory may apply to the problem of AI agency perception. I argue that there are two distinct types of thinking (processing) involved in human reasoning on AI agency: Type 1 and Type 2. The first one is fast, automatic, routine, and often unconscious; the second is a slower, controlled, more conscious one. These two distinct types of processing can yield differing and sometimes conflicting results for human cognition and interaction.

Type 1 processing triggers a point-type agency perception during HCI. Attempting to interpret intelligent machines' agent-like behavior entails making attributions about possible causal relationships and mechanisms governing their operation. In the Type 1 mode of thinking, we are likely to attribute agential status to a technical artifact thanks to a mental mecha-

[1] "Attribution" here means a process ascribing agential status (also called an "agency judgment," (Nomura et al., 2019) to artificial actors based on the perception of AI action-outcome contiguity and causality that triggers a sense of external agency. Claims on the phenomenon of attributing human-typical qualities to machines are widely supported by empirical pieces of evidence pointing to a human tendency towards both mindful and mindless anthropomorphism of artificial entities (e.g., Kim, Sundar, 2012; Nass et al., 1995; Nowak, Biocca, 2003).

nism I propose naming *the agential stance.*[2] Daniel Dennett identifies three basic mental strategies to explain the behavior of external-world objects, which he calls "stances:" the physical stance, the design stance, and the intentional stance (Dennett, 1987). The intentional stance focuses on a reason-giving explanation of an action in terms of assumed mental properties of an agent—namely that an agent has certain desires and beliefs and acts rationally towards their completion. Indeed, studies seem to indicate that humans to some extent tend to adopt an intentional stance towards artificial agents and give mentalistic explanations of their actions (Marchesi et al., 2019). The proposed agential stance can be regarded as yet more folk-psychological, instrumentally rational heuristics for predicting, explaining, and generating quasi-stable interpretations of the external world phenomena, by a mindful as well as a mindless attribution of agency (agential status assumption) to its objects[3]. Accordingly, a human user in the agential stance tends to perceive AI agents as genuine, even human-like agents, because agency attribution makes an overall HCI easier and less effortful, improving the overall first-person user experience.

One may assume Type 1 processing works spontaneously and automatically in HCIs involving AI agents and user interfaces designed for smooth social interactions, such as social robots or socially intelligent agents (SIAs) (e.g., Dautenhahn, 1998, Dautenhahn et al., 2006; Persson et al., 2001) or affective social robots (e.g. Sengers, 2002; cf. Marchesi et al., 2019). It is less likely that Type 1 thinking will occur equally in interactions with machines devoid of similar anthropomorphic cues or a human-centered design.

In those circumstances, when agency attribution towards an artificial agent is too cognitively demanding for a human user, Type 2 processing would be expected to occur. This mode of thinking triggers a different, network-type agency perception during HCI. My argument here is that, as the human mind tends to interpret and model the behavior of external-world objects for cognitive reasons, to rationalize external agency, we automatically perceive AI agents as entities entangled in stable alliances of many actors

[2] Agency attribution in the agential stance should not be confused with the similar category of intentional binding effect, which refers to the subjective experience of the compressed temporal interval between voluntary action and its external sensory consequence (as a result action and outcome are perceived as being closer together in time) and is sometimes used as in an implicit measure of the first-person sense of agency (SoA) (e.g., Desantis et al., 2012; Moore, Obhi, 2012; Obhi, Hall, 2011; Suzuki et al., 2019). Agential stance, on the other hand, describes a folk-psychological mechanism of attribution based on the perception of AI action-outcome contiguity and causality, triggering the sense of external action ownership (third-person SoA) in the one observing an action similar to one's own. Thus, if intentional binding is experienced in another agent's action, it might contribute to agency attribution.

[3] This approach is in line with the cognitive miser theory in psychology, claiming that a human mind has a natural tendency to avoid spending too much of a cognitive effort and simplify the thinking process whenever possible (Stanovich, 2009; 2011). As Fiske and Taylor put it: "People are limited in their capacity to process information, so they take shortcuts whenever they can" (Fiske, Taylor 1991, p. 41).

involved in the human-machine network.[4] In the agency-status analysis, a human observer looks for both a net of external relations around the agent and its internal structure. These agential powers may comprise mediated "agentive participants" such as human designers, hardware and software architectures, algorithms, datasets, and end-users with their material practices, as well as relations with other machines and human agents that are invisible in Type 1 "black-boxed" perceptions.[5] Type 2 processing is therefore a more complex and controlled way of reasoning on external action seeking the consistency and quasi-stability that comes from a coherent view of a larger "agential structure." It consists of more detailed and nuanced thought processes focused on decomposing, deconstructing, measuring, and analyzing the relational nature of HMN agency with more of a scientific or theoretic (albeit in a naïve sense) approach (cf. Crisp, Turner, 2014).

3. POSSIBLE TRIGGERS OF VARIOUS AGENCY PERCEPTIONS

The immediate question of "What can trigger this range of AI agency perceptions?" is vast and remains mostly unanswered in this paper. There may be many (or very few) potential factors influencing the transition between Type 1 and Type 2 thinking about AI agency, resulting in point and network types of perceptions. An empirical study could outline a possible direction for further research. Instead, I would like to pose a hypothesis about one of the potential breaking points in the perception of AI agency that is rooted in the principles of AI design.

In *Being and Time*, Martin Heidegger differentiates two phenomenological modes of tool-being that are constituted through Dasein's varying attitudes toward objects in the world: "presence-at-hand" and "readiness-at-hand" (Heidegger, 1962). As Graham Harman puts it:

> "The latter term, ready-to-hand, refers to equipment that remains concealed from view insofar as it functions effectively. Present-at-hand, the opposite term, refers to at least three different sorts of situations. In Heidegger's writings objects present in consciousness are called present-at-hand, and so are 'broken tools' that become obtrusive once they no longer function effectively, and so is the physical concept of objective matter occupying a distinct point in space-time.
>
> At any rate, present-at-hand and ready-to-hand are not two different types of entities. Instead, all entities oscillate between these two separate modes: the

[4] What I mean here, is that faced with no easy access to individual agency perception we tend to look for a bigger picture in order to make sense.

[5] If external action cannot be easily rationalized by assigning causative status to an object, interpretations based on network (relational, structural) characteristics become an epistemically useful mean to an end.

> cryptic withdrawal of readiness-to-hand and the explicit accessibility of presence-at-hand." (Harman, 2019, pp. 18–19)

A tool is ready-at-hand when it is perceived as a handy piece of equipment to be used for achieving some goal. It consists of multiple parts (such as a hammer comprising a head, claw, or handle) but they are usually "hidden or withdrawn realities performing their labors unnoticed" (Harman, 2019, p. 18). Most frequently, we deal with tools within this kind of practical relation, taking them for granted as items of everyday use. The moment the tool is broken, however, it becomes present-at-hand. The tool then reveals its secrets to Dasein, who now perceives it in a more scientific or theoretic perspective, concerned only with the bare factuality of its constituent parts, regardless of its usefulness, and with no subjective context involved.

Per analogiam, a tool malfunction can be one of many potential triggers for a dual-process AI agency perception (or a machine agency perception in general). Another potential candidate closely related to AI design patterns could be AI *interpretability*.[6] A hypothesis I would like to pose concerns a possible 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 interpretable black-box (while still providing smooth interaction and good user experience), the more likely a human user will adopt Type 1 thinking along with point-type agency perceptions.

On the other hand, the more AI systems implement interpretability showing their mechanisms of operation (inner workings as well as outer relations), as more explainable without the need of a mentalistic approach to generate behavioral predictions ("agential stance"), the more likely a human user will turn to Type 2 processing and network-type perceptions, trying to rationalize the role of an AI system within a more complex HMN agentive structure.

Dual-processing in AI agency perception influenced by factors such as AI interpretability may sometimes yield conflicting results. Spontaneous agency attribution in Type 1 processing may improve the overall user experience and even trigger social reactions while interacting with AI (Appel et al., 2012; Araujo, 2018; Cowley, Gahrn-Andersen, 2021; Lucas et al., 2018). Controlled and rationalized network thinking may impede agency attribution and result in "opening the black box," and objectifying AI's agential potential.

[6] AI is interpretable when humans can easily understand the reasoning behind predictions and decision making of the model. The more interpretable and transparent the AI agent is, the easier it is for the user to comprehend it and trust it.

4. CONCLUSIONS

The aim of this article was to develop a dual-process approach to AI agency perception that was previously suggested in (Rabiza, 2022), and to discuss possible triggers of various agency perceptions.

I argue that there are two distinct types of thinking (processing) involved in human reasoning on AI agency: Type 1 and Type 2. The first is fast, automatic, routine, and often unconscious; the second is slower, controlled, and more conscious. These two distinct types of processing can yield differing and sometimes conflicting results for human cognition and interaction. Type 1 processing triggers a point-type agency perception during HCI. In the Type 1 mode of thinking, we are likely to attribute agential status to a technical artifact. However, when agency attributJion towards an artificial agent is too cognitively demanding for a human user, Type 2 processing is expected to occur. This mode of thinking triggers a different, network-type agency perception during HCI. It is a more complex way of reasoning on external action, seeking the consistency and stability that comes from a coherent view of a larger "agential structure."

Using an analogy of Heidegger's broken tool analysis I propose a hypothesis on a possible 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 interpretable blackbox, the more likely a human user will adopt Type 1 thinking along with point-type agency perception. The more AI systems implement interpretability showing their mechanisms of operation (inner workings as well as outer relations) as more explainable without the need of a mentalistic approach to generate behavioral predictions, the more likely a human user will turn to Type 2 processing and network-type perception, trying to rationalize the role of an AI system within a more complex HMN agentive structure.

The preliminary philosophical findings may contribute to further investigations in philosophy of mind or cognitive psychology and could also be empirically tested in HCI and UX studies.

REFERENCES

J. Appel, A. von der Pütten, N.C. Krämer, J. Gratch, *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*, Advances in Human-Computer Interaction, 2012, pp. 1–10.

T. B. Araujo, *Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions*, Computers in Human Behavior, 85, 2018, pp. 183–189.

T. Araujo, N. Helberger, S. Kruikemeier, C. de Vreese, *In AI we Trust? Perceptions about Automated Decision-making by Artificial Intelligence*, AI & Society, 35, 2020, pp. 611–623.

A. Bandura, *Social Cognitive Theory: An Agentic Perspectiv*e, Annual Review of Psychology, 52, 2001, pp. 1–26.

J. Banks, *A Perceived Moral Agency Scale: Development and Validation of a Metric for Humans and Social Machines*, Computers in Human Behavior, 90, 2019, pp. 363–371.

K. M. Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, 2nd ed., Duke University Press, Durham–London 2007.

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

R. D. Beer, *A Dynamical Systems Perspective on Agent-environment Interaction*, Artificial Intelligence, 72, 1995, pp. 173– 215.

R. A. Brooks, *Intelligence without representation,* Artificial Intelligence*,* 47, 1991, pp. 139–159.

V. Chambon, N. Sidarus, P. Haggard, *From Action Intentions to Action Effects: How Does the Sense of Agency Come about?*, Frontiers in Human Neuroscience, 8, 2014, p. 320.

F. Ciardo, F. Beyer, D. De Tommaso, A. Wykowska, *Attribution of Intentional Agency towards Robots Reduces One's Own Sense of Agency,* Cognition, 194, 2020.

S. J. Cowley, R. Gahrn-Andersen, *Drones, Robots and Perceived Autonomy: Implications for Living Human Beings,* AI & Society, 2021.

R. J. Crisp, R. N. Turner, *Essential Social Psychology*, 3rd ed., SAGE Publications, Thousand Oaks, 2014.

K. Dautenhahn, *The Art of Designing Socially Intelligent Agents: Science, Fiction, and the Human in the Loop,* Applied Artificial Intelligence, 12 (7–8), 1998, pp. 573–617.

K. Dautenhahn, A. Bond, L. Cañamero, B. Edmonds, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, in: Socially Intelligent Agents. Multiagent Systems, Artificial Societies, and Simulated Organizations, vol. 3, K. Dautenhahn, A. Bond, L. Cañamero, B. Edmonds (eds.), Springer, Boston 2002, pp. 1–20.

D. C. Dennett, *The Intentional Stance*, 1st ed., MIT Press, Cambridge–London 1987.

A. Desantis, G. Hughes, F. Waszak, *Intentional Binding Is Driven by the Mere Presence of an Action and Not by Motor Prediction*, PLoS One, 7 (1), 2012.

V. Engen, J. B. Pickering, P. Walland, *Machine Agency in Human-Machine Networks; Impacts and Trust Implications*, in: Human-Computer Interaction. Novel User Experiences, Proceedings of the 18th International Conference on Human-Computer Interaction, 2016.

S. T. Fiske, S. E. Taylor, *Social Cognition: From Brain to Culture*, 2nd ed., McGraw-Hill, New York 1991.

F. Förster, K. Althoefer, *Attribution of Autonomy and Its Role in Robotic Language Acquisition,* AI & Society, 2021.

K. Frankish, *Dual-process and Dual-system Theories of Reasoning*, Philosophy Compass, 5 (10), 2010, pp. 914–926.

S. Franklin, A. Graesser, *Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents,* in: Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages, Lecture notes in computer science, Springer, Berlin 1193, 1996, pp. 21–35.

R. Glanville, *Black Boxes*, Cybernetics & Human Knowing, 16, 2009, pp. 153–167.

G. Harman, *Prince of Networks: Bruno Latour and Metaphysics*, re.press, Melbourne 2009.

G. Harman, *Technology, Objects and Things in Heidegger*, Cambridge Journal of Economics, 34 (1), 2010, pp. 17–25.

M. Heidegger, *Being and Time*, J. Macquarrie, E. Robinson (trans.), Blackwell Publishing, Oxford 1962.

R. B. Jackson, T. Williams, *On Perceived Social and Moral Agency in Natural Language Capable Robot*s, in: 2019 HRI Workshop on The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI, 2020.

D. Kahneman, *Thinking, Fast and Slow*, 1st ed., Farrar, Straus and Giroux, New York 2011.

S. Kauffman, *Investigations*, Oxford University Press, Oxford 2000.

Y. Kim, S. S Sundar, *Anthropomorphism of Computers: Is It Mindful or Mindless?*, Computers in Human Behaviour, 28 (1), 2012, pp. 241–250.

B. Latour, *Reassembling the Social: An Introduction to the Actor-Network Theory*, Oxford University Press, New York 2005.

R. Legaspi, Z. He, T. Toyoizumi, *Synthetic Agency: Sense of Agency in Artificial Intelligence,* Current Opinion in Behavioral Sciences, 29, 2019, pp. 84–90.

G. M. Lucas, N. Krämer, C. Peters, L. S. Taesch, J. Mell, J. Gratch, *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; Association for Computing Machinery, New York, 2018, pp. 247–254.

P. Maes, *Modelling adaptive autonomous systems*, Artificial Life, 1, 1994, pp. 135–162.

S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, A. Wykowska, *Do We Adopt the Intentional Stance Toward Humanoid Robots?*, Frontiers in Psychology, 10, 2019.

J.E. McEneaney, *Agency Attribution in Human-Computer Interaction*, in: Engineering Psychology and Cognitive Ergonomics, D. Harris (ed.), Springer, Berlin–Heidelberg 2009, pp. 81–90.

D. Mcquillan, *Data Science as Machinic Neoplatonism*, Philosophy & Technology, 31, 2018, pp. 253–272.

Y. Moon, C. Nass, *Are Computers Scapegoats? Attributions of Responsibility in Human-computer Interaction*, International Journal of Human-Computer Interaction, 49 (1), 1998, pp. 79–94.

J. W. Moore, S.S. Obhi, *Intentional Binding and the Sense of Agency: A Review,* Consciousness and Cognition, 21 (1), 2012, pp. 546–561.

A. Moreno, A. Etxeberria, *Agency in Natural and Artificial Systems*, Artificial Life, 11 (1–2), 2005, pp. 161–175.

C. I. Nass, J. Steuer, E.R. Tauber, *Computers Are Social Actors*, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York 1994, pp. 72–78.

C. I. Nass, M. Lombard, L. Henriksen, J. Steuer, *Anthropocentrism and Computers*, Behaviour & Information Technology, 14, 1995, pp. 229–238.

O. Nomura, T. Ogata, Y. Miyake, *Illusory Agency Attribution to Others Performing Actions Similar to One's Own,* Scientific Reports, 9, 2019.

K. L. Nowak, F. Biocca, *The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments,* Presence: Teleoperators and Virtual Environments, 12 (5), 2003, pp. 481–494.

S. S. Obhi, P. Hall, *Sense of Agency in Joint Action: Influence of Human and Computer Co-actors*. Experimental Brain Research, 211, 2011, pp. 663–670.

P. Persson, J. Laaksolahti, P. Lönnqvist, *Understanding Socially Intelligent Agents — A multilayered Phenomenon,* IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 31 (5), 2001, pp. 349–360.

M. Rabiza, *Point and Network Notions of Artificial Intelligence Agency,* Proceedings, 81, 2022, p. 18.

W. Rammert, *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, U. Seifert, J.H. Kim, A. Moore (eds.), transcript Verlag, Bielefeld 2015, pp. 62–91.

J. Rose, M Jones, *The Double Dance of Agency: A Socio-Theoretic Account of How Machines and Humans Interact,* Systems, Signs and Actions, 1, 2005, pp. 19–37.

J. Rose, D.P. Truex, *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, Kluwer, 2000, pp. 371–390.

S. J. Russell, P. Norvig, *Artificial intelligence: A modern approach*, Englewood Cliffs, Prentice Hall, New York, 1995.

P. Sengers, R. Liesendahi, W. Magar, C. Seibert, B. Muller, T. Joachims, W. Geng, P. Martensson, K. Hook, *The Enigmatics of Affect. Anonymous*, in: Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, London 2002, pp. 87–98.

J. Silva, *Increasing Perceived Agency in Human-AI Interactions: Learnings from Piloting a Voice User Interface with Drivers on Uber*, in: Ethnographic Praxis in Industry Conference Proceedings, 2019, pp. 441–456.

T. Smithers, *Are Autonomous Agents Information Processing Systems?*, in: The Artificial Life Route to Artificial Intelligence: Building Situated Embodied Agents, L. Steels, R. A. Brooks (eds.), Erlbaum, New Haven 1995.

C. Speed, M. Disley, *Intra-actions in Data-driven Systems: A Case Study in Creative Praxis*, in: Distributed Perception: Resonances and Axiologies, N. Lushetich, I. Campbell (eds.), Routledge Studies in Science, Technology and Society, Routledge, 2021, forthcoming.

K. E. Stanovich, *The Cognitive Miser and Focal Bias*, in: Rationality and the Reflective Mind, Oxford University Press, New York 2011, pp. 65–71.

___ , *The Cognitive Miser: Ways to Avoid Thinking*, in: What Intelligence Tests Miss: the Psychology of Rational Thought, Yale University Press, New Haven 2009, pp. 70–85.

K. Suzuki, P. Lush, A.K. Seth, W. Roseboom, *Intentional Binding without Intentional Action,* Psychological Science, 30 (6), 2019, pp. 842–853.

D. Swanepoel, *Does Artificial Intelligence Have Agency?,* in: The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts, R. Clowes, K. Gartner, I. Hipólito (eds.), Studies in Mind and Brain, Springer, 2021, pp. 83–104.

M. Taddeo, L. Floridi, *How AI Can Be a Force for Good*, Science, 361, 2018, pp. 751–752.

M. van Rijmenam, D. Logue, *Revising the 'Science of the Organisation': Theorizing AI Agency and Actorhood*, Innovation: Organization & Management, 23, 2020, pp. 127–144.

S. Zafari, S. T. Koeszegi, *Attitudes Toward Attributed Agency: Role of Perceived Control*, International Journal of Social Robotics, 13 (5), 2020, pp. 2071–2080.

ABOUT THE AUTHOR — PhD student, Institute of Philosophy and Sociology, Polish Academy of Sciences, Nowy Świat 72, 00-330 Warsaw, Poland;

ORCID: https://orcid.org/ 0000-0001-6217-6149

Email: marcin.rabiza@gmail.com