REFLECTIONS

FILOZOFIA
Volume 79, Number 5, 2024

DOI: https://doi.org/10.31577/filozofia.2024.79.5.5

Reconsidering Agency in the Age of AI

GERHARD SCHREIBER, Faculty of Humanities and Social Sciences, Helmut Schmidt University/University of the Federal Armed Forces Hamburg, Germany

SCHREIBER, G.: Reconsidering Agency in the Age of AI
FILOZOFIA, 79, 2024, No 5, pp. 529 – 537

The expansive development of AI technologies challenges our conventional understanding of agency, which has traditionally been anchored in the human capacity for autonomous action and decision-making. As human-AI interactions become increasingly complex, the boundary between human and machine agency is continuously breached, prompting a reconsideration of the concept of agency as potentially no longer a human proprium. This paper offers preliminary reflections on the prerequisites and conditions for a post-anthropocentric theory of agency in the age of AI, beginning with a historical reconstruction and conceptual validation of the evolving notion of agency.

Keywords: AI – agency – human-AI interaction – autonomy – Actor-Network Theory

The expansive development in the field of AI technologies, with their disruptive potential for almost all areas of our work and life, challenges us to reconsider capabilities, skills, and competencies that have previously been attributed exclusively to humans and, albeit to a lesser extent, to non-human animals as well. This is particularly evident in the concept of agency. It is in their ability to act independently and thus influence a given environment that AI technologies consummately challenge our notions of what has traditionally been considered the exclusive domain of humans. In view of the rapid and unstoppable integration of AI systems even into everyday tasks and actions (Coeckelbergh 2020, 3f.), traditional, anthropocentrically conceived or influenced concepts of agency are proving increasingly inadequate. This calls for a reconsideration of agency that appropriately reflects and mediates both the capabilities of current and future AI systems and the changing role of humans in an AI-pervaded world.

Before outlining the essential prerequisites and conditions for such a post-anthropocentric agency theory in the "Golden Age of AI" (Kaynak 2021), this evolution of the agency concept will be historically reconstructed and conceptually validated.

I. Evolving Non-Human Agency

The inclusion of non-human beings and entities in the concept of agency marks a fundamental shift in the understanding of agency altogether. This extension of the concept of agency to animals and machines took place in different scientific-historical contexts, each with its own specific impulse. The targeted exploration of the "actancy" (*Handlungsträgerschaft*) (Wiedenmann 2020, 127) or "shaping capability" (*Gestaltungsbefähigung*) (Bossert 2022, 146) of animals in variants of *Animal Studies* was significantly influenced by *New Materialism,* aiming to overcome the categorical distinction between humans and animals as such (Borgards 2016). In parallel and largely independently, the concept of machine agency emerged with the development of automated systems capable of performing tasks without human intervention.[1]

Adopting the distinction between operational autonomy and behavioral autonomy[2] introduced by Ziemke (1998) for differentiating levels of autonomy, it can be said that in both Human-Animal Studies and Cultural Animal Studies, the recognition of *animal agency* was pursued from the outset with the aim of attributing a certain form of behavioral autonomy to animals, which at the same time marked a break with the traditional dichotomy between free action (of humans) and instinctual behavior (of animals). In the discourse between computer science and the humanities, meanwhile, the concept of *machine agency* initially referred primarily to a specific form of

[1] Automation refers to the ability of systems to perform predefined tasks (e.g. achieve the target) without human intervention, based on fixed rules or algorithms. Autonomy, on the other hand, implies greater independence. Autonomous systems do not act solely on the basis of fixed algorithms but can also make decisions and perform actions based on an analysis of their environment and state (Adler 2019). This also shows that autonomy, as it is understood in computer science, does not correspond to the understanding of autonomy as an ethical concept of freedom. While autonomy in computer science is often reduced to the ability of a system to make decisions based on programmed criteria and the analysis of environmental data, an ethical understanding of autonomy also includes aspects such as self-awareness, responsibility, and the ability to reflect on one's own actions, which makes a differentiated view of autonomy indispensable.

[2] While operational autonomy describes the ability to perform pre-defined tasks independently (i.e., without direct external control), behavioral autonomy describes the ability to control one's own behavior based on intrinsic motivation, internal states, or self-imposed goals and also to adapt to dynamic environments.

operational autonomy. It was only with developments in the fields of machine learning and cognitive AI that the possibility of a functional equivalence between machine behavior and the behavioral autonomy of human agents was considered. Although machine behavior is much simpler, even simplistic, compared to the notorious complexity of human behavior – "comparable to a child's sailboat in relation to a high-tech ship" (Misselhorn 2019, 42; translation mine) – the question now became, more broadly, whether and to what extent machines are also capable of self-initiated behavior. Behavioral autonomy presupposes operational autonomy, but not vice versa (Chantemargue 2002, 205f.).

II. Human and Machine Agency

The transition from simple, predictive, and rule-based approaches to complex, autonomous AI systems capable of responding independently and flexibly even to unpredictable situations and of adapting to dynamic environments poses the challenge of how to adequately conceptualize such an ability of systems not only to perform tasks that would otherwise require human intelligence, but also to exceed their original programming (Gaon 2021, 239) and the expectations directed at them – in human terms: to "learn" from "experience." This point of exceeding original programming and adapting to new challenges brings to the fore the debate initiated by Ada Lovelace's objection to the idea of machine creativity (Dormehl 2016, 185f.) and Alan Turing's indirect response that suggested the potential for machines to exhibit behavior that could not have been explicitly programmed (Bown 2021, 60). Given the real-world implications and practical consequences of such AI systems, which cannot be considered merely passive tools[3] nor "constrained solely by human-defined parameters" (Brandtzaeg 2023, 5), the discussion of machine agency is not only understandable but imperative. Perhaps the greatest challenge in formulating and communicating such an expanded concept of agency is to clarify that something is being attributed to non-human entities without taking anything away from humans. In any case, the methodological elevation of machines to agents does not necessarily imply the marginalization of humans, although they are no longer seen as the only autonomously acting subjects (Sundar 2020, 78).

[3] See Rammert (2008, 69) for a distinction between different degrees of machine agency: passive, semi-active, reactive, pro-active and co-operative.

Concerns, sometimes expressed in drastic terms, about "losing the human element" (Caspar 2023, 217; translation mine) or even "the death of humanity" (Hassan 2021, 232) through AI become virulent when human and machine agency are conceived as two opposing poles on the same level and thus placed in competition with each other. This view implies that the recognition of machine agency would undermine and diminish the (sovereignty of) autonomy of human agents. However, in opposition to this assumption and consequent efforts to limit machine agency and reclaim human agency (Sundar 2020, 77), it could be argued that the issue regarding the agency of human *and* non-human entities is not primarily a question of *distribution,* but rather one of *participation*.[4] It is not about levelling differences through uniformity ("humachine"), but about recognizing performative or functional-pragmatic equality despite ontological inequality. In other words, it is not a matter of relativizing the significance of being human, but of precisely defining the role – or the "part" – of human beings within a complex network of conditions and effects.

This includes avoiding the confusion of levels between human and machine and the conceptual ambiguity that results from anthropomorphizing technical processes and artifacts (for example, by analogizing neural networks to neural tissue, or comparing a television camera to a human eye), or informatizing human processes and conditions, "biofacts," as it were[5] (for example, by describing mental processes in analogy to computer-based processing, or comparing the human brain to a computer).[6] However, once agency is released from its anthropocentric embrace (Barad 2012, 54), it opens up the possibility of conceptualizing it in a way that can equally encompass human, non-human, and extra-human entities – including AI systems. A first step toward such an inclusive understanding of agency might be to consider agency in virtue of its effects, rather than focusing solely on actions themselves.

[4] This view of agency describes action "from the outside," so to speak. It is less a theory that explains action as such, but rather a perspective that attempts to describe and interpret action. In this respect, Hitzler – Knoblauch (2006, 3090) agree that agency is an *action-describing* concept rather than an *action-theoretical* concept.

[5] "Biofacts" (i.e. *given entities* [*Gegebenheiten*] that can be influenced by human intervention or technological processes) are understood here in direct analogy to "artifacts" (i.e. objects created by humans, *human-made entities* [*Gemachtheiten*]) and thus not as Karafyllis (2006) uses it to denote an ontologically intermediate realm of natural-artificial beings.

[6] On the scientific place and role of analogies in knowledge acquisition and problem solving, see Kirchartz (2023, 15 – 47), who shows that and to what extent analogies can serve not only as a method of logical reasoning, but also as a powerful heuristic tool for generating new hypotheses and tackling complex scientific problems.

III. From Actions to Effects

In an effect-oriented approach, agency is understood as the ability to exert influence on a given environment[7] and bring about effects. This greatly broadens the spectrum of who or what can have agency. Agency, then, is no longer defined by or confined to the ability to act intentionally but means "the ability to bring about effects" (Ahearn 2001, 113). However, agency is not just causality and should not be confused with simple cause-and-effect relationships since agency always implies *authorship* – i.e. being both cause *and* occasion. The crucial point is that authorship can be conceived both personally and non-personally, i.e. as analogous to a subject (Gerhardt 1996, 8). The condition for the possibility of agency is *power*, understood as "disposition to effects" (Gerhardt 1996, 10; translation mine), whereby both the possibilities of effects and the means of realizing them can vary considerably.

While agency in the traditional understanding is closely linked to intentionality, the effect-oriented approach goes further and recognizes that non-intentional processes, structures, and systems can also function as authors – in the sense of both sources and occasions – of effects that shape and change a given environment (Schreiber 2022, 201). This perspective marks a paradigm shift with far-reaching theoretical and practical implications, which are briefly outlined below.

IV. Preliminaries to an Integrated Theory of Agency

In an effect-oriented understanding of agency, which can be understood as a productive reception of Actor-Network Theory (ANT) under digital auspices, not only are individual and collective human actors recognized as carriers of agency, but also non-human, socio-material hybrid[8] entities such as AI

[7] "Exerting influence" is understood as a middle way between "influence" and "interference." While "influence" merely describes the determining effect or impact of something/someone on something/someone else, "exerting influence" more strongly expresses the exercise of influence, in which – not always, but also – a voluntary moment can manifest itself; "interference," on the other hand, more strongly emphasizes the manipulative character of an influence, excluding disorderly, unplanned, and spontaneous forms of such a determining effect. Influence, by its very nature, is neutral, neither inherently positive nor negative. However, it holds the potential to impact outcomes either positively or negatively to varying degrees.

[8] Regarding the idea of the socio-material hybridity of things, see Latour 2002, 70. The use of this terminology underlines that AI systems are composed not only of material components (such as hardware) and digital components (software), but are also embedded in social structures and processes. AI systems interact not only with human actors, influencing their decisions, behaviors, and social norms, but are also shaped by social practices, ethical considerations, and regulatory frameworks.

systems, which can likewise exert influence on a given environment and bring about effects. This shifts the focus to interaction and impact contexts in which AI systems no longer function as passive objects or mere instruments of externally set purposes, but as actants that can influence human behavior and thus also social, economic, and political processes. To adequately capture these complex impact structures of heterogeneous entities, a multidisciplinary approach is required that transcends the boundaries of established fields of research. Such an integrative theory of agency will therefore bring together perspectives from computer science, social sciences, philosophy, and technology ethics to achieve a holistic understanding of agency in the complex reality of human-AI interactions.

The development of such a theory first requires critical reflection on the vocabulary used in order to establish an interdisciplinary basis for communication and understanding. This includes adopting the term "actant" instead of "actor" so as to move away from anthropocentric thought structures and to capture equally the "agency" or "actancy" of both human and machine entities. Beyond terminological clarity as an essential prerequisite for conceptual clarity, formulating a common "grammar" to depict human-AI interactions also plays an essential part in theory formation. This grammar serves as a structuring framework that enables a coherent description of the complex relationships and interactions between human and machine actants and the systematic identification of differences and similarities.

Concurrently, it will be necessary to examine whether the interactions between humans and artificial intelligence simultaneously require moving beyond the traditional dyadic framework (human-machine) in favor of a triadic model to accurately represent the multifaceted relationships and interactions arising from the integration of AI systems into human activities.[9] Such a triadic model would not only consider the direct interactions between humans and machines, but also embrace the emergent properties and dynamics these interactions generate as an additional, distinct dimension. This further emphasizes the inadequacy of classical dualisms like subject/object, action/structure, or active/passive, autonomous/heteronomous in fully capturing the concept of agency in this evolving landscape.

[9] On the concept of triadic agency in the sense of a combination of causal agency (causal efficacy) and intentional agency (the capacity for intentional action), i.e., the combination of the contributions of users, designers, and artifacts to produce states of affairs see Johnson – Verdicchio (2019, 642f.).

The elaboration of such a post-anthropocentric theory of agency is undoubtedly a challenging endeavor, whose "effectiveness" must first and repeatedly be proven in practice and continuously critically reflected upon against the backdrop of new technological developments and changing social contexts. Ideally, a theory spelled out in this way would not only provide guidance for the ethically responsible development and implementation of AI systems, but also create a regulatory framework that promotes innovation and prevents misuse.

V. Shifting Boundaries

The history of AI is a story of constantly shifting boundaries. From the early beginnings, when AI was still a fascinating, almost nerdy concept within the scientific community, through the first successes in pattern recognition and language processing, to today's systems capable of solving complex problems, writing sophisticated texts, creating innovative works of art, and simulating human emotions, it is a continuous process of exploring the boundaries of what is possible – more precisely, of what is considered possible – and the human endeavor to exceed those boundaries. We find ourselves at the threshold of a new era of human-AI interaction in which the ontologically unbreachable boundary between human and machine appears to have been breached – in agency as a medium in which the divisiveness of human-machine differences is suspended. Therefore, ethical reflection on how we design and manage these interactions is indispensable to ensuring that they are conducted responsibly. What is being discussed is nothing less than a Copernican revolution, challenging the anthropocentrism of traditional concepts and ways of thinking and promoting a perspective in which human and machine capabilities will be not in opposition but complement each other.

Bibliography

ADLER, R. (2019): Autonomous or Merely Highly Automated – What is Actually the Difference? Available at: https://www.iese.fraunhofer.de/blog/autonomous-or-merely-highly-automated-what-is-actually-the-difference (visited 25.02.2024).

AHEARN, L. (2001): Language and Agency. *Annual Review of Anthropology*, 30, 109 – 137.

BARAD, K. (2012): Interview with Karen Barad. In: Dolphijn, R. – van der Tuin, I. (eds.): *New Materialism: Interviews and Cartographies*. Ann Arbor, Michigan: Open Humanities Press, 48 – 70.

BORGARDS, R. (2016): Einleitung: Cultural Animal Studies. In: Borgards, R. (ed.): *Animals: Handbook of Cultural Studies*. Berlin: J. B. Metzler, 1 – 5.

BOSSERT, L. (2022): *Gemeinsame Zukunft für Mensch und Tier: Tiere in der Nachhaltigen Entwicklung*. Freiburg: Alber.

BOWN, O. (2021): *Beyond the Creative Species: Making Machines That Make Art and Music*. Cambridge, Massachusetts: The MIT Press.

BRANDTZAEG, P. B. – YOU, Y. – WANG, X. – YUCONG, L. (2023): "Good" and "Bad" Machine Agency in the Context of Human-AI Communication: The Case of ChatGPT. In: Degen, H. – Ntoa, S. – Moallem, A. (eds.): *HCI International 2023 – Late Breaking Papers. 25th International Conference on Human-Computer Interaction, HCII 2023, Copenhagen, Denmark, July 23-28, 2023. Proceedings, Part VI*. Cham: Springer Nature Switzerland, 3 – 23.

CASPAR, J. (2023): *Wir Datensklaven: Wege aus der digitalen Ausbeutung*. Berlin: Econ.

CHANTEMARGUE, F. (2002): Conflicts in Collective Robotics. In: Tessier, C. – Chaudron, L. – Müller, H.-J. (eds.): *Conflicting Agents: Conflict Management in Multi-Agent Systems*. New York et al.: Kluwer, 203 – 220.

COECKELBERGH, M. (2020): *AI Ethics*. Cambridge, Massachusetts: The MIT Press.

DORMEHL, L. (2016): *Thinking Machines: The inside Story of Artificial Intelligence and Our Race to Build the Future*. London: WH Allen.

GAON, A. H. (2021): *The Future of Copyright in the Age of Artificial Intelligence*. Cheltenham and Northampton: Edward Elgar.

GERHARDT, V. (1996): *Vom Willen zur Macht: Anthropologie und Metaphysik der Macht am exemplarischen Fall Friedrich Nietzsches*. Berlin and New York: De Gruyter.

HASSAN, A. (2021): The Usage of Artificial Intelligence in New Media. In: Al-Sartawi, A. et al. (eds.): *Artificial Intelligence Systems and the Internet of Things in the Digital Era. Proceedings of EAMMIS 2021*. Cham: Springer Nature Switzerland, 229 – 240. DOI: https://doi.org/10.1007/978-3-030-77246-8_23

HELFFERICH, C. (2012): Von roten Heringen, Gräben und Brücken – Versuche einer Kartierung von Agency-Konzepten. In: Bethmann, S. – Helfferich, C. – Hoffmann, H. – Niermann, D. (eds.): *Agency: Qualitative Rekonstruktionen und gesellschaftstheoretische Bezüge von Handlungsmächtigkeit* (Reihe: *Edition Soziologie*). Weinheim: Juventa-Verlag, 9 – 39.

HITZLER, R. – KNOBLAUCH, H. (2008): Handlungsträgerschaft. In: Rehberg, K.-S. (ed.): *Die Natur der Gesellschaft: Verhandlungen des 33. Kongresses der Deutschen Gesellschaft für Soziologie in Kassel 2006*. Frankfurt am Main and New York: Campus, 3089 – 3090.

JOHNSON, D. G. – VERDICCHIO, M. (2019): AI, Agency and Responsibility: The VW Fraud Case and Beyond. *AI & SOCIETY*, 34, 639 – 647.

KARAFYLLIS, N. (2006): Biofakte – Grundlagen, Probleme, Perspektiven. *Erwägen Wissen Ethik*, 17 (4), 547 – 558.

KAYNAK, O. (2021): The golden age of artificial intelligence: Inaugural Editorial. *Discover Artificial Intelligence*, 1 (1). DOI: https://doi.org/10.1007/s44163-021-00009-x

KIRCHARTZ, M. (2023): *Riskantes Denken. Zur Funktion der Mensch-Maschine-Analogie in der Medienwissenschaft*. Bielefeld: transcript.

LATOUR, B. (2002): Zirkulierende Referenz. In: Latour, B. (ed.): *Die Hoffnung der Pandora: Untersuchungen zur Wirklichkeit*. Frankfurt: Suhrkamp, 36 – 95.

MISSELHORN, C. (2019): Maschinenethik und Philosophie. In: Bendel, O. (ed.): *Handbuch Maschinenethik*. Wiesbaden: Springer VS, 33 – 55.

RAMMERT, W. (2008). Where the Action Is: Distributed Agency between Humans, Machines, and Programs. In: Seifert, U. – Kim, J. H. – Moore, A. (eds.): *Paradoxes of Interactivity*. Bielefeld: transcript, 62 – 91.

SCHREIBER, G. (2022): Datentoxikalität. Eine technikethische Herausforderung. In: Augsberg, S. – Gehring, P. (eds.): *Datensouveränität: Positionen zur Debatte*. Frankfurt am Main and New York: Campus, 199 – 217.

SUNDAR, S. (2020): Rise of Machine Agency: A Framework for Studying the Psychology of Human-AI Interaction (HAII). *Journal of Computer-Mediated Communication*, 25 (1), 74 – 88. DOI: https://doi.org/10.1093/jcmc/zmz026

WIEDENMANN, R. (2020): Action-Theoretical Approaches to Human-Animal Sociality: A Comparative Sketch. In: Jaeger, F. (ed.): *Humans and animals. Foundations and Challenges of Human-Animal Studies*. Berlin: J.B. Metzler, 111 – 137.

ZIEMKE, T. (1998): Adaptive Behavior in Autonomous Agents. *Presence: Teleoperators and Virtual Environments*, 7 (6), 564 – 587.

Gerhard Schreiber
Faculty of Humanities and Social Sciences
Helmut Schmidt University/University of the Federal Armed Forces Hamburg
Holstenhofweg 85
22043 Hamburg
Germany
e-mail: schreiber@hsu-hh.de
ORCID ID: https://orcid.org/0000-0003-1178-1802