Philosophy in Informatics VIII

Frontiers of Artificial Intelligence — Philosophical Explorations

Organizers:

Commission on Philosophy of Science, Polish Academy of Arts and Sciences

Chair of History & Philosophy of Science, Pontifical University of John Paul II

online

Kraków, Dec 1-2nd, 2023

Short history of 'Philosophy in informatics' conferences

The first conference in the series took place at the Warsaw University of Technology in 2015; the second at the Faculty of Philosophy of the Pontifical University of John Paul II in Krakow in 2016; the third was organized by the Adam Mickiewicz University in Poznan in 2017; the fourth by the Faculty of Administration and Social Sciences of the Warsaw University of Technology in 2018, the fifth by the Maria Curie-Skłodowska University in Lublin in 2019. Sixth conference was organized jointly by Pontifical Univeristy of John Paul II in Kraków with Warsaw University of Technology in 2021. Seventh conference was organized in 2022 by Adam Mickiewicz University in Poznań.

About the conference

The conference explores the philosophical problems at the frontiers of Artificial Intelligence (AI). The assumption behind the conference is that AI technology itself cannot explain what AI technology does and can accomplish, why AI systems do what they do and not do and why they cannot (so far) move beyond a certain class of problems. The claim is that to progress, AI needs to understand the philosophical background of the problems it attempts to model. Thus, AI needs philosophy; the question "What philosophy can do for AI?" may be another title of this conference.

We are proud to invite you to the eighth edition of our "Philosophy in Informatics" conference. This edition of the event will take place on 1-2 December, in a virtual form. The main organiser is Commision for Philosophy of Science of Polish Academy of Arts and Sciences in Kraków and Chair of History and Philosophy of Science of Pontifical University of John Paul II in Kraków.

The possible topics include but are not limited to:

- Roads to AGI – is there one or many or none?
- AI like us? Do we really want it?
- AI creativity – AI art, music, literature, philosophy – is there one?
- Synthetic philosophy; philosophy of AI, or by AI?
- Epistemic States in AI systems- are there any?
- Belief forming processes in AI systems
- Deep ethics in AI systems, ethical problem space, dual use AI- what are we talking about?

- What is agency? AI threats to human agency, creativity, truth, and society – realities, sci-fi visions, dystopias or utopias; reading tea-leaves of technology, who should be in control of whom, do we lose ontological status of primary epistemic agents.

- Do we need philosophy in AI or having just clever programmers enough?

Submissions

The proposals for the presentations should be submitted for the anonymous review to <phil.in.AI.2023[at]gmail.com>. On the separate page the author(s) should submit the full name, affiliation and a short CV (up to 150 words). The proposals should not exceed 300-400 words (including references) in (*.docx) format. We accept presentations in English.

Important dates:

Nov 10th – Deadline for abstracts

Nov 15th – The authors of accepted abstract are notified.

Nov 20th – Program finalized; Zoom contact sent out.

Dec 1-2nd – Conference online (ZOOM)

Abstracts

Keith Begley, *Two Problems for Political Representation by Artificial Intelligence*.....7

Alexandre Bretel, *Why do we need to put an end to techno-utopia?*.....8

Mariela Destefano & Anna Trifonowa, *A Multidimensional Curriculum for Arficial Intelligence in Primary Education*.....9

David Gamez, *The Myth of AGI*.....11

Nathaniel Gan, *Can AI systems imagine? A conceptual engineering perspective*.....12

Brittany A. Gentry, *The Art of Representation: AI and Human Choice*.....14

Wojciech Głażewski, *Cognition* in silico *is far from reality*.....15

Zekiye Goz, *A Non-Anthropomorphic Comprehensive Approach to Evaluating Creativity in AI: Revision of Rhodes's 4Ps*.....16

Jumbly Grindrod, *The Transformer Picture of Language*.....17

Roman Krzanowski & Paweł Polak, *Is Whole Brain Emulation (WBE) a road to SuperIntelligence (SI)?*.....18

Łukasz Mścisławski, *Is it possible for human beings to lose the status of primary epistemic agent in the context of AI systems?*.....19

Maciej Musiał, *What should we take into account while deciding whether we want AI like us?*.....20

Hadeel Naeem, *AI cognitive extension and epistemic integration*.....21

Adam Olszewski, *Philosophy of AI and Solipsism*.....23

Antonio Oraldi, *The Co-Constitution of Humans and Technology: On the Concept of Agency in the Age of AI*.....24

José Antonio Pérez-Escobar & Deniz Sarikaya, *AI safety, value alignment and the later Wittgenstein*.....25

Luka Poslon & Anto Čartolovni, *Outpacing the Trustworthiness in LLM Use in Medicine by Addressing Opacity and Enhancing Explainability*.....26

Dakota Root, *Inbetweenness: the existence of artificial intelligence systems*.....28

Kristina Šekrst, *Do computers hallucinate electric Fata Morganas?*.....30

Paweł Stacewicz & Krzysztof Sołoducha, *The Turing test and the issue of trust in AI systems*.....31

Ljupcho Stojkovski, *AI Companionship: A Step forward or backwards in addressing loneliness?*.....32

Luca Tenneriello, *AI, adjustable autonomy, and human responsibility: the case of authorship and intellectual property*.....33

Rui Vieira da Cunha, *Technomoral change and the case for the AI alignment problem as a transformative experience*.....34

Dilek Yargan, *To what extent are LLMs creative?*.....36

ABSTRACTS

in alphabetical order of first authors

Keith Begley,

Two Problems for Political Representation by Artificial Intelligence

(Department of Philosophy, Durham University)

The notion of representative democracy faces a challenge from artificial intelligence. AI could potentially provide a more efficient conduit for the democratic representation of one's views than a member of parliament. It could also supersede traditional forms of direct democracy that require a high-level of participation and many costly ballots. However, there are potential theoretical issues involved with using AI or machine learning in this way, because it leads to various forms of inherited epistemic defeat. This paper will outline, discuss, and compare two such forms of problem in relation to political representation.

One problem is an instance of the now well-known explainability or opacity problem for black-box algorithms (Danaher 2016). Although we would be able to provide the exact cause of the representations or political choices made on behalf of an individual or population, that is, the exact series of matrix operations leading to an output, we would not be able to provide a reasoned explanation of those representations. That is, we would be unable to articulate the rationale for a particular choice or output.

There is a further problem that has been mentioned more sparsely in the machine learning literature (Hinton 2018; D'Amour et al. 2022), which I think has not been fully recognised as a distinct form of inherited epistemic defeat. An example of this is when two models have the same level of empirical adequacy, and have the same outputs, yet have different internal structures. Properly understood, this is a form of underdetermination. In the case of political representations, a model need only correspond to the population's political views in a way that is empirically adequate. Some such models could eventually diverge radically from human preferences, while others could continue providing accurate representations indefinitely, but there would be not be a way to distinguish between them.

Select Bibliography:

Danaher, J. (2016) 'The Threat of Algocracy: Reality, Resistance and Accommodation'. Philosophy & Technology 29: 245–268.

D'Amour, A., Heller, K., Moldovan, D. et al. (2022) 'Underspecification Presents Challenges for Credibility in Modern Machine Learning', Journal of Machine Learning Research 23: 1–61.

Hinton, G. (2018) 'Deep Learning—A Technology With the Potential to Transform Health Care', JAMA 320(11): 1101–1102.

Alexandre Bretel,

Why do we need to put an end to techno-utopia?

("Ethics & AI" Chair of Multidisciplinary Institute in Artificial Intelligence in Grenoble)

Certain movements such as long-termism, singularitarianism and transhumanism project a future in which the technological promises are multiplied by the development of artificial intelligence. These movements could be described as techno-utopian, in that they project a better world into the future, conditioned mainly by certain technological advances. However, some authors have questioned the very relevance of the concept of utopia. Whether it was Günther Anders (The Outdatedness of Human Beings, 1956), Hans Jonas (The Imperative of Responsibility, 1979) or Hannah Arendt (The Human Condition, 1958), these twentieth-century German thinkers strongly criticised the concept of utopia. So, for Günther Anders, we are already living in an anti-utopian world, in the sense that the complexity of our environment is already preventing us from understanding the present world. It's not a dystopia, in other terms a world to be avoided but one that we can imagine. Our mental representations are already unable to keep up with changes in the present world, and it is even more unlikely that we will be able to imagine or anticipate other alternatives. This does not mean, however, that projects cannot be undertaken to improve current conditions, but a conception that is too all-encompassing, and above all places too much hope in certain technologies, to the detriment of other social aspects for example, proves to be illusory and even dangerous. For Hannah Arendt, utopia begins with Plato, who conceives of the world in terms of eternal ideas that are instantiated in the world. Utopia would be the transcription of this metaphysics, but would at the same time forget the complexity of human relations. For Hans Jonas, utopia remained harmless as long as humanity did not have the technical means to try to achieve it. Now that the means are available, the temptation to achieve it is strong, with the risk of subordinating human and living beings to the achievement of this ideal. This temptation is a challenge to responsibility, which is only supposed to be assumed for what it is possible to answer for, which seems impossible for applications with consequences beyond our reach. In the course of this presentation, we will link certain techno-utopian projects to the philosophical critique of utopia, to better understand the scope of what we can expect from future advances in AI.

Mariela Destefano & Anna Trifonowa,
A Multidimensional Curriculum for Arficial Intelligence in Primary Education

(National Council for Scientific and Technical Research,
& Institute of Philosophy, University of Buenos Aires)

A Multidimensional Curriculum for Artificial Intelligence in Primary Education Artificial intelligence (AI) is a multidisciplinary field that has the potential to transform primary education (Luger and Stubblefield 1993). However, there's a lack of research and comprehensive curricula focused on integrating AI into primary education. Currently, a few fragmentary curricula in informatics exist, but they don't address the cultural context and human capacities related to AI (Williams et al 2019; Su and Yang 2022, Yang 2022). This paper introduces a multidimensional curriculum implemented in a primary school in Barcelona, Spain, which covers technological, philosophical, cognitive, and cultural dimensions.

Technological Dimension: The curriculum is practical and teaches children to create and use AI engines through block-based coding. Students engage in project-based learning, gaining hands-on experience in solving various problems using digital solutions.

Philosophical Dimension: This human-centered curriculum emphasizes the importance of understanding AI through a critical comparison with human intelligence. It aligns with the Beijing Consensus on AI and Education, stressing that AI should be developed with human control and consideration of human intelligence's features.

Cognitive Dimension: The curriculum fosters critical reflection and enhances children's executive functions, combining critical thinking with creative problem-solving skills.

Cultural Dimension: This multilingual curriculum caters to the specific needs of Catalan primary students, designed to work in a plurilingual community where Spanish and Catalan coexist in formal education. This multidimensional curriculum draws on various fields, including the Philosophy of Mind, the Philosophy of Children, Cognive and Developmental Psychology, and coding tools like Scratch and App Inventor, as well as Machine Learning for Kids. This holistic approach to AI education for children sets this proposal in the trends of the digital education. It not only promotes digital literacy but also provides a broader liberal education for children.

Bibliography:

Beijin Consensus on Artificial intelligence and Education (2019), 16 – 18 May 2019 Beijing, People's Republic of China, UNESCO

Luger & Stubbelfield (1993) "AI: Structures and Strategies for Complex Problem Solving", Benjamin Cummins

Williams, R., Park, H. W., Oh, L., Breazeal, ,C. (2022) "PopBots: Designing an Artificial Intelligence Curriculum for Early Childhood Education", The Ninth AAAI Symposium on Education Advances in Artificial Intelligence (EAAI-19)

Yang, W. (2022) "Artificial Intelligence Education for Young Children: Why, What, and How in Curriculum Design and Implementation", Computer and Educaos: Artificial Intelligence, 3.

David Gamez,

The Myth of AGI

(Department of Computer Science, Middlesex University, London, NW4 4BT, UK)

Intelligence is often believed to be an absolute property that is independent of a system's environment. However, real human intelligence varies considerably with the environment. A person who has played video games all their life has a higher level of intelligence in video game environments than a member of an Amish community, who has never used a computer. All humans have low levels of intelligence in hundred-dimensional environments and in environments containing petabytes of numerical data. As Chollet (2019) points out, the human brain evolved to help us survive in a hunter-gatherer environment and it has a limited ability to generalize beyond this environment. If human intelligence is not completely general, then there is very little reason to believe that completely general artificial intelligence is possible. A much more plausible view is that there are many different types of natural and artificial intelligence that are optimized for different environments. This idea has often been discussed in the literature on intelligence. For example, Gardner (2006) claims that there are multiple types of intelligence, including musical intelligence, linguistic intelligence and emotional intelligence. Warwick (2000) frames this more generally with his idea that intelligence is a high-dimensional space of abilities. If AGI is a myth, natural and artificial intelligence should be compared according to their degree of generality - there is not an absolute distinction between narrow AI and artificial general intelligence.Intelligence measures, such as IQ and g, attribute an amount of intelligence to systems independently of the environments that they are in. If human intelligence is not completely general, then systems' levels of intelligence should be indexed to the sets of environments in which they have these levels of intelligence. The last part of the talk outlines a new universal measure of intelligence that is based on the number of accurate predictions that an agent makes in a set of environments (Gamez 2021). This environment-indexed measure could make a significant contribution to intelligence research and AI safety.

References

Chollet, F. (2019) On the measure of intelligence. arXiv: 1911.01547.

Gamez, D. (2021). Measuring Intelligence in Natural and Artificial Systems. Journal of Artificial Intelligence and Consciousness 8(2): 285-302.

Gardner, H. (2006) Multiple Intelligences: New Horizons. Basic Books: New York.

Warwick, K. (2000). QI: The Quest for Intelligence. London: Piatkus.

Nathaniel Gan,

Can AI systems imagine? A conceptual engineering perspective

(postdoctoral research fellow at the National University of Singapore)

Some AI systems perform their target tasks using simulated representations of real-world scenarios; these simulations are sometimes called 'imagination' in the engineering literature (e.g., Wu, Misra, and Chirikjian 2020). If AI systems have imagination, this may have implications for their resemblance to humans and for the possibility of other capacities not typically associated with AI. We might thus consider if the term 'imagination' is appropriate in this context. This presentation will consider this question from a conceptual engineering perspective (for more on conceptual engineering, see Chalmers 2020). From a conceptual engineering perspective, the target question can be framed in terms of our goals regarding our concept of imagination. Our concept of imagination presently has two primary functions: it helps us understand the nature of imagination by facilitating comparison between imagination and other mental states (Currie and Ravenscroft, 2002), and it helps us understand how imagination aids our reasoning by facilitating consideration of imagination's epistemic roles (Kind and Kung, 2016). AI simulations align somewhat, but not perfectly, with these functions, hence they underdetermine the question of whether we should attribute imagination to AI. Nevertheless, a possible new function for this concept emerges when we think of AI systems as models of human intelligence. The more popular deep learning models are limited in the extent to which they can be said to model human intelligence, but simulation-based AI involve more human-like processes. Recognising AI simulations as instances of imagination may serve to highlight this similarity and perhaps even explain some resemblances.

Suggestions for adopting an AI-friendly concept of imagination will be offered. Besides broadening the way we typically use the term 'imagination,' we might also become open to the possibility of AI having capacities not typically associated with artificial systems. Countenancing AI imagination might also affect the way we think about our relation to AI and our understanding of human imagination.

References

Chalmers, D. (2020). What is conceptual engineering and what should it be? Inquiry: An interdisciplinary journal of philosophy. doi:10.1080/0020174X.2020.1817141

Currie, G., & Ravenscroft, I. (2002). Recreative Minds: Imagination in Philosophy and Psychology. Oxford: Oxford University Press.

Kind, A., & Kung, P. (2016). Knowledge Through Imagination. Oxford: Oxford University Press.

Wu, H., Misra, D., & Chirikjian, G. S. (2020). Is that a chair? Imagining affordances using simulations of an articulated human body. IEEE International Conference on Robotics and Automation,

About the Author

Nathaniel Gan is a postdoctoral research fellow at the National University of Singapore, working on spatial reasoning and imagination in robotics.

Brittany A. Gentry,

The Art of Representation: AI and Human Choice

(Assistant Professor of Philosophy, Utah State University, Logan, Utah)

The use of images in artistic and scientific modeling and representation is ubiquitous and ubiquitously complex. Both fields undertake the same kind of process and face similar challenges and choices when representing a given subject (van Fraassen 2008). Images and representations are the result of complex editorial processes that choses which features are relevant and necessary to communicating a particular idea or model or representation of some subject. Representation, both scientific and artistic is a process of choice with respect to things like what to represent, what counts as unnecessary information to the given image or representation, and what counts as relevant information to the image. This paper considers three implications for human choice and creativity that come with using generative AI for representation generally and, specifically, in modeling and artistic representation.

The first implication focuses on how outsourcing decisions around composition and subject in images and representations, artistic and scientific, reduces human awareness of meaning and symbolism. The second implication is focused on how reduced awareness of what Generative AI has been used on the image or model reduces human capacity to evaluate and engage with the quality of representation and abstract involved in a given project or output. The third implication that this paper is concerned with is how using AI might result in long-term underutilization of creative decision making that potentially reduces creative capacity in human beings. These implications raise further concerns about the ethical and developmental implications of using AI to make decisions about imaging and representation in artistic and scientific fields, which we will point towards in the conclusion of the paper.

References

Fraassen, Bas C van. Scientific Representation: Paradoxes of Perspective. Oxford University Press, 2008.

Hughes, Richard IG. “Models and Representation.” Philosophy of Science 64 (1997): S325–36.

Knuuttila, Tarja. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” Studies in History and Philosophy of Science Part A 42, no. 2 (June 2011): 262–71. https://doi.org/10.1016/j.shpsa.2010.11.034.

Knuuttila, Tarja, and Andrea Loettgers. “Modelling as Indirect Representation? The Lotka–Volterra Model Revisited.” The British Journal for the Philosophy of Science 68, no. 4 (2017): 1007–36. https://doi.org/10.1093/bjps/axv055.

Suárez, Mauricio. “Scientific Representation.” Philosophy Compass 5, no. 1 (January 2010): 91–101. https://doi.org/10.1111/j.1747-9991.2009.00261.x.

Wojciech Głażewski,

Cognition in silico *is far from reality*

(Institute of Philosophy, University of Bialystok, Poland)

Computers have always been called automatic data processing machines. The name of the microprocessor comes from the word "processing", which means to collect data, calculate (in the sense of arithmetic and logical operations), store it and output it. The structure of the processor does not allow anything else. It's a silicon wafer with fixed metal traces on it, creating billions of miniature transistors.

Arithmetic and logical operations are deductive in nature - the result is contained in the premises. In the case of computers, this means that the amount of information put into the system is equal to the amount of information taken out from the system. Processing changes data form only. This happens because the physical properties of the computer's computational substrate do not change while it performs informational operations. Each of the billion transistors remains in place and each of the paths connecting them does not change while it runs. Only formal operations are performed, therefore AI systems create nothing. They generate data by processing other data previously entered. Artificial Intelligence is sometimes called computational intelligence, but the proper technical term should be calculational.

The brain constitutes another informational system, based on an active, biological computational substrate. The transmission of impulses between nerve cells is reducible to formal operations, but the formation of new connections between them is not. The informational processes occurring during human thinking are not exclusively computational. They are accompanied by active changes in the structure of the substrate, modifying the paths along which signals flow. The new structure of the substrate constitutes a new form of data, and therefore new information. A phenomenon impossible to implement in silico.

References

Dodig-Crnkovic, G.; Miłkowski, M. (2023). Discussion on the Relationship between Computation, Information, Cognition, and Their Embodiment. Entropy 2023, 25, 310.

Floridi, Luciano (2019). What the near future of artificial intelligence could be. Philosophy and Technology 32 (1):1-15.

Nicolelis, Miguel A. L.; Cicurel, Ronald (2015). The Relativistic Brain: How it Works and why it cannot be simulated by a Turing Machine. Montreux: Kios Press.

Schuman, Catherine & Kulkarni, Shruti & Parsa, Maryam & Mitchell, J. & Date, Prasanna & Kay, Bill (2022). Opportunities for neuromorphic computing algorithms and applications. Nature Computational Science. 2. 10-19.

Thagard, Paul (2022). Energy Requirements Undermine Substrate Independence and Mind-Body Functionalism. Philosophy of Science 89 (1):70–88.

Zekiye Goz,

A Non-Anthropomorphic Comprehensive Approach to Evaluating Creativity in AI: Revision of Rhodes's 4Ps

(Department of Philosophy, Durham University, 50 Old Elvet, DH1 3HN, United Kingdom)

The development of machine learning-based AI systems has led to the labelling of many products or even the systems themselves as creative in various fields such as art, literature, etc. This phenomenon has raised significant concerns regarding the essence of creativity and the criteria used for assessment. My main goal here is to show the necessity for a non-anthropomorphic comprehensive approach in order to achieve more accurate evaluation, while highlighting some notable limitations that are inherent in the traditional perspective.

The notion of creativity is primarily described as "the ability to come up with ideas that are new, surprising, and valuable" (Boden 2004, p. 1). However, this definition is mostly interpreted in terms of agent-based, process-based, or product-based evaluations for creativity. Hence, such evaluations are not sufficient because they do not account for the social context. As a result, in recent years, there has been a subtle shift in traditional approach, which emphasizes the importance of judgements provided by external world.

With regard to this point, it is obvious that we need a more comprehensive approach that encompasses both the components of traditional approach and the impact of society. In the literature, this approach is known as the 4Ps of creativity, introduced by Rhodes (1961) – person, process, product, and press. However, this approach is human-centred.

My main objective is to explore how to develop a non-anthropomorphic comprehensive approach to assess the notion of creativity in AI. In my view, the best way to achieve this is through a revision of Rhodes's 4Ps of creativity. The new version of this approach consists of producer, process, product, and perceiver. To consider all these points, the revised version of Rhodes's 4Ps of creativity presents us a framework to have a better understanding of the notion creativity, and then assessment criteria.

Bibliography

Boden, M. (2004). The Creative Mind: Myths and Mechanisms, Taylor & Francis e-Library.

Jordanous, A. (2012). 'A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on what it is to be creative', Cognitive Computation, Volume 4(3), pp. 246-279.

Rhodes, M. (1961), 'An Analysis of Creativity', Phi Delta Kappan International, Volume 42(7), pp. 305-310.

Jumbly Grindrod,
The Transformer Picture of Language

(Lecturer in Philosophy, University of Reading)

My research background lies primarily within philosophy of language and epistemology. A great deal of my previous research has focused on issues around epistemic contextualism i.e. the view that the meaning of "know" is dependent upon the context of its use. More recently, I have focused on whether methodologies in corpus linguistics and computational linguistics can help inform philosophical debates. This includes employing corpus linguistic methodologies within experimental philosophy, as well as investigating the recent progress large language models to consider whether they can shed light on the nature of linguistic meaning. I have previously published in journals such as *Ergo*, *Episteme*, *Mind & Language*, *Topoi*, and *Inquiry* (full list of publications is available on my website).

Roman Krzanowski & Paweł Polak,

Is Whole Brain Emulation (WBE) a road to SuperIntelligence (SI)?

(Faculty of Philosophy, Pontifical University of John Paul II in Kraków)

This paper is a critical review of the current discussions on WBE and SI. WBE and SI are all so-called conceptual technologies. To put it simply, there are technologies that do not exist yet, and there is no clear path to developing them. On one hand, all the technologies that we have now were in this mythical phase at some point in time. So dismissing them would go against our experience. On the other hand, recent developments in brain studies (e.g., D. Eagelman on neuronal brain structure) and most recent artificial intelligence technologies (e.g., F. Chollet on human intelligence and LLMs) suggest that the concept of WBE and SI as formulated by Bostrom may have to be significantly revised, if not abandoned. In this paper, we look at the original assumptions behind WBE and the WBE road to SI as proposed a decade ago and discuss how these ideas hold ground today. We posit that one of the key problems with conceptual technologies like WBE and SI is the lack of a sound philosophical understanding of what we really mean by them. These ideas have been proposed by technology experts who assume tacitly that things can be done if they are thought of (vide McCarthy AI project). And, if history is any guide to the future, we claim that a lack of fundamental understanding of these ideas will lead to a "WBE and SI" winter or at least cooling down period, as it did with AI

Łukasz Mścisławski,

Is it possible for human beings to lose the status of primary epistemic agent in the context of AI systems?

(Wrocław University of Science and Technology)

AI systems proved to be incredibly efficient in multiple applications working much faster than human beings. Such a situation raises much excitation, justified or not, but also causes that much fear is entering the stage. Are we doomed to lose the status of primary epistemic agents and destined to be submitted to omnipresent and almost omnipotent technology, depending on it both in existential and epistemic dimension? The present paper is not aiming to present totally dystopian vision of the future of mentioned spaces of human activity. Its goal is to carefully examine some possibilities of very positive contribution of AI systems to human cognizance. Their limitations will also be taken into account, as well as some possible interesting issues related to acquiring knowledge by human beings and the problem of justification of the results (especially in the context of mathematics and physics, (Wójtowicz, 2012) , (Murawski, 2015) , (Leciejewski, 2013) , (Graczyk, Strzelczyk and Matyka, 2023)).

It seems that a philosophical approach to achievements of "digital paradigm" in the area of science can possibly result in valuable suggestions regarding possible new concept of scientific knowledge and being an epistemic agent, enriched by achievements from AI area. And it also seems that human beings are still holding the title of primary epistemic agents, however the meaning of the title has to be modified.

Keywords:

AI, epistemology, philosophy, science, knowledge

References:

Graczyk, K.M., Strzelczyk, D. and Matyka, M., 2023. Deep learning for diffusion in porous media. https://doi.org/10.48550/arXiv.2304.02104.

Leciejewski, S., 2013. Cyfrowa rewolucja w badaniach eksperymentalnych: studium metodologiczno-filozoficzne = Digital revolution in experimental research: methodological and philosophical study. Wydanie I ed. Seria Filozofia i logika. Poznań: Wydawnictwo Naukowe UAM.

Murawski, R. ed., 2015. Filozofia matematyki i informatyki. Kraków: Copernicus Center Press.

Wójtowicz, K., 2012. O pojęciu dowodu w matematyce. Monografie FNP. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Maciej Musiał,

What should we take into account while deciding whether we want AI like us?

(Zakład Filozofii Kultury, Wydział Filozoficzny UAM)

This presentation makes a far-fetched and controversial assumption that humans someday will design and develop AI entities that will be recognized as persons with moral status and rights analogous to those of human beings—Artificial Persons (APs for short). In other words, it is assumed that we will be able to develop AI with phenomenal consciousness, free will, intentionality, and any other properties that we find necessary for "being like us". Such a scenario is most often discussed in terms of APs' moral agency and in terms of how this agency should be shaped to protect the well-being of humans. However, here, I would like to focus on robots' moral patiency and developing them in a way that protects their well-being, since it is important to note that if humans could develop entities equal to themselves, they would have to care for them as much as for themselves. In other words, APs would no longer be our tools but persons deserving the same treatment as we do. Hence, I would like to highlight that bringing APs into existence would 1) obligate us to solve some philosophical, e.g. ethical, issues, especially in reference to the process of designing such entities (some of which are interestingly parallel to questions about—also far-fetched and controversial—prenatal human enhancement), and 2) require us to be ready to share some resources that currently are at our exclusive disposal. As for 1), it involves questions such as: should APs be designed to experience childhood?; should they be able to possess offspring?; should they be developed to experience and/or cause suffering? As for 2), it includes awareness that APs would compete with us both for material resources (they would probably have rights to work, to possess money and other material goods) and symbolic resources such as love (some people may choose to marry AP rather than a human being) or respect (some APs might be better than humans in sports, arts and other respectable activities). Hence, this presentation would like to emphasize that developing "AI like us" implies not only epistemic and technical challenges (how to create consciousness, etc.), but also equally or even more daunting ethical challenges. Whether or not we would like to face these challenges should be an important part of answering the question: "Do we want AI like us?".

Hadeel Naeem,

AI cognitive extension and epistemic integration

(junior fellow at the Käte Hamburger Kolleg Aachen: Cultures of Research
RWTH Aachen University)

Some have argued that our fluent and seamless reliance on AI systems may cause us to incorporate into our cognitive systems the strange errors these systems commit. For instance, AI systems based on deep neural networks (DNN) can succumb to adversarial exemplars, where even after being trained on a large dataset, they misidentify what to us is clearly an instance of some object (e.g. a car) as something else (e.g. a non-car) (Szegedy et al. 2014).

According to the thesis of extended cognition (Clark and Chalmers 1998), our cognitive states may sometimes be realised at least partially outside our bodies, for instance in notebooks, phones, and other devices. When the devices with which we extend our cognition are AI-enabled, then the AI's strange defects threaten to become defects in our own cognition. One proposed solution to this problem, owed to Michael Wheeler, is that we design DNN-based AI systems such that they are not employed transparently (Wheeler 2019). This concept of transparency is borrowed from the phenomenological literature, which describes how, when we fluently employ a tool (say, a hammer), the tool may disappear from our focus of attention so that we are directed at the task at hand (say, hammering a nail). Several proponents of the extended cognition thesis have argued that transparency is necessary for cognitive extension. It is therefore not surprising that Wheeler advocates intransparent use of AI technologies that render cognitive extension impossible.

While some aspects of AI extension are worrying, we think discouraging transparent employment and preventing all AI extension isn't a good solution. Most obviously unreliable AIs – we show – are unlikely to be employed transparently, firstly. Second, processes that are transparently employed may still be reflected on at a later time or the same time. This is true especially for processes epistemically integrated into our cognitive systems. We can use them transparently and still become aware when they turn unreliable. Agents can therefore responsibly employ AI systems that epistemically integrate into their cognitive systems.

According to our conclusion, we offer design and policy recommendations to motivate proper epistemic integration of AI systems into our cognitive systems.

References

Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” Analysis

Szegedy, Christian, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus. 2014. “Intriguing Properties of Neural Networks.”

Wheeler, Michael. 2019. “The Reappearing Tool: Transparency, Smart Technology, and the Extended Mind.” AI and Society

Adam Olszewski,

Philosophy of AI and Solipsism

(Faculty of Philosophy, Pontifical University of John Paul II in Kraków)

In my paper, I will inquire about what kind of philosophy is suitable for talking about AI. I will argue that solipsism is suitable for this role. In the presentation I provide the basic definitions of solipsism and their scheme. I also present the concept of the Subject in the versions of Hilbert, Brouwer and Turing. The latter will serve to reflect on the main problem of the paper.

Antonio Oraldi,

The Co-Constitution of Humans and Technology: On the Concept of Agency in the Age of AI

(University of Lisbon, Centre of Philosophy (CFUL), Praxis Research Group)

This presentation delves into the intricate relationship between human agency and technology, with a specific focus on artificial intelligence (AI). To begin with, I will approach the issue from a theoretical standpoint. Is there a human agency separate from technics and technology? The question of technology and agency has been approached from a variety of philosophical standpoints in classical accounts of technology (Marcuse, 1964; Habermas, 1968). Beyond the classical perspectives, I will point out how human agency is always mediated with and through technology, and, on other hand, technology is always mediated, filtered through human social values and power (Feenberg, Stiegler, 2011; Latour, 1994). In other words, there is a deep interrelation that establishes

a process of co-constitution between human and technical agency. Such a co-constitution implies a fundamental openness both of technics and human agency to forms of contestation and subjectivation. Exploring this dynamic co-constitution further, the presentation identifies two modes of agency in connection with technology and AI. Firstly, we can consider human agency as a technically embedded agency when the technology enables a new course of action that would not be possible without the technical embedding (e.g., protheses). Secondly, there is a foundational level of human agency when the sequence of technical actions remains traceable to human initiation (e.g., input to generative AI). The question then is not the elimination of agency as such but its transformation, as well as some of the disabling effects of social, economic, and political relations embedded in technical structures.

In the last part, I will discuss some empirical examples to substantiate the theoretical claims on technically embedded human agency in connection with generative AI. I will explore forms of agency mediated through AI, including instances of AI-powered contestation surrounding data extraction. Finally, I will conclude by considering learning as an enabler of agency and discussing the possibility of human learning in connection with machine learning.

References:

Habermas, J. (1970) "Technology and Science as 'Ideology'". In Toward a Rational Society. Beacon Press.

Feenberg, A. (2017) Technosystem: The Social Life of Reason. Harvard University Press.

Latour, B. (1994). On technical mediation. Common Knowledge 3 (2):29-64.

Marcuse, H. (1964) One-dimensional Man. Routledge.

Stiegler, B. (2011). Technics and time, Vol. III. Trans. Barker, S. F. Stanford University Press.

José Antonio Pérez-Escobar & Deniz Sarikaya,
AI safety, value alignment and the later Wittgenstein

(École Normale Supérieure, Paris Sciences et Lettres University
& Vrije Universiteit Brussel)

In this talk we argue that the later Wittgenstein's philosophy of language and mathematics, substantially focused on rule-following, is relevant to understand and improve on the Artificial Intelligence (AI) alignment problem: his discussions on the categories that influence alignment between humans can inform about the categories that should be controlled to improve on the alignment problem when creating large data systems to be used by supervised and unsupervised learning algorithms as well as when introducing hard coded guardrails for AI models. We cast these considerations in a model of human-human and human-machine alignment and sketch basic alignment strategies based on these categories and further reflections on rule-following like meaning as use. To sustain the validity of these considerations, we also show that successful techniques employed by AI safety researchers to better align new AI systems with our human goals are indeed congruent with the stipulations that we derive from the later Wittgenstein's philosophy. However, their application may benefit from the added specificities and stipulations of our framework: the categories of the model and the core alignment strategies presented in this work extend on the current efforts and provides further, specific AI alignment techniques.

The categories and alignment strategies outlined in the talk hold the potential to enrich the discourse on algorithmic bias. By delving into the categories underlying alignment, this approach offers a pathway towards cultivating fairer, more unbiased AI systems that align with human goals and values. Our approach may reduce algorithmic bias in several ways. For instance, a meaning-as-use-training based on the model parameters may reduce unintended generalizations like Google's black-people-labelled-as-gorillas fiasco. It can also help in cases

where two human populations have different moral standards, and the AI must respond in a way that adapts to the standards of a population despite being developed by the other population. An example of the latter situation that we discuss is the "Moral Machine experiment", an ambitious global study initiated by MIT to understand human preferences in the context of moral dilemmas faced by autonomous vehicles. Say, a collision is unavoidable, but depending on the action taken the outcomes differ. For instance, the car can either compromise the safety of young passengers in a car or elderly pedestrians. These judgements vary across cultures, subpopulations and even individuals, making misalignment likely, but we argue that our approach leads to an improvement.

Luka Poslon & Anto Čartolovni,

Outpacing the Trustworthiness in LLM Use in Medicine by Addressing Opacity and Enhancing Explainability

(Digital healthcare ethics laboratory (Digit-HeaL), Catholic University of Croatia, Zagreb, Croatia & Digital healthcare ethics laboratory (Digit-HeaL), School of Medicine, Catholic University of Croatia, Zagreb, Croatia)

Ensuring data security and patient privacy takes precedence in integrating AI, particularly large language models (LLMs) like ChatGPT, into medical applications. The obstacles presented by LLM opacity, particularly when considering its application in the medical field, highlight the moral challenges related to transparency and the "black-box problem." Ensuring that AI systems adhere to ethical standards like explainability and trustworthiness is imperative. We should protect patients' rights and interests while maximising the benefits of AI in medicine and healthcare by giving the highest importance to these principles. It is essential to inform patients and physicians about each algorithmic prediction LLMs make to enhance medical decision-making. Since there are everyday scenarios in medicine with high-risk outcomes, explainability as an ethical need needs to be widely accepted in the medical field. Some progress has been made in overcoming the obstacles caused by opacity. The "chain of reasoning" prompt for LLMs, which may provide sequential reasoning before delivering a final output, is one example of how progress has been made in resolving opacity difficulties. However, such an approach must be revised to fulfil the explainability criteria. Explainability should be prioritised, and we must also increase trust in using medical AI tools like LLMs. By doing this, we want to foster confidence in algorithmic solutions by providing quick and precise medical diagnoses and treatments. Explainable artificial intelligence (xAI) offers a human-centred solution to enhance trust and make algorithmic predictions more explainable, improving medical decision-making by decreasing LLMs opacity.

Keywords:

ChatGPT, xAI, LLM, Ethics, Medicine, Trust

Literature:

1. Bruckert S, Finzel B, Schmid U. The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. Front Artif Intell. 2020 Sep 24;3:507973. doi: 10.3389/frai.2020.507973. PMID: 33733193; PMCID: PMC7861251.
2. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, New York (2019)

3. Liévin, V., Egeberg Hother, C., Winther, O., Can large language models reason about medical questions?, arXiv:2207.08143, last accessed 2023/5/26
4. Ordish J, Hall A. Black Box Medicine and Transparency: Interpretable Machine Learning. PHG, Foundation. 2020, last accessed 2023/5/26
5. O'Neill, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. New York, NY: Crown Publishing Group, p. 31
6. Q. Vera Liao, Kush R. Varshney, Human-Centered Explainable AI (XAI): From Algorithms to User Experiences, arXiv:2110.10790

Dakota Root,

Inbetweenness: the existence of artificial intelligence systems

(postdoctoral researcher, Chair Ethics and AI,
Institut de Philosophie de Grenoble, Université Grenoble Alpes)

Recent research studies the possibilities of experience with and through artificial intelligence systems (AIS). For example, Wellner describes how generative AI has become the "writing I" producing text (2018, 218). Liberati's analysis of user relationships with chatbot XiaoIce emphasizes the possibility of "digital intimacy" with chatbots capable of situational, flowing dialogue (2022). Kanemitsu suggests that social robots introduce an another-other who influences human action and feels like "a real other" (2019, 54). Gunkel's work on robots has led him to argue that AIS may "deconstruct the existing logical order that differentiates person from thing" (2023, 162). In this presentation, we use Gunkel's point as a jumping off place to think about the existence of artificial intelligence systems. We will ask 1) are AIS ontologically different than other objects? We will use key features of AIS to distinguish them from other objects, illustrating our point with real-life examples. AIS process and identify data, for example Google maps uses neural networks to distinguish features of the environment (Lookingbill and Russell 2019; Boiling and Bohl 2022). AIS add new content into the world where the AIS is the creator, exemplified by Chat-GPT's text generation (Miroshnichenko 2018). AIS learn from past data for adjustment and improvement, illustrated by the Transformer deep learning architecture for natural language processing (Uszkoreit 2017). Finally, AIS can actively adapt the environment, such as Google Nest thermometer system that changes temperature in a room (Google Nest n.d.). We argue that these distinguishing features suggest that AIS are ontologically distinguished from other objects. We introduce the term inbetweenness, emphasizing the relationality that characterizes machine learning, to support our arguments. Finally, we will address what the existence of AIS means for subjective experience of humans.

Bibliography:

Andrew Lookingbill, and Ethan Russell. 2019. "Google Maps 101: How We Map the World." Google. July 22, 2019. https://blog.google/products/maps/google-maps-101-how-we-map-world/.

Bolling, Liam, and Kristi Bohl. 2022. "How AI and Imagery Build a Self-Updating Map."

Google. April 7, 2022. https://blog.google/products/maps/how-ai-and-imagery-build-self-updating-map/.

Google Nest. n.d. "Google Nest Connected Home." Google Store. Accessed June 4, 2020. https://store.google.com/us/category/connected_home.

Gunkel, David J. 2023. Person, Thing, Robot. Cambridge: MIT Press.

Kanemitsu, Hidekazu. 2019. “The Robot as Other: A Postphenomenological Perspective.” Philosophical Inquiries 7 (1): 51–61. https://doi.org/10.4454/philinq.v7i1.238.

Liberati, Nicola. 2022. “Digital Intimacy in China and Japan.” Human Studies, July, 1–15. https://doi.org/10.1007/s10746-022-09631-9.

Miroshnichenko, Andrey. 2018. “AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is ‘Yes’).” Information 9 (7): 183. https://doi.org/10.3390/info9070183.

Uszkoreit, Jakob. 2017. “Transformer: A Novel Neural Network Architecture for Language Understanding.” Google Research. August 31, 2017. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html.

Wellner, Galit. 2018. “From Cellphones to Machine Learning. A Shift in the Role of the User in Algorithmic Writing.” In Philosophy of Digital Media, edited by Alberto Romele and Enrico Terrone, 205–24. Ebook: Palgrave Macmillan.

Kristina Šekrst,

Do computers hallucinate electric Fata Morganas?

(University of Zagreb)

With the advent of large language models, the concept of artificial hallucinations has surfaced, denoting a confident AI model response which is not justified by its training data. For example, an AI language model might confidently state that the current net worth of a person is a certain amount of money, which is a made-up number, or the model might provide references that do not exist (cf. Bhattacharyya et al. 2023). The phenomenon was compared to human hallucinations, but instead of being a psychological issue, here it lies in the domain of epistemology: having unjustified 'beliefs'. First, we will observe how large language models are trained and used, and then see the problems arising with unjustified responses without any justification in the training data portion. Hallucinations can derive from incorrect data, but more often, they derive from the training process itself (Ji et al., 2023). For example, various errors in encoding and decoding between text and vector representations are one possible source of hallucinations. We will see how extra guardrail validations (cf. Varshney 2023) might be seen as either belief revisions or parts of reliabilistic processes. Second, we will posit that qualia-like states or similar seemingly emergent consciousness might also be seen as confabulatory responses, i.e., faulty ones or results of a specific training process.

References

Bhattacharyya M. et al. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus. 15 (5).

Ji et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. Association for Computing Machinery. 55 (12): 1–38

Varshney, N. (2023). A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. arXiv:2307.03987

Paweł Stacewicz & Krzysztof Sołoducha,

The Turing test and the issue of trust in AI systems

(Politechnika Warszawska, Wydział Administracji i Nauk Społecznych
& Wojskowa Akademia Techniczna)

Abstract. The Turing test, which is a verbal test of the indistinguishability of machine and human intelligence, is a historically important idea that has set a way of thinking about the artificial intelligence (AI) project that is still relevant today. According to it, the benchmark for AI is human intelligence, and the key skill of AI systems is supposed to be their communicative ability—which involves, among other things, explaining the decisions the system makes.

Adopting a contemporary point of view that goes beyond Turing's original idea, the paper will develop and justify the thesis that the ability of an AI system to explain the reasons behind its decisions is a factor that significantly increases the user's trust in the system. It is particularly important when the user accepts the principle of limited trust in the system —that is, he or she does not trust the system unreservedly, but is aware of its limitations and its potential for making mistakes.

Passing the original Turing test by a machine does not guarantee that the machine is trustworthy from a human point of view; on the contrary, the idea of a machine capable of "cheating" or "outsmarting" a human is implicit in the idea of the test. The main reason for this is the imitative and behavioural nature of the whole procedure. The AI system is designed to answer questions in a maximally 'human' way, without referring at all to its inherent internal processing patterns.

Despite the above, the thesis of the paper is that, both in Turing's original proposal and in its critique, there are some valuable elements that lead to the specification of the boundary conditions of a good trust test. A test that meets these conditions should be at least: a) non-imitative, b) non-behavioural, c) focused on explanatory ability, taking into account, however, the design and operation of the machine (and not just human expectations), d) focused on the machine's ability to learn.

At the end of the paper we will also outline an ethical thread. Assuming that, in addition to technology and communication factors (such as those above), trust in an AI system is also influenced by the system's compliance with certain ethical standards, we will look at some ideas for enriching the original Turing test with procedures for examining the 'ethics' of an AI system.

Ljupcho Stojkovski,

AI Companionship: A Step forward or backwards in addressing loneliness?

(assistant professor in international law and international relations
at the Faculty of Law "Iustinianus Primus" Skopje, Ss. Cyril and Methodius University)

Loneliness is becoming one of biggest problem of modern times. Today, it affects not only old people above 65, but it is widespread among young people as well (according to some researches in 1 in every 3 young people). Loneliness is found to contribute to many mental illnesses, such as depression, anxiety, dementia, but it worsens physical health too. Thus, immediate reflection is needed to address it.

One area of advancement of AI technology is the domain of "relation artifacts", that is computational objects explicitly designed to engage a user in a relationship (Turkle, 2007). These artifacts could take different forms, from software apps for smartphones to material things like (care) robots, sex dolls, etc. These "evocative objects" (Turkle, 2007) are (increasingly going to be) used to address some human desires and needs, such as love, sex or loneliness.

While the issue of loneliness is a complex one, affecting many individuals in different ages, periods of life and socio-economic (and technological) circumstances, and consequently, one measure cannot fit all cases, it is worth deliberating further whether AI companions can help or worsen the problem of loneliness. On the one hand, AI companionship artifacts do not possess the human qualities of understanding and empathizing with a human being. On the other hand, despite this, for many (observers and people suffering from loneliness) they are "better than nothing", better than feeling completely lonely. The concern, however, especially as AI progresses, is that these digital companions can become even better than the real thing, i.e. than human-human contact and relationship (Brooks, 2021). And there are a lot of things that point to this concern – AI's learn about the users' preferences and desires, what triggers the user's emotional buttons, probably better than human beings, and therefore always providing an emotionally satisfying feedback to the user; the AI companion is always there for the human being; it can never let the user down; it is not needy, and the user cannot hurt its feelings; etc. As a result, this initially "better-than-nothing" solution could soon turn into utopian but then also a dystopian scenario of individuals feeling not lonely but becoming even more isolated.

Therefore, the thesis in this paper is that while AI companionship should definitively not be overruled as a potential assistance to the problem of loneliness (in certain cases), its wide use is not without problems for human relationship and can, in fact, undermine the problem it tries to address by creating more human isolation (and thus loneliness) at the price of human-AI companionships.

Luca Tenneriello,

AI, adjustable autonomy, and human responsibility: the case of authorship and intellectual property

(Dipartimento di Filosofia, Sapienza Università di Roma, via Carlo Fea, 2 - 00161 Roma)

AI technologies are revolutionizing our everyday life in many ways, bringing out philosophical questions about normativity and responsibility, among others. In this framework, the implementation of adjustable autonomy into AI systems allows for a dynamic shift in control, enabling AI to make decisions autonomously or collaborate with humans, while respecting certain directions prompted by human operators. On one hand, this adaptation of autonomy enhances creativity and our epistemic comprehension of realty (Coeckelbergh and Gunkel, 2023); on the other hand, it complicates the legal and moral implications for human responsibility.

As a case study, I will take into consideration the problems of authorship and intellectual property. It has been noted that the emergence of AI technologies might undermine academic integrity (Eke, 2023). I shall try to argue that the use of AI-driven text generators in writing or editing texts to be covered by intellectual property (papers, books...) does not morally and legally alter authorship, as long as this use is merely instrumental, i.e. implemented alongside the active role of the human author. In other terms, intellectual property (and creativity) remains safe if the human author uses AI only as a mere brainstorming tool: that is, to acquire suggestions on the topic of the paper, broaden its perspectives and possible implications, and receive a textual basis to be humanly elaborated and enriched.

Essential References:

Coeckelbergh, M., Gunkel, D.J. (2023), "ChatGPT: deconstructing the debate and moving it forward". AI & Society, 38.

Eke, D. O. (2023), "ChatGPT and the rise of generative AI: Threat to academic integrity". Journal of Responsible Technology, 13.

Farahani, N.A. (2023), The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology. St. Martin's Press: New York.

Floridi, L. (2023), The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities. OUP: Oxford.

Frankel, R., Krebs, V.J. (2022), Human Virtuality and Digital Life. Philosophical and Psychoanalytic Investigations. Routledge: London.

Greenfield, A. (2017), Radical Technologies: The Design of Everyday Life. Verso: London.

Katz, J., Floyd, J., Schiepers, K. (2023), Perceiving the Future through New Communication Technologies. Robots, AI and Everyday Life. Palgrave Macmillan: London.

Methnani, L. et al. (2021), "Let Me Take Over: Variable Autonomy for Meaningful Human Control". Frontiers in Artificial Intelligence, 4.

Rui Vieira da Cunha,
Technomoral change and the case for the AI alignment problem as a transformative experience

(MLAG (Mind, Language and Action Group) of the Institute of Philosophy of the Faculty of Arts of the University of Porto; Business School of the Catholic University of Porto)

The alignment problem in artificial intelligence, which concerns the challenge of ensuring that AI systems act in ways that are beneficial to humans, is often approached as a technical hurdle (Russell 2019). However, this paper posits that the true intractability of the problem lies less in its lack of objectivity and more in the dynamic nature of human values, which are continually reshaped by technological advancements (Vallor 2016).

The paper argues that as technology becomes deeply embedded in our lives, it not only serves as a tool but also actively shapes our understanding of the world and, consequently, our values. This fluidity of values in the face of technological change makes the static alignment of AI to a particular set of values a Sisyphean task (Bostrom 2014). In this regard, the insights of John Danaher on the process of technomoral change underscore the fluidity and evolutionary nature of morality in the face of technological progress (Danaher&Skaug Saetra 2022).

Moreover, the views of Ian Hacking (1999) on the looping effects of human kinds can be used in the context of AI and technology, where the categorizations and understandings propagated by these systems can lead to a recursive effect on human self-conception and values.

A further theoretical guide of the paper is the concept of transformative change, as introduced by L.A. Paul (2014). Paul's exploration of life-altering decisions, where the very act of undergoing an experience can change one's preferences and values, mirrors the challenges we face with AI and the possibility that they induce transformative changes in individuals and societies (Harari 2015), leading to shifts in values that are unpredictable and challenging to align with. Recognizing the intertwined nature of technology, values,

and transformative experiences is crucial for understanding the profound challenges of the AI alignment problem (Floridi 2013).

Keywords:

AI alignment problem, human kinds, looping effects, technomoral change, transformative experiences.

References:

Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press, 2014.

Danaher, J., and H. Sætra. "Mechanisms of Techno-Moral Change: A Taxonomy and Overview." Ethical Theory and Moral Practice 2023.

Floridi, Luciano. The Ethics of Information. Oxford: Oxford University Press, 2013.

Hacking, Ian. The Social Construction of What? Cambridge, MA: Harvard University Press, 1999.

Harari, Yuval Noah. Homo Deus: A Brief History of Tomorrow. New York: Harper, 2015.

Paul, L.A. Transformative Experience. Oxford: Oxford University Press, 2014.

Russell, Stuart. Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019.

Vallor, Shannon. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford: Oxford University Press, 2016.

Dilek Yargan,

To what extent are LLMs creative?

(Department of Philosophy, University of Rostock)

Suppose there was a chance to survey whether creativity in artificial intelligence is possible in a group that included people of all ages, both experts and non-experts in the field. In that case, there is no doubt that a remarkable and dramatic difference would be observed between the period before and after the introduction of ChatGPT-3. Of no surprise, there are a significant number of academic and non-academic mediums of communication that advocate the creativity of large language models (LLMs).

In this study, I want to question, from a philosophical perspective, whether the applications of these models are creative or not. For this purpose, I introduce the arguments against machine creativity based on two main points: Creativity is unique to humans among all living things, and an artifact cannot be creative; furthermore, machines only follow what their creators tell them to do. However, studies in philosophy, cognitive science, and psychology show that creativity arises from internal and external factors, which can pave the way for machine-generated creativity, regarded as a synthesis of diverse knowledge structures at varying levels.

Speaking of machine creativity requires discussing the parameters that can define it. Theoretical studies aligning with the potential for machine creativity show that there are common rules governing the creative processes of the human mind – rules that help develop cognitive approaches and research methodologies to foster machine creativity. Thus, in light of the common rules governing human creativity and the methodologies and approaches developed for machine creativity, I utilize a creative system approach for establishing machine creativity within the limits of machine intelligence.

Explaining this approach in this work, I discuss the philosophical principles of machine creativity. That is, machines that fulfill these principles should be considered creative. In the end, taking these principles into account, I analyze the features of LLMs, discuss whether they are agents, perhaps why they are not, and ultimately conclude whether LLM applications are creative.

About the Author

Dilek Yargan is a postdoc researcher in the project “Learning from Nature: Epistemological and Ontological Foundations of Biomimetics”, the Department of Philosophy, University of Rostock. https://www.iph.uni-rostock.de/mitarbeitende/homepage-dilek-yargan/