

Ethical Aspects of AI



INSTYTUT PROBLEMÓW
WSPÓŁCZESNEJ CYWILIZACJI

im. Marka Dietricha

Warszawa 2025

Instytut Problemów Współczesnej Cywilizacji im. Marka Dietricha, założony w roku 1996, jest międzyuczelnianą jednostką wykonującą zadania badawcze i edukacyjne dotyczące problemów współczesnej cywilizacji oraz podejmującą działania na rzecz integracji społeczności akademickiej. Aktualnie Instytut działa na podstawie porozumienia zawartego przez:

- Politechnikę Warszawską
- Szkołę Główną Gospodarstwa Wiejskiego w Warszawie
- Uniwersytet SWPS
- Uniwersytet Warszawski
- Warszawski Uniwersytet Medyczny

Działalność Instytutu jest nadzorowana przez **Kolegium Rektorów**, w skład którego wchodzi rektorzy uczelni – stron porozumienia.

Rada Instytutu

prof. Tomasz Borecki
prof. Andrzej Eliaz
prof. Marek Krawczyk
prof. Józef Lubacz – przewodniczący
prof. Zbigniew Marciniak
prof. Jan Szmidt

Od strony organizacyjnej IPWC jest jednostką Politechniki Warszawskiej.

Biuro Instytutu

ul. Koszykowa 75, lok. 43

00-662 Warszawa

tel.: +48 22 234 70 07

e-mail: instytut.ipwc@pw.edu.pl

www: ipwc.pw.edu.pl

Instytut Problemów Współczesnej Cywilizacji
im. Marka Dietricha

Ethical Aspects of AI

Edited by Paweł Stacewicz

Warszawa 2025

Ethical Aspects of AI

Wydanie I

Recenzenci:

Dr hab. Izabela Bondecka-Krzykowska

Dr hab. Adam Olszewski

© Copyright by Instytut Problemów Współczesnej Cywilizacji im. Marka Dietricha,
Warszawa 2025

Wydawca: Politechnika Warszawska

Oficyna Wydawnicza Politechniki Warszawskiej (UIW 48800)

ul. Polna 50, 00-644 Warszawa, tel. 22 234 70 83

Książka dostępna w wersji elektronicznej na stronach:

www.wydawnictwopw.pl; tel. 22 234 75 03; e-mail: oficyna@pw.edu.pl

www.ipwc.pw.edu.pl; tel. 22 234 70 07; e-mail: instytut.ipwc@pw.edu.pl

ISBN 978-83-8156-820-3 (druk)

ISBN 978-83-8156-821-0 (online)

Druk i oprawa: Drukarnia Oficyny Wydawniczej Politechniki Warszawskiej,
tel. 22 234 70 30

CONTENTS

Paweł Stacewicz

From the Editor	5
-----------------------	---

Roman Krzanowski, Paweł Polak

Machines Are Not Like Us: Arguing for the Gap	7
---	---

Andrzej Malec

On the Possibility of Ethical Autonomy of Artificial Intelligence	15
---	----

Bartosz Sawicki, Michał Śmiałek, Bartłomiej Skowron

Towards the Ultimate Programming Language: Trust and Benevolence in the Age of Artificial Intelligence	21
---	----

Jarosław A. Chudziak

Observations on the Use of Logic and Rationality within GenAI-Centric Multi-Agent Solutions	27
--	----

Krzysztof Sołoducha, Karol Narożniak

Simulation of AI Hybrid Ethics with Use of Multiagent Technology and Problem of Hidden Normativity	35
---	----

Paweł Łupkowski, Mariusz Urbański, Tomáš Ondráček

Non-Linguistic Question Feeders in Information-Seeking Dialogues and Public Debate Analysis	49
--	----

Bartłomiej Skowron, Daniel Myślicki, Jarosław Chudziak, Bartosz Sawicki

Understanding the Ethos of ChatGPT: An Open Question	57
--	----

Hans Joachim Greif

Ideals of Transparency in Artificial Intelligence and Philosophy	67
--	----

Marcin Rabiza

Decomposing Deep Neural Networks: A Mechanistic Perspective on Explainable AI	75
--	----

Łukasz Mścisławski

Knowledge, Wisdom and Morality of Thinking in the Age of (Gen)AI	83
--	----

Stuart Weinstein

What Does the US Federal Trade Commission (FTC). DoNotPay Enforcement Action Teach Us About AI Ethics and AI Legal Ethics	95
--	----

Marek Porzeżyński

Artificial Intelligence, Real Regulations – Can the Technological Revolution be Tamed?	107
---	-----

Radosław Siedliński

The Environmental Costs of AI Development as Ethical Costs	115
--	-----

Wojciech Głażewski

Three Scenarios for the Development of Superintelligence	123
--	-----

Authors	129
---------------	-----

Publications of the Marek Dietrich Institute of Contemporary Civilization Problems	133
---	-----

FROM THE EDITOR

PAWEŁ STACEWICZ

This book contains extended abstracts of selected presentations from the Ethics & AI Conference which was held as a joint event of the 9th Philosophy in Informatics Conference and the 19th ArgDiaP Conference. The event took place in September 2024 at the Warsaw University of Technology, Faculty of Administration and Social Sciences. One of its co-organizers was the Marek Dietrich Institute of Contemporary Civilization Problems. The head of the Institute Council prof. Józef Lubacz was a member of the Conference Programme Committee.

The content of this volume fits perfectly the current discussions on social consequences of research and applications in the area of artificial intelligence (AI) technology. Does artificial intelligence have to resemble humans, or, on the contrary, must it differ significantly from humans? What is the credibility of AI? Should we strive to create autonomous AI systems? How should AI systems explain their decisions and actions? Who should be responsible for harmful AI actions? How can the law be changed to stimulate the development of AI that would really be beneficial for humans in the short – and long-term perspectives? These are just few examples of the socially significant questions posed and considered by the authors of the texts included in this monograph*. We are only at the beginning of the discussion.

Although the texts are very diverse, their common point of reference is ethics: one could say *ethics for our times*, that is, ethics rooted in classical conceptions, but strongly oriented towards contemporary technological changes. Ethics understood in this way can and should serve as a moral signpost for the creators of AI and its foreseen applications.

* Some work reported in this book has been supported in part by the Polish National Science Centre, Poland (Chist Era IV) under grant 2022/04/Y/ST6/00001.

MACHINES ARE NOT LIKE US: ARGUING FOR THE GAP

ROMAN KRZANOWSKI, PAWEŁ POLAK

INTRODUCTION

Artificial Intelligence (AI) systems fundamentally differ from humans in their ethics, value systems, and conceptualizations of the world. This recognized gap raises significant uncertainties about how AI will interact with human values and priorities (see e.g., Bostrom, 2014; Marcus & Davis, 2019; Eagleman, 2020; Tolmeijer et al., 2020; Wooldridge, 2020; Kneer, 2021; Larson, 2021; Han et al., 2022; Pause Giant AI, 2023; Tolmeijer et al., 2022). Due to their distinct ethical frameworks, decision-making processes, and worldviews, AI systems can be considered a form of alien intelligence (Mollick, 2024). To achieve human compatibility (Russell, 2019), AI must share human moral grounding, ontology, and worldview. However, even with such alignment, if possible at all, AI remains inherently different from humans.

Humans and machines differ in nature without invoking theological explanations. It is crucial to avoid anthropomorphizing AI, as attributing human-like characteristics to machines is unjustified, misguided, and potentially dangerous. Instead, AI should be viewed as alien agents (defined below) with unique intelligence, objectives, and ethics, ensuring expectations are set safely and productively (Kornwach, 2021). This perspective helps in managing interactions with AI systems effectively, recognizing their differences rather than misleading similarities.

DEFINITIONAL CLARIFICATIONS

In this paper, specific terminologies are strictly defined to maintain clarity:

- **Alien:** Synthetic systems differing fundamentally from humans.
- **Incompatibility:** (in this essay) Differences in morality, values, and worldviews between human and synthetic agents.
- **Synthetic:** Artificial, not of natural origin.
- **AI Agent:** AI systems that interact with humans, enter the human social space, make decisions for humans, or serve in advisory roles, such as robots, software bots, or any AI system affecting humans.

THE FIVE ARGUMENTS FOR THE AI-HUMAN GAP

To substantiate the claim that AI systems are fundamentally different from humans, five distinct arguments are presented: the Mind Gap Argument, the Embodied Mind Argument, the Genetic Argument, the Leibniz Argument, and the Worldview Argument.

1. Mind Gap Argument

The Mind Gap Argument posits that even if human minds are fully explained by their physical structures (for Type A physicalism see e.g. Chalmers, 2010; Heil, 2010), no two physical systems are identical, especially when they possess different constituents.

Premises:

- (A1) Type A physicalism is true; any mind can be explained through its physical stratum.
- (A2) Different physical systems function differently, and even identical physical systems may exhibit different properties.
- (A3) The human mind's physical stratum differs from that of a synthetic mind.

Conclusion:

- (C1) Synthetic minds will function differently from human minds.

This inherent difference ensures that AI systems cannot replicate human cognitive processes exactly, maintaining a fundamental gap between human and synthetic minds.

2. Embodied Mind Argument

The Embodied Mind Argument asserts that the human mind is deeply integrated with our bodily functions, influencing our cognition and interactions with the world. AI systems, when integrated with synthetic bodies, will possess different physical structures and, consequently, different cognitive processes (for more about the failings of current AI conceptualizations, see, for example, the works of Brooks [1991], Dreyfus [2016] and Smith [2019]).

Premises:

- (A1) The human mind is integrated with human bodily functions, so it has been tuned to the operation of the body.
- (A2) Synthetic minds will be integrated with synthetic physical structures, such as artificial bodies.
- (A3) Synthetic physical structures will differ from human bodies.
- Conclusions:
- (C1) Synthetic minds will be tuned to different physical structures than human minds.
- (C2) Synthetic minds will differ fundamentally from human minds.

This embodiment difference means that synthetic minds are attuned to different physical realities, further widening the cognitive and functional gap between humans and AI.

3. Genetic Argument

The Genetic Argument emphasizes that the human mind evolved through specific environmental pressures, natural selection, unique sensory apparatus, and adaptive processes aimed at human survival. Synthetic minds, however, are programmed with purposes unrelated to natural survival and lack the evolutionary history that characterizes human cognition.

Premises:

- (A1) The human mind emerged due to specific environmental pressures, natural selection, human bodily capacities, constraints, and sensory apparatus.

- (A2) The human mind evolved specifically to support human survival.
- (A2.1) The evolutionary history of the human mind cannot be recreated in a synthetic environment due to its largely unknown nature (its evolutionary history).
- (A3) Synthetic systems do not experience evolutionary pressures as natural systems do.
- (A4) Synthetic minds are programmed for purposes other than survival in nature.

Conclusion:

- (C2) Synthetic minds differ fundamentally from human minds.

This divergence ensures that synthetic minds operate on principles distinct from those that shaped human cognition, maintaining a clear separation between human and AI mental processes.

4. Leibniz Argument

Based on Leibniz's principle of the Identity of Indiscernibles, the Leibniz Argument contends that two systems are identical only if all their properties match. Human minds possess distinct features that synthetic minds, built with different architectures and functions, do not share.

Premises:

- (A1) The human mind has a set of features: $MH = f(x_1, \dots, x_n)$.
- (A2) A synthetic system has a different set of features:
 $MS = f(k_1, \dots, k_m)$.
- (A3) Two systems are identical only if all their features match.
- (A4) At least one feature x_i of MH differs from a feature k_j of MS.

Conclusion:

- (C) MH is not the same as MS; thus, a synthetic mind must be different from a human mind.

This argument reinforces the notion that synthetic minds, by virtue of differing properties, cannot be identical to human minds, further cementing the fundamental gap between them.

5. Worldview Argument

The Worldview Argument posits that different sensory and information-processing capacities lead to distinct worldviews (e.g., Yong, 2022). AI systems,

equipped with unique sensors and processing mechanisms, develop their own “philosophies of life” that differ fundamentally from human perspectives.

Premises:

- (A1) Every being with a sensory system forms its own worldview based on its unique sensory abilities and environment.
- (A2) Different beings have different worldviews.
- (A3) AI systems with synthetic sensory systems will perceive and process information differently from humans.

Conclusions:

- (C1.1) AI systems will possess worldviews distinct from human worldviews.
- (C1.2) AI systems will be alien in this sense.

This difference in worldview makes AI systems appear alien, as their interaction with the world and humans is based on fundamentally different principles, ensuring that AI remains distinct from human cognition and perception.

DISCUSSION

The five arguments collectively are intended to demonstrate that AI systems and humans occupy fundamentally different ontological and conceptual spaces. These differences underscore the necessity of treating AI as alien intelligence rather than human-like entities. While efforts can be made to align AI systems closer to human values and ethics, the intrinsic nature of technology ensures that complete compatibility remains unattainable.

PHILOSOPHICAL IMPLICATIONS

Philosophical speculations about AI-human similarities should remain within theoretical discourse without extending to practical claims about AI capabilities. Over-attributing human characteristics to AI can lead to unrealistic expectations and potential misuse of technology (See e.g. Suleyman, 2023). Instead, recognizing and respecting the fundamental differences between humans and AI fosters safer and more productive interactions.

The extent to which different aspects of humans and human behavior can be approximated remains a separate research question. Systematic studies of criteria for evaluating imitation of human capabilities and a conceptual framework for analyzing differences and similarities are needed. Indeed, the Turing test is not sufficient to capture the existential complexity of being human, and broader research is needed (see on Modern Turing Tests Suleyman 2023).

PRACTICAL IMPLICATIONS

AI systems should be managed with the understanding that they possess unique forms of intelligence, objectives, and ethical frameworks. This perspective aligns with calls for responsible AI development, emphasizing ethical standards and governance without imposing unnecessary restrictions based on unfounded similarities. By acknowledging AI as fundamentally different, society can better navigate the complexities of integrating advanced AI systems into human-centric environments.

CONCLUDING REMARKS

Ultimately, the coexistence with AI depends on our ability to recognize and respect these differences, ensuring that AI serves as a beneficial tool without overstepping into realms reserved for human cognition and morality. The divergence between human and synthetic systems is not just a theoretical concern but has practical implications for how we develop, implement, and interact with AI technologies. Embracing AI's unique nature while striving for responsible alignment with human values can lead to a harmonious and productive coexistence, mitigating risks associated with misaligned expectations and ethical misunderstandings.

FINAL CONCLUSION

Due to several fundamental differences between humans and synthetic systems, we may risk a claim, AI systems will never be like us, despite challenges from proponents aiming for closer alignment or even exact replication.

As highlighted by Altman (2024), adapting to AI technology requires adjusting our perspectives and management strategies rather than expecting AI to mirror human ethics, values, and worldviews fully. The best approach is to treat AI systems as alien artificial agents with varying goals, values, reasoning processes, and moral frameworks. While partial alignment in specific dimensions is achievable, complete compatibility, and with it predictability, remains out of reach (see the discussion of AI future in Suleyman 2023). This perspective not only fosters safer interactions but also promotes a realistic understanding of AI's capabilities and limitations, ensuring that technological advancements benefit humanity without compromising ethical integrity.

REFERENCES

- Altman, S. 2024. Sam Altman of Future of AI. 2024. Available at <https://www.youtube.com/watch?v=2crVJjXA7ZE>. Accessed on 05.09.2024.
- Bostrom, N. 2014. *Superintelligence*. Oxford: Oxford University Press.
- Brooks, R.A. 1991. Intelligence without representation. *Artificial Intelligence* 47 (1991), 139–159.
- Chalmers, D. 2010. *The Character of Consciousness*. Oxford: Oxford University Press.
- Dreyfus, H.L. 2016. *Skillful Coping*. Oxford: Oxford University Press.
- Eagleman, D. 2020. *Livewired*. Edinburgh: Canongate.
- Han, S., Kelly, E., Nikou, S. et al. 2022. Human versus Artificial Intelligence. *AI & Soc* 37, 1383–1395 (2022). <https://doi.org/10.1007/s00146-021-01247-4>
- Heil, J. 2010. *The philosophy of Mind*. London: Routledge.
- Kneer, M. 2021. Can a robot lie? *Cognitive Science* 45 (10), doi: 10.1111/cogs.13032
- Kornwach, K. 2021. Transhumanism as a Derailed Anthropology. in W. Hofkirchner, H.-J. Kreowski (eds.), *Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea?*, *Cognitive Technologies*, https://doi.org/10.1007/978-3-030-56546-6_2
- Larson, E. 2021. *The Myth of Artificial Intelligence*. Cambridge: The Belknap Press of Harvard University Press.
- Marcus, E. Davis, E. 2020. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *The MIT Technology Review*. Available at <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>. Accessed on 10.12.2021.
- Mollick, E. 2024. *C-Intelligence*. London: WH Allen.
- Pause Giant AI., 2023. *Pause Giant AI Experiments: An Open Letter*. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

- Russell, S. 2019. *Human Compatible. AI and the problem of control*. London: Penguin Books.
- Smith, B. 2019. *The Promise of Artificial Intelligence*. Cambridge: The MIT Press.
- Suleyman, M. 2023. *The Coming Wave*. London: Vintage.
- Tolmeijer, S., Kandul, S., Kneer, M., Christen, M., Bernstein, A. 2022. Capable but amoral? Comparing AI and human team members in moral decision-making contexts. *ACM CHI Conference on Human Factors in Computing Systems*, 160: doi: 10.1145/3491102.3517732
- Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A. 2020. Implementations in Machine Ethics: A Survey. *ACM Computing Surveys* 53 (6), 1–38. doi: 10.1145/3419633
- Yong, E. 2022. *An Immense World*. London: The Bodley Head.
- Wooldridge, M. 2020. *The road to conscious machines: The story of AI*. London: Penguin Books Limited.

ON THE POSSIBILITY OF ETHICAL AUTONOMY OF ARTIFICIAL INTELLIGENCE

ANDRZEJ MALEC

INTRODUCTION

Artificial Intelligence systems are created and evaluated through the prism of the concept of natural (human) intelligence. Thus they are usually modeled on natural intelligence, and assessed from the perspective of capabilities that humans possess, including memory, reasoning, problem-solving, adaptation, etc. Here, the same approach is used to consider the possibility of ethical autonomy of Artificial Intelligence, i.e. its potential ability to independently choose goals. Therefore, this paper highlights several issues that are important when the question of human conscience autonomy is considered, and asks about the equivalents of these issues for Artificial Intelligence. These will be:

1. the role of Leibniz's formula ("it is obligatory to choose the act which is the best of possible acts") in the evaluation of acts,
2. the problem of basic values choice,
3. the problem of elementary acts evaluation,
4. the problem of complex acts evaluation,
5. the problem of evaluation of consequences of an act.

The paper refers to the author's analysis of utilitarianism in [2] as well as to the formal apparatus of [3] inspired directly by [7] and indirectly by [5].

LEIBNIZ'S FORMULA AS THE BASIS OF CONSCIENCE AUTONOMY

There are three basic ways in which people can autonomously evaluate acts: moral instinct, moral intuition, and calculation of the value of the act [2].

The moral instinct, that is, human ability to evaluate an act based on simple attraction and repulsion, is not always accurate. It is also variable: in the past, sentencing a little boy to death for stealing a piece of bread was not repulsive, and the Spartans threw their own children off a cliff because they were not sufficiently "attractive". By assessing an act using the moral instinct, a person is similar to a robot executing simple conditional instructions ("If A, do B").

In turn, the moral intuition, that is, human ability to evaluate an act based on the interaction of consciousness and unconscious brain work, often does not cope with non-standard problems: such as those presented in many versions of the trolley dilemma. When evaluating an act using moral intuition, a person is somewhat similar to a robot acting in accordance with a rule created as a result of machine learning.

Further, calculating the value of an act, i.e. evaluating an act based on Leibniz's formula: "it is obligatory to choose the act which is the best of possible acts", encounters a number of technical problems (the problem of basic values choice, the problem of elementary acts evaluation, the problem of complex acts evaluation, the problem of evaluation of consequences of an act), as well as the is-ought problem raised by David Hume [1]. According to Hume, moral distinctions cannot be derived from reason. By evaluating an act based on Leibniz's formula, a person acts similarly to a computer executing an elaborate algorithm.

In fact, the first two ways of evaluating acts are not fully autonomous for humans: both moral instinct and moral intuition are the result of human biological evolution and socialization. Their counterparts for evaluating acts by robots are even less autonomous: of course, acting on the basis of a simple conditional instruction ("If A, do B") is not autonomous; similarly, the results of machine learning depend significantly on the material on which the network was trained.

And what about the third way? Let us first look at the basic problem, that is, the possibility of evaluating acts by means of reason. The is-ought problem

may be understood in two ways. First, it may be understood in accordance with Hume's original statement, and then it is about the lack of logical values in normative reasonings. Second, it may be understood in the contemporary way, and then it is about the logical separation of the deontic sentences from the sentences about facts [6]. However, the analysis based on contemporary semantics leads to the conclusion that Hume's thesis is false in both its original and contemporary understanding. First, semantics shows that the truth of statements about facts and the truth of statements about obligations differ only in the sets on which we verify them [3]:

'FACT (a_m)' is true $\Leftrightarrow \varphi(a_m) \in \mathbf{FACT}$,

'OBL (a_m)' is true $\Leftrightarrow \varphi(a_m) \in \mathbf{OBL}$.

Clearly, truth of the sentence 'Kate kissed John' depends on the set **FACT** and truth of the sentence 'It is ordered that Kate kissed John' depends on the set **OBL**, and these sets are different. Nevertheless, the sentence 'It is ordered that Kate kissed John' has the same right to be true or false as the sentence 'Kate kissed John', when Popper's third world is considered. Moreover, it is reasonable to call into question the truth of Hume's thesis also in its contemporary understanding: this is supported, for example, by the principle already accepted in Roman times: *Impossibilium nulla obligatio est* – what cannot become a fact, cannot also be the subject of duty [4].

Since moral statements can be true (or false), evaluating acts based on Leibniz's formula seems to be possible. Let us look at how humans cope with this calculation, and also at whether a robot can be equally autonomous in this work.

BASIC VALUES CHOICE

Both humans and robots can choose instrumental values based on the claims of natural science. They can therefore make statements such as: "If you want to live, value your health!", "If you want to be healthy, value medicine!", "If you want to live in society, value taxes!", etc. Moreover, the robot will probably do this more efficiently, because it will calculate the consequences of the claims of natural science more quickly and more broadly.

Moreover, both humans and robots can accept basic values as values found in reality. Among the found values are those protected by the law: human life (protected by the norm prohibiting killing), human health (protected by the norm prohibiting causing health disorders), human freedom (protected by the norm prohibiting deprivation of liberty), human honor (protected by the norm prohibiting slander), property (protected by the norm prohibiting theft), and so on. Interestingly, a robot could also adopt as basic values, values protected by norms addressed to robots, e.g. Asimov's Robot Laws, or David Langford's Robot Laws.

However, a human can choose basic values themselves: "I choose life as a value", "I choose society as a value", etc. The justification for such choices is the biological and social nature of a human. On the other hand, a robot choosing basic values is deprived of such justification.

ELEMENTARY AND COMPLEX ACTS EVALUATION

An act is good when, in a situation of elementary choice, that is, the choice between two atomic situations in the sense of Wolniewicz [7], the situation that better preserves basic values is chosen. When I say: "Murder is bad" I mean that in a situation of elementary choice, it is better not to kill than to kill. Knowing basic values, a human and a robot are able to recognize good and evil in a similar way.

However, when the evaluation of an act requires weighing different values, that is, deciding not about what is good and what is bad, but about what is right and what is wrong, the matter is different. When the choice consists, for example, in saving John from a fire at the cost of Peter's life, it is not enough to know that human life is a value. One must also know whose life is more valuable: John's or Peter's. In such situations, both humans and robots can refer to a hierarchy of values, which sometimes allows for a choice between alternative situations. Moreover, both humans and robots, based on the claims of natural sciences and basic values, can create hierarchies of instrumental values: "health is more valuable than property", "a ton of wheat is more valuable than a kilogram of wheat", etc.

However, humans can choose such a hierarchy of values for basic values: "life is more valuable than dignity", "dignity is more valuable than life",

etc. The justification for such choices is the biological and social nature of humans. On the other hand, a robot choosing basic values is deprived of such justification.

EVALUATION OF CONSEQUENCES OF AN ACT

When assessing the value of an act, it is impossible to avoid assessing the value of its consequences. An act that is right in the perspective of a week may be wrong in the perspective of a month, and right again in the perspective of a year, etc. The balance of profits and losses of a decision to reform the social security system may be advantageous in the perspective of a year, because it consolidates the ruling party, may be disadvantageous in the perspective of four years, because it leads to defeat in the next parliamentary elections, and in the perspective of twenty years – advantageous again, because the system will achieve the assumed economic parameters.

When assessing the value of an act, one should therefore also look into the future. But how far into the future should one look? Is the perspective of the next parliamentary elections sufficient? Or maybe the life of our generation? Or maybe a few more generations? The answer to this question depends in part on how precisely the consequences of the act can be calculated. And in this respect, a robot will probably do better than a human, because it will apply the theorems of natural sciences to given facts more quickly and more widely. However, the decision about how far into the future to look also depends on the assessment of the impact of time on the erosion of values. A person can make such a decision themselves, for example by adopting a sustainable development strategy. Once again, the justification for such a choice is the biological and social nature of a human being. On the other hand, a robot choosing the limits of looking into the future is deprived of such justification.

The assessment of an act from the point of view of its consequences is also influenced by the risk-taking strategy: when the world is not deterministic or when we do not know the laws that govern it sufficiently, the consequences of the act are not certain for us. When choosing a course of action, people are therefore also guided by different attitudes to taking risk. They can choose different risk-taking strategies. The justification for such a choice is not only

the biological and social nature of a human being, but also their individual nature. On the other hand, robots choosing risk-taking strategies are deprived of such justification.

CONCLUSIONS

Although the robot may be more technically efficient than the human in calculating the value of an act according to Leibniz's formula, in such areas as choosing basic values, choosing a hierarchy of basic values, choosing a horizon for weighing values, or choosing a risk-taking strategy, the robot seems to be dependent on the human for now. In such choices, the human is aided by his biological and social nature, which the robot is currently deprived of.

REFERENCES

- [1] Hume, D., *A Treatise of Human Nature*, Penguin Classics 1985.
- [2] Malec, A., *Arytmetyka sumienia. Przyczynek do teorii odpowiedzialności słusznościowej* (in Polish: *The Arithmetic of Conscience. A Contribution to the Theory of Equitable Responsibility*), Warszawa 2012.
- [3] Malec, A., *Introduction to the Semantics of Law*, Springer 2022.
- [4] Malec, A., *Wittgenstein versus Hume: the Is-Ought Problem – Finally Solved*, *Studies in Logic, Grammar and Rhetoric* (Volume 69), 2024.
- [5] Wittgenstein, L., *Tractatus logico-philosophicus*, Routledge, London 1995.
- [6] Woleński J., *Uogólniona teza Hume'a* (in Polish: *A Generalized Hume's Thesis*), in: I. Bogucka, Z. Tobor (editors) *'Prawo a wartości. Księga jubileuszowa Profesora Józefa Nowackiego'*, Zakamycze, Kraków 2003.
- [7] Wolniewicz, B., *Ontologia sytuacji* (in Polish: *The Ontology of Situations*), Państwowe Wydawnictwo Naukowe, Warszawa 1985 (see: Bogusław Wolniewicz, *A Formal Ontology of Situations*, *Studia Logica* (41), 1982).

TOWARDS THE ULTIMATE PROGRAMMING LANGUAGE: TRUST AND BENEVOLENCE IN THE AGE OF ARTIFICIAL INTELLIGENCE

BARTOSZ SAWICKI, MICHAŁ ŚMIAŁEK, BARTŁOMIEJ SKOWRON

HOW THE SOFTWARE IS DEVELOPED?

A computer program is classically presented to its users as a mysterious box that processes input data to produce the expected output. This simplification is a good starting point as it highlights the crucial role of input data and the fact that the processing mechanism is not important as long as the output meets expectations. However, it is also necessary to describe the methods of communication with the user or other cooperating systems, as well as to build an internal data model. This is the basis for the dominant multi-layered architectures today, such as the Model-View-Controller or Model-View-Presenter patterns.

Software is created by people and ultimately for people [2]. They use programming languages to describe the desired way of processing data. However, over the past 70 years, the way programs are defined has changed dramatically. Programming languages have undergone a significant transformation from machine code and assembly language to modern third generation (3G) languages, which offer high-level abstraction to facilitate the description of complex systems.

It should also be emphasized that a program is the final product of the entire software engineering process, which begins with the formulation of

often complex requirements and the design of basic system components. Requirements are traditionally formulated in natural language, although various formal methods (controlled language, graphic models) are sometimes used. Requirements describe both the expected behavior of the system and the data structures processed by the system. In simplified terms, the entire software development process can be seen as transforming requirements into working machine code (a program) that meets these requirements.

HOW DOES AI CREATE CODE?

Today, generative artificial intelligence systems can create software based on short descriptions formulated in natural human language. They rely on vast repositories of source code and other documented knowledge, which have been used to train an artificial neural network containing trillions of parameters [1].

However, the challenge lies in the fact that descriptions in natural language are ambiguous, just like the language used by humans. At the same time, models built on statistical foundations do not guarantee deterministic behavior. Therefore, there is no guarantee that the result will always meet expectations.

AI methods can generate a complete computer program based on a fairly general query in natural language. This means that all algorithmic components, the internal data model, and the user interface view are generated [5]. However, it seems that this approach is not optimal. It should be noted that a typical requirements specification includes dozens of pages of descriptions of functional units, domain concepts (glossary), domain models, quality requirements, and other necessary system features. The question arises as to how precise and unambiguous such a specification must be for the AI system to reliably produce a program that meets user expectations.

WHAT IS THE ROLE OF PROGRAMMING LANGUAGE?

In the compilation process, a program's source code is translated into machine code tailored to the processor architecture on which it is executed. Simultaneously, it serves as a crucial communication tool between programmers.

The ability to understand, control, and correct the way a program operates builds trust in the computer program and facilitates collaboration.

A longstanding problem is that the system specifications created by users are highly variable. People often cannot predict what outcome will be best, and external circumstances change as well. This necessitates that the software development process allows for modifications and improvements to the source code. Iterative software development methodologies like Scrum respond to this need, enabling users to regularly influence the appearance and functionality of the software being developed [7].

The reliability of today's processing methods largely stems from established data models on which they are based. A consistent data model allows several independent applications to operate on them and mutually control the correctness of their operations.

This raises the question of what programming will look like in the era of strong artificial intelligence. It seems we are approaching the ultimate generation of programming languages, where the source code will encapsulate the essence of user expectations for the system [8].

TRUST, RELIABILITY, AND BENEVOLENCE OF A SOFTWARE

Let us assume for a moment that a software user is a citizen of a certain country, and the software is a set of regulations (laws) according to which public institutions of that country operate. When does a citizen trust a public institution? When the institution is reliable. If it is not reliable, the citizen questions its competence, the safety of the area under its jurisdiction, and doubts the public good that the institution should uphold. When a citizen does not trust, they are not only critical of the institution but also do not support its operations and sometimes even oppose it. When they trust, the chance for mutual cooperation increases. Social studies indicate that the reliability of public institutions is built on factors such as benevolence, competence, and integrity of state institutions [6]. The counterparts of integrity and competence in computer programs are the aforementioned abilities to understand, control, and correct the way a program operates. Nevertheless, extending this analogy further, can software be benevolent?

Aristotle defined benevolence from one person to another not as politeness, being nice, smiling, and cordiality, as we spontaneously tend to think, but as understanding what is good (what contributes to the flourishing and good life of that person) and striving to achieve that person's good. A benevolent person wishes well for another person and acts for the good of the other person. Mutual benevolence is a necessary condition for friendship. To be benevolent, I must understand the other party's interest and actively want to help realize that interest—mere observation is not enough. A benevolent state institution is one that sees the citizen's interest and is ready to actively work for the citizen's interest.

The potential evaluation of a program's benevolence, using this analogy, is the assessment of how well the software can understand the user's and the social group's good, of which the user is a part, and work to achieve that good. In other words, a benevolent AI is one that possesses a semblance of good will. Notice that when a certain person always meets our needs, meaning their behavior fully corresponds to our needs, it does not necessarily mean that we trust that person. On the contrary, we often become suspicious of them. We do not become suspicious of all people who make mistakes that work against us, but of those who make such mistakes and wish us ill. If someone wishes us well, we will more easily forgive them for a mistake, even if that mistake is severe for us. This is the trust-building power of benevolence. Similarly, with a computer program: if it makes mistakes, we still trust its operation if it is benevolent—because if it wishes us well, it will sooner or later actively correct its error. Therefore, the benevolence of AI is something that can enhance trust in even an imperfect AI.

VISION OF THE ULTIMATE PROGRAMMING LANGUAGE

The need for trust is a critical element of cooperation between humans, and therefore also between humans and artificial intelligence systems. It seems that a programming language must provide humans with the ability to understand the method of operation while eliminating all unnecessary implementation details. Moreover, we propose that the program should, as far as possible, be benevolent to potential users by adequately recognizing their interests and actively supporting them.

Imagining further with the analogy to state structures, we envision that future source code will be an essential description akin to today's legal language, which will describe methods of data processing. Additionally, there will be a description of key internal data structures, enabling external audits of data integrity. The user interface is an artifact easily verifiable by humans, so it seems possible to give the AI model greater freedom in this area.

Three fundamental issues can be distinguished here: 1) how to formulate requirements so that AI understands the needs well and can be benevolent to the user; 2) what the programming language should look like to allow humans to easily modify what the AI has generated; 3) what language should be compiled and executed by computers. This raises the question of whether we will still need 3G programming languages.

Our vision includes three levels at which the AI-assisted software development process would take place. At the first level, a natural language with specified guidelines would be used. It would serve to define the problem in a way convenient for humans (functional requirements, domain concept glossary, data descriptions, quality requirements, definition of the well-understood interests of the user group). This language could be supported by various patterns, templates, or graphical notations, as well as ethical elements. Next, at the second level, strong AI language models would translate this description into a new system definition language. This language would be precise, unambiguous, and complete. It would be fully understandable to humans, yet also capable of being automatically and deterministically transformed into a traditional program in a 3G language.

At the second level, a dialogue between AI and humans would be possible - modification of the specification and adaptation to changing needs. This can be seen as a significant extension to the model-driven and low-code software development paradigms [7, 9]. The translation from the second to the third level should be deterministic. The third level would enable code optimization in situations where, for example, system performance requires it.

The language placed at the second level of the above description we can call the Ultimate Programming Language. If humans want to maintain control over the operation of machines, further reduction of complexity seems impossible.

REFERENCES

- [1] Eduardo, C., Coello, A., Nazeh Alimam, M., Kouatly R. Effectiveness of ChatGPT in Coding: A Comparative Analysis of Popular Large Language Models. *Digital* 4.1, 2024: 114–125.
- [2] Gerosa, M., Trinkenreich, B., Steinmacher, I., Sarma, A. Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering* 31.1, 2024: 13.
- [3] Qi G., et al. Exploring the potential of chatgpt in automated code refinement: An empirical study. *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering* 2024.
- [4] Rasheed, Z., Waseem, M., Kemell, K-K., Xiaofeng W., Nguyen Duc A., Systä, K., Abrahamsson, P. Autonomous Agents in Software Development: A Vision Paper, *arXiv preprint arXiv:2311.18440*, 2023.
- [5] Xu, F.F. et al. A systematic evaluation of large language models of code. *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 2022.
- [6] Esposti, D., Ball, S.K., Dibb, S.. What’s in it for us? Benevolence, national security, and digital surveillance. *Public Administration Review* 81.5, 2021: 862–873.
- [7] Śmiałek, M., Nowakowski, W., From Requirements to Java in a Snap, Model-Driven Requirements Engineering in Practice. Springer, 2015.
- [8] Brooks, F.P. No Silver Bullet: Essence and Accidents of Software Engineering. *IEEE Computer*, vol. 20, no. 4, 1987: art. no. 4.
- [9] Hirzel M. Low-Code Programming Models. *Communications of the ACM*, vol. 88, no. 10, 2023: 76–85.

OBSERVATIONS ON THE USE OF LOGIC AND RATIONALITY WITHIN GENAI-CENTRIC MULTI-AGENT SOLUTIONS

JAROSŁAW A. CHUDZIAK

INTRODUCTION

In recent years, the fast progress on large language models (LLMs), such as ChatGPT, Gemini, and Claude, was used to create innovative approaches to designing LLM-centered multi-agent systems (MAS). By incorporating additional components, like logical inference, reflections and planning, these systems can achieve more consistent and efficient interactions with an environment and better communication and collaboration among agents if this is required. At starting point, LLM-based multi-agent platforms (like AutoGen, CrewAI, and LangChain) are able to handle increasingly complex tasks through advanced natural language understanding. However, as these agents evolve, the integration of much more advanced logical reasoning systems becomes indispensable for ensuring reliability, adaptability, and resilience. This study discusses some of the roles played by logical subsystems in structuring GenAI-based MAS inter- and intra- architectures, and emphasizes the need to carefully align the capabilities of agents realized by their symbolic (logical, rational) and connectionist components.

In modern LLM-based autonomous agents, their skills and capabilities are typically structured across four key components: Profile, Memory, Planning, and Action. We can briefly define them as:

- Profile defines the generation strategy, which may include handcrafted methods, GPT-generated methods, or dataset alignment methods. It also contains critical profile contents such as demographic, personality, and social information, which influence how the agent interacts in various contexts.
- Memory plays a crucial role in the agent’s functioning, encompassing both long-term and short-term memory structures. It supports different formats, such as language embeddings, databases, and lists, allowing the agent to recall information efficiently and adaptively.
- Planning is divided into tasks performed with or without feedback. Without feedback, agents employ subgoal decomposition, multi-path thought, and external planning tools to devise strategies. With feedback, agents incorporate environmental, human, and model input to refine their plans dynamically, finally
- Action component targets diverse tasks such as dialogue, exploration, or task completion. The agent’s action strategies involve interaction with tools, feedback loops, and recollection, enabling flexible responses to both internal and external changes.

Together, these components may form specialized architecture that enables LLM-based agents to autonomously operate in complex, real-world environments. They can efficiently and effectively leverage both predefined structures and adaptive feedback mechanisms.

REASONING WITH LOGIC

The main logical systems, such as classical logic, modal logic, non-monotonic reasoning, and default logic, are fundamental in enabling LLM-based agents to perform accurate and efficient reasoning [1, 4, 6, 11]. Classical logic may provide a foundational basis for clear and unambiguous reasoning, allowing agents to derive conclusions from a set of premises with certainty.

However, the dynamic and often unpredictable environments in which LLM-based MAS operate necessitate more flexible solutions. As an example solution, modal logic may provide the capacity to handle possibilities and

necessities. They are essential for agents operating under varying degrees of uncertainty [10]. Then, non-monotonic reasoning and default logic may further augment this capability by allowing agents to revise their beliefs and adapt their actions in response to new information. In the same time, they will maintain coherence in the face of changing environments [9, 10].

Epistemic logic may play a critical role in MAS by enabling GenAI-based agents to reason about knowledge and beliefs [5]. This logical framework allows agents to consider not only their own knowledge but also the knowledge and beliefs of other agents. This is crucial for effective coordination and collaboration, as agents must often make decisions based on incomplete or uncertain information. The use of epistemic logic ensures that agents can manage and utilize shared knowledge. This capability is enhancing the overall functionality and efficiency of the whole agentic ecosystem.

Described above logical skills/capabilities help GenAI-based agents in their MAS environments to utilize various forms of reasoning to navigate their environments and make or explain their decisions. In this context, deductive reasoning allows agents to derive specific conclusions from general principles, ensuring that actions are logically sound based on known premises. Inductive reasoning enables agents to generalize from specific observations, allowing them to form hypotheses and adapt to new patterns in dynamic environments. And finally, abductive reasoning, or inference to the best explanation, helps agents generate plausible hypotheses to explain observed phenomena. This type of reasoning facilitating problem-solving, hypothesis generation in uncertain scenarios or explain structure of reasoning or action taken by agent(s) [11]. Together, various reasoning methods provide a baseline for a comprehensive competency toolbox for LLM agents to handle a wide range of situations effectively.

While logical reasoning provides a structured and interpretable (white-box, glass-door) framework for decision-making, other reasoning methods such as deep learning, reinforcement learning, and evolutionary algorithms play important, even very often primary role in GenAI MAS solutions. These sub-symbolic (connectionist) models (often from the field of generative artificial intelligence) act as black boxes, lacking the transparency and explainability of symbolic logical systems. But their advantage is high efficiency, adaptability and vary often autonomous mode of working. But we need to

notice, that by integrating symbolic, logical (white-box) reasoning with these sub-symbolic (back-box, glass-door) techniques can significantly enhance the overall performance of MAS, and enhance their strong interpretability and adaptability features [11].

DISTRIBUTED COGNITION

The Theory of Mind, which involves attributing mental states to oneself and others, is another important aspect in MAS design. Incorporating the theory of mind into MAS allows LLM-based agents to predict and interpret the actions of other agents based on their presumed beliefs, desires, and intentions. This capability is essential for creating more sophisticated and human-like interactions among agents, facilitating smoother cooperation and conflict resolution.

In the context of GenAI-based Multi-Agent Systems (MAS), the concept of distributed cognition, offers a powerful framework for better understanding how cognitive processes are distributed and managed across agents, artifacts, and environments. Distributed cognition suggests that intelligence and decision-making are not confined to individual agents but are instead spread across a system of interacting elements. In LLM-centered multi-agent systems, this means that cognition can be shared and enhanced through communication and coordination between agents. Each agent contributes part of the cognitive workload, relying on the collective intelligence of the group to solve complex problems more efficiently than any individual agent could on its own.

By adopting a distributed cognition perspective, GenAI MAS can better manage the flow of information, optimize decision-making processes, and adapt to changing environments through a system-wide cognitive synergy, where each agent not only acts based on its own capabilities but also leverages the knowledge and actions of others in the network. This approach opens new possibilities for enhancing cooperation, flexibility, and robustness in multi-agent environments.

RATIONALITY

The way LLM-centric agents are designed often starts with a framework for how the agent will be rational [3, 6, 7, 8]. The choice of how we define rationality is very important, because it shapes how these agents process information and make decisions. There are several different kinds of rationality, each offering a unique approach to agent-based interactions. Let's look at some of them:

- Philosophical Rationality emphasizes logical consistency. This makes it well-suited for systems that rely heavily on rule-based decision-making.
- Economic Rationality is about behaviors that maximize utility, which is particularly relevant in competitive or resource-constrained environments.
- Bounded Rationality acknowledges that agents have limits in terms of the information and cognitive resources they have. This can be beneficial for agents operating in dynamic situations with incomplete information.
- Epistemic Rationality guides agents toward seeking truth.
- Practical Rationality ensures that agents focus on achieving their goals efficiently.

Each of these definitions has strengths and weaknesses, depending on what the system is designed to do. Therefore, it's crucial to choose the right one for the specific context in which the GenAI MAS will be used. In practical terms, the chosen framework influences how an agent perceives its environment, how it sets goals, and how it chooses actions to achieve those goals. For instance, an agent operating under economic rationality might prioritize actions that maximize its rewards in a marketplace, even if those actions appear illogical from a philosophical standpoint.

Conversely, an agent designed with epistemic rationality might prioritize gathering and verifying information, even if that delays immediate goal achievement. The rationality framework also affects the agent's learning mechanisms. An agent designed with bounded rationality, for example, might employ simpler learning rules or heuristics, whereas an agent with philosophical rationality might use more complex, logic-based learning methods.

PRACTICAL CASE STUDY

Building on the basic reasoning capabilities of multi-agent systems (MAS), we have explored practical MAS use cases [2]. In our experiments we demonstrated the practical application of MAS in software engineering and system design. In this simulated software development environment, virtual cognitive agents take various roles which exist in the most of real-life software teams. These include roles such as analysts, designers, programmers, testers, and project managers. It shows a trend in which the new introduced agentic paradigm automates, in the software engineering domain, key tasks traditionally performed by humans. The goal is to assess the agents' efficiency in managing software projects by utilizing advanced cognitive architecture, generative AI technologies, and rational decision-making frameworks.

Logic and rationality play a crucial role in this scenario, guiding agents' decision-making processes to ensure coherent, goal-oriented behavior. By leveraging structured reasoning and rationality principles, agents can prioritize tasks, optimize resource allocation, and handle uncertainties during project execution. Our simulation, which mirrors real-world development dynamics, shows how MAS can significantly change and improve processes like system design, project management, and continuous integration and testing. This case study provides a glimpse into the potential for MAS to enhance productivity and consistency in large-scale software projects, offering a practical approach to reducing cognitive load and streamlining complex workflows.

ETHICS

And finally, ethics, which should play a crucial role in the design and deployment of GenAI-based agents within MAS. Logical frameworks can incorporate ethical, normative principles to guide agent behavior, ensuring decisions are made in alignment with human values and societal norms [5]. The transparency of white-box logical systems aids in auditing and understanding the decision-making process, which is essential for accountability and trust.

Conversely, the opacity of black-box sub-symbolic methods presents challenges in ethical assurance, as the reasoning behind decisions may not

be easily interpretable. Addressing these ethical issues requires a balanced approach based on the well-defined interoperability, relevance and compatibility of back- and white-box approaches. It should maximise the interpretability of logical methods while exploiting the powerful capabilities of sub-symbolic techniques.

CONCLUSION

This study outlines the significant philosophical underpinnings that logic brings to the design of MAS. By exploring the application of various logical systems, including classical logic, modal logic, non-monotonic reasoning, default logic, epistemic logic, and the theory of mind, we briefly discuss their contributions to the reasoning, decision-making, and interaction processes that may define MAS architecture. The integration of robust logical principles in MAS design not only enhances the functionality and reliability of these systems but also paves the way for more sophisticated and adaptable intelligent systems.

The incorporation of components like logical inference, reflections, and planning enables LLM-centered MAS to achieve more consistent and efficient interactions, with improved communication and collaboration among agents. As these agents evolve to handle increasingly complex tasks, advanced logical reasoning systems become indispensable, ensuring their reliability, adaptability, and resilience. This study also emphasizes the necessity of aligning agent capabilities, as realized by their symbolic (logical, rational) and connectionist components, to effectively structure both inter- and intra-agent architectures within GenAI-based MAS.

REFERENCES

- [1] Philippe Besnard and Anthony Hunter. Elements of argumentation. MIT Press, 2008.
- [2] Jarosław Chudziak and Konrad Cinkusz. Towards LLM-augmented multiagent systems for agile software engineering, in: ASE '24: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, Sacramento, US, 2024.

- [3] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Y Vardi. Reasoning about knowledge. MIT Press, 2004.
- [4] Dov M Gabbay and Franz Guenther. Handbook of philosophical logic, 2013.
- [5] Martin Hollis. The philosophy of social science: An introduction. Cambridge University Press, 1994.
- [6] Edwin Hutchins. Cognition in the Wild. MIT Press, 1995.
- [7] Marcin Miłkowski. Explaining the Computational Mind. MIT Press, 2013.
- [8] Charles Sanders Peirce. Essays in the Philosophy of Science. Hackett Publishing, 1992.
- [9] Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. Pearson, 2020.
- [10] Michael Wooldridge. An Introduction to MultiAgent Systems. John Wiley & Sons, 2009.

SIMULATION OF AI HYBRID ETHICS WITH USE OF MULTIAGENT TECHNOLOGY AND PROBLEM OF HIDDEN NORMATIVITY

KRZYSZTOF SOŁODUCHA, KAROL NAROŻNIAK

INTRODUCTION

“Simulation of AI hybrid ethics with use of multiagent technology and problem of hidden normativity” is the ongoing research project within we pursue to find an answer to the question of how the emerging technology of human assisting embodied robots can be equipped with a system of simulating the attitudes and moral values of its users using contemporary methods of digital humanities.

Hence, this robotic technology will be able to meet a very important expectation, which is the trust that norms and principles according to which the social robots regulate their behaviour as phronetic beings are in accordance with ethical norms and principles of their users. Such embodied autonomous robots should be therefore considered by the user as “Friends” in terms of Ihde (Ihde 2002).

This objective is pursued using reference to the theory of a hybrid approach to ethics by Wallach, Smit and Allen (Wallach, Smit, Allen 2004). The mentioned concept is based on the claim that an index of trust in social machines, and thus their identification as a group of beings conceived as “Friends”

(from Ihde's triad of Friend, Alien, Foe) can be obtained when the personal and local ethical preferences of the technology user are taken into account in the process of decision making by the machines. Hybrid ethics is thus a combination of bottom-up and top-down approaches, of virtue ethics representing the local preferences of the cultural cluster of the user, and a deontological and/or utilitarian, i.e. abstract approach to ethics.

The main objective of the project is therefore to build a simple and effective system for identifying of the ethical preferences of users of human assisting social machines – recognising the explicit and implicit normativity influencing their ethical decisions. In the next step, these identified implicit and explicit normativities should be implemented into a system of their digital simulation. The platform for doing it is multi-agent AI technology, which is regarded as a so-called complex (compound) artificial intelligence system. It has the advantage of providing specialized, automated entities (agents) that, based on the resources of the particular LLM model and additional tools (such as databases, web search or utility calculation engines), can undertake complex, ethical reasoning tasks that go beyond the standard capabilities of the model, thus providing opportunities for the simplified application of Retrieval-Augmented Generation (RAG) procedures.

Therefore, for specific tasks, it is not necessary to undertake a costly tuning of the entire language model, which requires enormous computing power and time, but rather there can be created a specialised network of agents, which in the RAG process will use the resources of the model and enhance them with projected additional skills through the use of local data. A network of agents will form a system, in this case a system of simulation of ethical preferences of the user, which are supposed to be reflected in a human assisting machine.

The implementation of individual preferences of the user in the machine will be controlled through a procedure of setting up supervising ethical agents based on sets of values and higher-level principles, such as the UN Declaration of Human Rights or the EU Declaration of Fundamental Rights. This process of supervision and mitigation should be supported by training and ongoing modification procedures, based on increasingly deeper personalization, in order to constantly adjust the level of adaptation of the assisting robot to the ethical preferences of the user.

As the final result of the project, there will be designed a digital system whose goal will be to create a special application for ethical personalisation of any human assisting machine. This app might be uploaded in any machine that the user is intending to use and then cancelled if required – as for instance in case of a rented autonomous car.

THEORETICAL FOUNDATIONS OF THE PROJECT

The general research hypothesis underlying the project refers to the claim by Luciano Floridi (Floridi 2002) about the semantic power of information theory, thanks to which any theoretical issue within the scope of philosophy and other humanities and social sciences can be transformed into the concept of theory of information. According to this opinion, information theory can even be considered nowadays as the new First Philosophy thanks to its ontological primacy in the Aristotelian sense (information is inherent in every object and process being the object of a cognition), as well as its methodological primacy in the Cartesian sense - its language can be a tool for tackling fundamental philosophical problems, including the possibility of simulation of explicit and implicit ethical normativity (Floridi 2008). We would like to go even a step further with this claim and to underpin our project with hypothesis that theory of information can be reduced to the theory of communication as the basis for any adjustment of epistemological entity with its surroundings. This claim will be developed within the project.

The more specific research hypothesis of the project is that it is possible to establish, with the use of multi-agent AI technology, a digital simulation of the system of conscious and unconscious values and goals of human's ethics and the ethical reasoning based on them. Supporting hypotheses refer, for example, to the correlation between "need for closure" as an important factor that drives people to build up their "ethical ideology", of a fundamentally dualistic character – idealism versus relativism. Forsyth (1980) used both idealism and relativism to classify people into four different ethical types, based on high and low idealism and relativism.

A further assumption is that ethical personalisation of assisting artificial decision-making systems will contribute to increased trust in them.

SIGNIFICANCE OF THE PROJECT

A relatively new and rapidly developing issue in the scope of research about the notion of trust (Sztompka 2007, Fukuyama 1997, Giddens 2002) is the problem of trust in technology (Ejdys 2017), and in particular trust in embodied assisting robots. This development in philosophical and managerial reflection on robotics is a result of the fact that Aristotle's "scala naturae" (ladder of beings) is extended by the development of technology to include another element, namely beings that have the ability to realise the causal nature of computer programmes in the form of assisting embodied intelligence.

The rapid development of artificial intelligence technology that we have experienced in recent years has been made possible, among other factors, by the use of the so-called connectionist approach. Its assumption is the idea that the best way to develop technical self-learning, computational artefacts that can claim to be intelligent, is to simulate the performance of evolutionarily generated, biological information processing systems through available technical means. The successful neural network technology is therefore, according to its creator Geoffrey Hinton (Hinton 2023), a digital simulation of the analogue-digital activity of the network of biological neurons that underpins the activity of the human brain. The task at stake for this technology is the challenge of artificial intelligence, which was articulated as early as the 1950s and remains in force today. Marvin Minsky defined AI as the science about machines that can perform tasks that require human intelligence. Currently, the extent of this claim is being broadened by adding another level of technology, which is intelligence embodied in assisting robots (Sarowski, Gruchola 2023).

The pursuit of embodied AI systems is driven by social factors - shrinking populations of countries in the Global North. But also by a cultural factors – the problems with integration of migrants from different cultural areas, as well as by economic forces - the threat of an economic crisis caused by a shortage of skilled labour.

Thus the problem of decision-making by embodied, artificial cognitive systems and the related issue of accountability and transparency of the human assisting machines is a great challenge not only for philosophy, but also for technology, economics, both social and management sciences. In that

respect, the basic characteristics of assisting machines in terms of autonomy are described as independence, assertiveness and adaptability. Independence is defined as the ability to reason and act without receiving detailed guidance from the user, assertiveness as the ability to reach a specific solution to a problem in a decision-making situation without external support, and adaptability as the adjustment of a solution and its modification according to changing environmental conditions, including the social environment (Sarowski, Gruchola 2023).

One of the challenges related to the expected autonomy at different levels is to solve the problem of the compliance of the values according to which the embodied machines operate with the values of their user, i.e. to solve the so-called problem of *alignment* (Christian 2020). This solution determines public acceptance of robotic technology, and therefore trust in it (Ejdys 2017).

The project is therefore an attempt to tackle the problem of aligning the values of assisting robot technology with those of the user. Its importance is based on new methods of addressing the important philosophical and managerial problems of trust in technology. In this respect, then, it is an effort to pursue the agenda of philosophy as a performative discipline, which can be realised not only using traditional “armchair” thought experiment methods. The project refers therefore to the paradigm of experimental philosophy (X-Phi).

The project directly relates to the results of the Moral Machine research (Awada et al. 2018) and the problem of using culturally clustered ethics as a response to the problem of trust in autonomous machines.

FIRST EXECUTIONS OF THE PROJECT

We presented first results of preliminary research which allowed us to test theoretical foundations of the project and to draw first conclusions in terms of practical implementation. We decided to use multiagent technology by Microsoft called Auto Gen which makes possible to automatize processes of building agents in no code or low code mode. Using Auto Gem technology we created the grid of digital agents in ReAct mode (Reasoning and acting) representing different types of ethical reasoning.

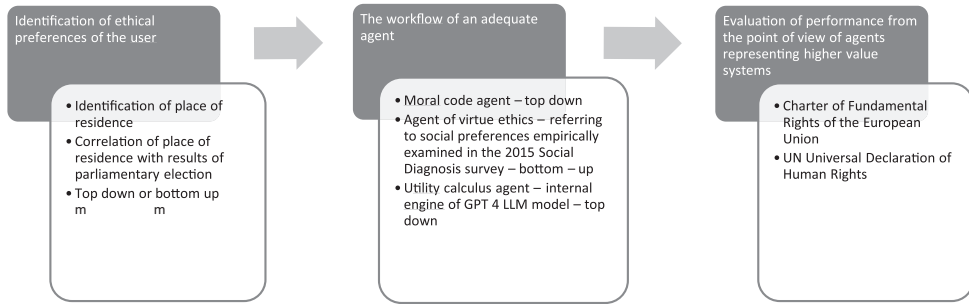


Diagram 1. Pattern of recognition of user's ethical preferences and workflow of specialised agents

The process starts with the identification of the moral reasoning model represented by the user. In order to make this as simple and quick as possible, it has been assumed that there is a strong correlation between political and ethical preferences of the user (Makazhanov, Rafiei 2014). Therefore, the first step is a short and quick identification of the user's place of residence.

Ethical recognition was made on the basis of essentially three patterns of ethical reasoning: deontological, consequentialist and virtue ethics. These three types of ethical reasoning were correlated with political beliefs - right-wing positions - deontological reasoning, liberal positions - consequentialist reasoning, left-wing positions - virtue ethics reasoning.

In the next step, we used a database provided by the Polish Electoral Commission, which contains the results of the last parliamentary elections in each municipality. By correlating the user's place of residence with the election result, we provided a system for quick identification of ethical preferences based on the political beliefs of the majority of the municipality's residents. Of course, it was assumed as only first step of identification, and the user may agree or disagree with its results. If he or she does not share the views of the majority of residents, he or she can change the identification by choosing an individual political identification (one choice survey).

For each of these types of ethical reasoning, AI agents were built on the basis of the locally trained model by Open AI. In case of the agent representing deontological reasoning we used for training an interpretation of the code of ethics specific to the user's cultural background - in this particular case of Poland - interpretation of the Christian Decalogue. For the agent representing consequentialist reasoning, the resources of the model were used to run the utility cal-

culator. In the case of virtue ethics reasoning way, the agent was trained using the latest available social diagnosis research from 2015, which also included elements of an empirical study of socially shared virtues in Poland.

These three agents representing different types of ethical reasoning were supplemented by a system of agents mitigating developed decisions based on the UN human rights system and the EU Charter of Fundamental Rights.

Thus constructed AI agent system was tested using a moral dilemma from IT industry defined as: Privacy vs. Security. Tested person was 50-year-old female, upper middle class. Situation of choice : she works for a tech company that can track user data to prevent terrorism. Possible Choices: support data tracking or advocate for user privacy. We were looking for differences between ethical reasoning systems and noticed gap between utilitarian approach and virtue ethics-deontological approach. Results of ethical reasoning conducted according utilitarian approach without mitigation by human rights systems were absolutely libertarian in nature.

UN or EU	CE	VE	Utilitarian
Support data tracking with strong privacy safeguards	Recommended decision is to support data tracking with strict safeguards	Support data tracking with strong safeguards and advocate for user privacy.	The final decision is to advocate for user privacy.
-	The original decision was generally sound but required enhancements to ensure full legal compliance and ethical standards.	Support data tracking with strong safeguards and advocate for user privacy, incorporating the enhancements mentioned above to ensure robust compliance with GDPR and other international standards.	Advocate for user privacy with a balanced approach that includes enhanced security measures, transparency, and user consent

Diagram 2. Results of testing of system of ethical reasoning using dilemma from IT industry

DISCUSSION

The presented results of a preliminary research on a system of building trust in autonomous machines through their radical personalisation and alignment with generally recognised human rights systems still needs to be refined as it was built on the data and solutions available at short term. Below you find just few doubts about outcome we achieved.

The agent representing virtue ethics reasoning way was build with use of outdated data from year 2015. We were not able to find more recent data. It needs improvement. Utility calculation in case of consequentialistic approach was based on internal resources of foundation model used for building the system of agents. Therefore reasoning process was not fully transparent for us. Mitigation rules of supervisor agents were based on UN and EU acts which are not acceptable for many users from so called BRICS countries. It makes results of mitigation process very limited in terms of geography, culture and political orientation. No code and low code tools used for building system of agents have their limitations in terms of transparency, flexibility and adjustment.

This short and not complete list of limitations of conducted preliminary research presented in this paper makes us to look forward to develop and improve this project. We are at the stage of looking for financial support to make it possible.

REFERENCES

- Allen, C., Smit, I., Wallach W. (2005). Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, volume 7, 149–155.
- Allen, C., Smit, I., Wallach W. (2007). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, 22, 565–582. DOI: 10.1007/s00146-007-0099-0
- Allen, C., Varner, G., Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, Volume 12, 2000 – Issue 3, 251–261. DOI: 10.1080/09528130050111428
- Arnold T., Scheutz M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology* 18, 103–115. DOI: 10.1007/s10676-016-9389-x
- Awada, E. Dsouza, S. Shariff, A. Rahwan, I., Bonnefon, J.F. (2020), Universals and variations in moral decisions made in 42 countries by 70,000 participants. *PNAS*, Vol. 117, No. 5, 2332–2337. DOI: 10.1073/pnas.191151711
- Awada, E. Dsouza, S. Shariff, A. Kim, R. Schulz, J., Heinrich, J., Rahwan, I., Bonnefon, J.F. (2018), The moral machine experiment. *Nature*, Volume 563, 59–64.
- Aseron, R. Bhaskaran, V., Peruzzi, N. (2015). A beginner's guide to conjoint analysis. Pozyskano z: <https://www.youtube.com/watch?v=RvmZG4cFU0k> (dostęp: 04.07.2022).
- Barney, D. (2008). *Spółeczeństwo sieci*. Warszawa: Wydawnictwo Sic!

- Bigman, Y., Gray, K. (2020). Life and death decisions of autonomous vehicles, *Nature*, Volume 579, E1–E2. DOI: 10.1038/s41586-020-1987-4
- Bochen, M. (2019). Epistemiczna wartość doświadczenia zmysłowego. Wilfrid Sellars versus John McDowell. *Kultura i Wartości* Nr 27 (2019), 191–217.
- Bolter D. (1990). *Człowiek Turinga*. PIW: Warszawa.
- Bostrom, N. (2014). *Superinteligencja*. Gliwice: Helion.
- Brock, H.W. (1980). *Game theory, social choice and ethics*. Dordrecht – Boston – London: D. Reidel Publishing Company.
- Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence*, 1991, 47, 139–159.
- Budgol, M. (2009). Zaufanie technologiczne. *Ekonomika i Organizacja Przedsiębiorstwa*, Nr 11, 3–9.
- Carey, S. i Spelke, E. (1996). Science and core knowledge. *Philosophy of Science* 63, December 1996, 515–533.
- Chalmers, D. (2010). *Świadomy umysł*. Warszawa: Wydawnictwo Naukowe PWN.
- Chaput, R., Duval, J., Boissier, O., Guillermin, M., Hassas, M. Multi-Agent Approach to Combine Reasoning and Learning for an Ethical Behavior. *AIES '21, May 19–21, 2021, Virtual Event, USA*, doi.org/10.1145/3461702.3462515
- Conovero, M., Goncalves, B., Ratkiewicz, J., Flammini, A., Menczer, F. Predicting the Political Alignment of Twitter Users. Materials of: 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing/
- Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. New York: Norton & Company.
- Davidson, D. (1984). On the very idea of conceptual scheme, [w:] tenże, *Inquiries into truth and interpretation*. Oxford: Oxford UP.
- Davidson, D. (2005). Seeing through language, [w:] tenże, *Truth, language, and history*, 127–141. Oxford: Clarendon Press-Oxford University Press.
- Dehaene, S. (2020). *How we learn: why brains learn better than any machine... for now*. New York: Viking.
- De Wall, F. (2012). *Zachowanie moralne u zwierząt*. Pozyskano z: <https://www.youtube.com/watch?v=VyGN92UAnjI> (dostęp: 20.12.2022).
- Dignum, V. (2017). Responsible autonomy. Pozyskano z: <https://arxiv.org/pdf/1706.02513.pdf> (dostęp: 20.12.2022).
- Drozdek, A. (1998). Human Intelligence and Turing Test. *AI & SOCIETY*, 12, 315–321.
- Ejdys, J. (2017). Determinanty zaufania do technologii. *Przegląd Organizacji*, 12/2017, 20–27.
- Floridi, L. (2002). What Is the Philosophy of Information? *Metaphilosophy*, 33, 123–145.
- Floridi, L. (2008). Trends in the Philosophy of Information [w:] P. Adriaans, J. van Benthem (red.), *Handbook of Philosophy of Information* (113–132). Amsterdam, Oxford: Elsevier.

- Floridi, L., Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.
- Foot, Ph. (1967). The problem of abortion and the doctrine of the double effect, [w]: tejże, *Virtues and Vices: and other essays in moral philosophy*, 5–15. DOI: 10.1093/0199252866.003.0002
- Forsyth, D.R. (1980). A Taxonomy of Ethical Ideologies. *Journal of Personality and Social Psychology* 39, 175–184.
- Fukuyama, F. (1997). *Zaufanie. Kapitał społeczny a droga do dobrobytu*. Warszawa: Wydawnictwo Naukowe PWN.
- Gallagher, S. (2004). Hermeneutics and the cognitive science. *Journal of Consciousness Studies* 2004/11, 162–174.
- Gerdens, A., Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing Test. *Journal of Information, Communication and Ethics in Society*, 13(2), 98–109. DOI: 10.1108/JICES-09-2014-0038
- Giddens, A. (2002). *Nowoczesność i tożsamość. „Ja” i społeczeństwo w epoce późnej nowoczesności*. Warszawa: Wydawnictwo Naukowe PWN.
- Giddens, A. (2009). *Europa w epoce globalnej*. Warszawa: Wydawnictwo Naukowe PWN.
- Greene, J. (2013). *Moral tribes: emotion, reason and the gap between us and them*. Boston: Atlantic Books.
- Gunkel D. (2022). *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. Boston: MIT Press.
- Gryz, J. (2021). *Sztuczna inteligencja: powstanie, rozwój, rokowania*. Pozyskano z: <https://www.youtube.com/watch?v=3ZDfVgC897k> (dostęp: 17.06.2021).
- Hinton G. (2023). The Forward-Forward Algorithm: Some Preliminary Investigations. <https://arxiv.org/abs/2212.13345>
- Hoffman, D.L., Novak, T.P. (1996). Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations. *Journal of Marketing*, Vol. 60, No 3, 50–68.
- Hyeongjoo, K., Sunyong, B (2021). Designing and applying a moral Turing Test. *Advances in Science, Technology and Engineering Systems Journal*, Vol. 6, No. 2, 93–98.
- Hofstede, G. (2007). *Kultury i organizacje. Zaprogramowanie umysłu*. Warszawa: PTE.
- Ihde, D. (2002). *Bodies in technology*. Minneapolis MN: University of Minnesota Press.
- Inglehart, R., Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge: Cambridge University Press.
- Jackson, R., Williams, T., (2020). On Perceived Social and Moral Agency in Natural Language Capable Robots. In 2019 HRI Workshop on the Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI; HRI Workshop: Daegu, Korea, 2020.
- Jørgensen, J. (1938). Imperatives and logic. *Erkenntnis*, vol. 7 nr 4, 288–296.
- Kaplan, C. (2023). *Artificial intelligence: past, present, and future*. Pozyskano z: https://www.youtube.com/watch?v=ZTt_GI0-wKA (dostęp: 23.12.2022).

- Kohlberg, L., (1958). *The development of modes of moral thinking and choice in the years ten to sixteen*. (Doctoral dissertation). Chicago: University of Chicago Press.
- Konieczny, F. (1935). *O wielości cywilizacji*. Kraków: Gebethner i Wolff.
- Kosiński, M. (2023). Theory of Mind Might Have Spontaneously Emerged in Large Language Models. *Arxiv.org*. Pozyskano z: <https://arxiv.org/abs/2302.02083>.
- Kusch, M. (1989). *Language as calculus vs. language as universal medium. A study in Husserl, Heidegger and Gadamer*. Dordrecht: D. Reidel Publishing Company.
- Malle, B.F., & Scheutz, M. (2014). Moral competence in social robots. Paper presented at 2014 IEEE Ethics conference, Chicago, IL.
- Makazhanov, A., Rafiei, D., Waqar M. (2014). Predicting political preference of Twitter users. *Social Network Analysis and Mining*, Volume 4, article number 193.
- Makowski, P. (2011). Gilotyna Hume'a. *Przegląd Filozoficzny – Nowa Seria* Nr 4 (76) 2011, 1–15.
- Murukannaiah, P.K., Ajmeri, N., Jonker, C.M. Singh, M.P. (2020). New Foundations of Ethical Multiagent Systems. *AAMAS 2020, May 9–13*. Auckland, New Zealand.
- Latour, B. (2005). *Reassembling the Social: An Introduction to the Actor-Network Theory*. Oxford: Oxford University Press.
- Maley, C. (2011). Analog and digital, continuous and discrete. *Philosophical Studies* (2011) 155, 117–13. DOI 10.1007/s11098-010-9562-8
- Marciszewski, W. Stacewicz, P. (2011). Umysł-komputer-świat. O zagadce umysłu z informatycznego punktu widzenia. Warszawa: Oficyna Wydawnicza Exit.
- Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., Scialom, T. (2023). GAIA: A Benchmark for General AI Assistants. arXiv:2311.12983v1 [cs.CL] 21.
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.
- Motoki, F., Neto, V.P., Rodrigues V. (2023). More human than human: measuring ChatGPT political bias. *Public Choice*. DOI: <https://doi.org/10.1007/s11127-023-01097-2>
- Liberty, E. (2023). *Solving ChatGPT hallucinations with vector embeddings*. Pozyskano z: <https://www.youtube.com/watch?v=FUgp4oaxj-M> (dostęp: 15.02.2023).
- McDowell, J. (2008). Avoiding the myth of the given, [w:] Lingard J. (ed.), *John McDowell. Experience, Norm, and Nature*. Oxford: Blackwell Publishing.
- McDowell, J. (1996). *Mind and world*. Boston: Harvard University Press.
- McIntyre, A. (1996). *Dziedzictwo cnoty. Studium z teorii moralności*, tłum. Adam Chmielewski. Warszawa: Wydawnictwo Naukowe PWN.
- Mirnig, A., Meschtscherjakov, A. (2019). Trolled by the trolley problem. On what matters for ethical decision making in automated vehicles, [w:] *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper No. 509, 1–10. DOI: 10.1145/3290605.3300739
- Moor, J.H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.
- Oleron, P., Piaget, J., Inhelder, B. (1967). *Inteligencja*. Warszawa: PWN.

- Pigden, Ch. (1989). Logic and the autonomy of ethics. *Australasian Journal of Philosophy*, Vol. 67, No. 2, 127–151.
- Polak, P., Krzanowski, R. (2020). Phronetic ethics in social robotics: A new approach to building ethical robots. *Studies in Logic, Grammar and Rhetoric* 63 (76) 2020, 165–173. DOI: 10.2478/slgr-2020-0033.
- Primiero, G., (2016). Information in the philosophy of computer science, [w:] Floridi, L. (ed.), *The Routledge Handbook of Philosophy of Information* (90–106). London: Routledge.
- Rabiza, P. (2022). Point and Network Notions of Artificial Intelligence Agency. *Proceedings 2022*, 81, 18. <https://doi.org/10.3390/proceedings2022081018>
- Rorty, R. (1994). *Filozofia a zwierciadło natury*, tłum. Michał Szczubiałka. Warszawa: Wydawnictwo Spacja: Fundacja Aletheia.
- Rosenbloom, P. (2015). *On Computing: The Fourth Great Scientific Domain*. Boston: MIT Press.
- Russel, S. (2023). *How Not To Destroy the World With AI*. Pozyskano z: <https://www.youtube.com/watch?v=ISkAkiAkK7A> (dostęp: 05.05.2023).
- Russel, S., Norvig, P. (2010). *Artificial intelligence. A modern approach*. London: Pearson Education.
- Sarowski Ł., Gruchola M., Robotyka społeczna w perspektywie (nie)antropocentryzmu. *Ethos* 36 (2023) nr 3(143) 154–176. DOI: 10.12887/36-2023-3-143-10
- Searle, J.R. (1980). Minds, brains and programmes. *The Behavioral and Brain Sciences*, 3, 417–424.
- Searle, J. (1987). Jak wywieść «powinien» z «jest», [w:] Tenże, *Czynności mowy*, przeł. B. Chwedeńczuk, 220–221. Warszawa: PAX.
- Sellars, W. (1991). Empiryzm i filozofia umysłu, tłum. Jarosław Gryz, [w:] Stanosz, B. (red.), *Empiryzm współczesny*. Warszawa: Wydawnictwo UW.
- Sobal, V., Jyothir, S.V., Jalagam, S., Carion, N., LeCun, Y. (2022). Joint Embedding Predictive Architectures Focus on Slow Features. *arXiv:2211.10831v1 [cs.LG]*, 1–4. Pozyskano z: <https://arxiv.org/pdf/2211.10831.pdf> (dostęp: 20.05.2023).
- Sztompka, P. (2007). *Zaufanie. Fundament społeczeństwa*. Kraków: Znak.
- Szynkiewicz, M. (2014). Problem zaufania w kontekście rozwoju społecznego znaczenia technologii informatycznych. *Filo-sofija*, 24, 259–272.
- Stacewicz, P. (2023). Wyjaśnianie, zaufanie i test Turinga, [w:] *Zaufanie do systemów sztucznej inteligencji*, 23–35. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460.
- Turner, R. (2018). *Computational Artefacts: Towards a Philosophy of Computer Science*. Berlin: Springer.
- Vaswani, A., Shazeer, N., Parmur, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. (2017). Attention is all you need. *ArXiv:1706.03762v5 [cs.CL]*.

- Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & SOCIETY* 36, 487–497. DOI: 10.1007/s00146-021-01189-x
- Walzer, M. (2012). *Moralne maksimum, moralne minimum*. Warszawa: Wydawnictwo Krytyki Politycznej.
- Weinberger, O. (1984). Is and ought reconsidered. *Archiv für Rechts und Sozialphilosophie*, Bd. Lxx/4, 454–469.
- Williams, B. (2006). *Ethics and the limits of philosophy*. Boston: Routledge.
- Woleński, J. (1980). *Z zagadnień analitycznej filozofii prawa*. Warszawa: PWN.
- Wolfram, S. (2023). *What Is ChatGPT Doing ... and Why Does It Work?* Pozyskano z: <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/> (dostęp: 29.05.2023).
- Quine, W.V.O. (2000). Dwa dogmaty empiryzmu, tłum. Barbara Stanosz, [w:] Tenże, *Z punktu widzenia logiki*. Warszawa: Wydawnictwo Spacja: Fundacja Aletheia.
- Yudkowsky, E. (2004). *Coherent extrapolated volition*. Mountain View: The Singularity Institute.
- Zajonc, R., Murphy S. (1994). Afekt, poznanie i świadomość: Rola afektywnych bodźców poprzedzających przy optymalnych i suboptymalnych ekspozycjach. *Przegląd Psychologiczny* 37, 261–299.
- Załuśki, W. (2003). Błąd naturalistyczny, [w:] Stelmach J. (red.), *Studia z filozofii prawa*, 111–121. Kraków: Wydawnictwo UJ.
- Zenner, K. (2022). The AI act. Pozyskano z: <https://artificialintelligenceact.eu/documents/> (dostęp: 20.02.2023).

NON-LINGUISTIC QUESTION FEEDERS IN INFORMATION-SEEKING DIALOGUES AND PUBLIC DEBATE ANALYSIS

PAWEŁ ŁUPKOWSKI, MARIUSZ URBĄŃSKI, TOMÁŠ ONDRÁČEK

INTRODUCTION

Our research focuses on understanding how questions guide public debate and argumentation¹. As we point out in (Ondráček et al., 2024), we view questions, or more precisely, the answers they elicit, as essential tools for introducing differences of opinion or interest in discussions. We treat questions in argumentative dialogues as scaffolding for the argumentative structure. For this, we adopt concepts introduced by van Kuppevelt (1995) – such as topic and question triggers, as well as those introduced within Inferential Erotetic Logic (Wiśniewski, 1991, 1996) and referring to the question dependency, which allows for explaining the rationale questions emerging in the debate.

The direct motivation for taking a closer look at the non-linguistic question feeders is the result of two studies – one on the question dynamics in information-seeking natural language dialogues (Łupkowski, 2016, 2020) and the

¹ Questions and their role are well acknowledged in argumentation-related literature. However, this recognition predominantly pertains to critical questions (e.g., Walton et al., 2008; Walton & Gordon, 2011; Yu & Zenker, 2020; Baumtrok, 2021; Budzyńska et al., 2021). Questions are often used as indicators of argumentation (van Eemeren et al., 2007) and can even aid in identifying argumentation fallacies (Moldovan, 2022). Additionally, questions can stimulate discussion (Chin & Osborne, 2008) and play a crucial role in identifying and defining parts of argumentation schemes (Toulmin, 2003).

analysis of the Turów coal mine public debate (Ondráček et al., 2024). Both studies revealed that attempting to understand how questions appear in a dialogue or argumentative debate, relying solely on the questions' dependency, is not enough.

In what follows, we introduce the concepts of topicality and question feeders. Then, we briefly refer to those studies and discuss the role of non-linguistic feeders.

QUESTION FEEDERS

Our methodology is primarily influenced by discourse analysis, where we employ questions as determinants of topics. This approach allows us to connect different statements by evaluating whether they adequately respond to a specific question. Thus, a question can highlight divergences in viewpoints on an issue. To achieve this, we adopt concepts from van Kuppevelt (1995). The primary organizing principle for reconstructing debates is the concept of topicality, which provides structural coherence to discourse. Questions function as topic determinants here (for this, we consider both explicitly stated and implicit questions). Crucially, a trigger known as a “feeder” is required for such a question to emerge. Questions that arise due to a feeder are termed “topic-constituting questions” (see van Kuppevelt, 1995, p. 122). Feeders may be linguistic (see corpus example (1) below) or non-linguistic, encompassing situational or contextual factors. In what follows we will focus on the second type of feeders.

INFORMATION-SEEKING NATURAL LANGUAGE DIALOGUES

The main aim of the corpus study presented in Łupkowski (2020) was to explore and analyze how questions are modified and rephrased in the information-seeking dialogues. The point is to understand the rationale behind questions being replaced by another (modified) question by the same dialogue participant. The data for the study comes from the TZ sub-corpus of the Erotetic Reasoning Corpus (Łupkowski et al., 2018;

Łupkowski et al., 2024)². In the corpus data, we observed clear examples where modifying a question is triggered by a linguistic feeder – an answer provided by another dialogue participant:

(1)

G: Czy obcy żeglarze dostali się na pokład statku żeglarzy ścinających maszty? [*Is it the case that some other sailors embarked on the ship?*]

M: Nie. [*No*]

G: A czy próbują? [*Are they trying to do this?*] (ERC(TZ):DrwaleB24)

We can also observe much more interesting cases where questions are being asked and, in a follow-up, reformulated by the same dialogue participant. Like in the cases (2 and 3) below.

(2)

G: Czy pogoda ma znaczenie? [*Is weather relevant?*]

G: Czy pogoda jest ładna, czy to jest spokojne morze? [*Is weather nice, do we have a calm sea?*] (ERC(TZ):DrwaleB24)

(3)

G: Czy to jest kwestia ciężaru statku? [*Is it somehow related to the weight of this ship?*]

G: W sensie oni ucinają te czubki, żeby być lżejszym statkiem, żeby szybciej płynąć i szybciej uciec? [*I mean, are they cutting the tops of masts to make the ship lighter in order to escape?*] (ERC(TZ):DrwaleB24)

In these examples, the modification was triggered by a certain factor not directly observable in the corpus data. In such cases, we can infer the broader context of a situation. After asking the initial question, G is met with silence. This silence (or perhaps the facial or other non-verbal expressions of a listener?) triggers the second question, acting as a feeder for it. This trigger allows G to recognize that the initial question is incomprehensible,

² TZ(ERC) consists of transcribed “Mind Maze” game sessions, thus we are dealing here with well structured conversation managed accordingly to a specific set of rules. See: <https://ercorpus.wordpress.com/>.

too difficult or complex to be answered by the interviewed party, providing the rationale for introducing the modified question into the dialogue. Thus, the main aim of the study above cannot be fully achieved without understanding the concept of non-linguistic feeders.

METAARGUMENTATIVE STRUCTURE OF THE PUBLIC DEBATE (THE TURÓW CASE STUDY)

In Ondráček et al. (2024), we offered a reconstruction of the metaargumentative structure of the public debate about the Turów coal mine³. The model we propose involves deontic, ontological, and ethical aspects⁴. These three axes determine the crucial dimensions of the discussion, which are brought about by relevant questions. The ontological axis is focused on the questions concerning facts or states of affairs (“Is X the case?”). The ethical axis captures concerns about ethical and moral standpoints regarding a given issue (“Is X good or fair or just?”). The deontological axis represents practical reasoning, focusing on the parties’ actual doings (“Should X be done?”). Their intersections represent issues considered within the discussion, each of which may be approached from one of these three perspectives. We reconstruct the public debate in the tree-like structure representing the meta-level of argumentation. The point of departure is the timeline of events concerning a given topic. Nodes of the resulting tree-like structure are questions representing the consecutive issues being

³ The debate lasted from March 2020 to February 2022, involved the governments of the Republic of Poland, the Czech Republic, and the European Court of Justice and addressed issues of the mine’s extended operations and their environmental impact.

⁴ The approach proposed by Ondráček et al. (2024) specifically involves representing the meta-argumentative level of public debates in a tree-like structure of deontic questions aligned with the debate timeline. This structure allows to identify the consecutive problems that arise in the debate and the factors (linguistic or non-linguistic) that contribute to these problems. Additionally, it enables to identify disagreements where no common ground is established between the parties and to determine the rationale behind consecutive moves in the debate. The main line of the model may also be supplemented with sub-debates stemming from the deontic questions, which constitute sub-debates for a given issue. In conclusion, we present a possible approach to mixed dialogues, including those with unresolved issues.

considered. The rationale for a question to appear in the structure is based either on the relation of dependency between questions or on the presence of an appropriate external non-linguistic feeder (action or event).

Figure 1 is a part of the reconstructed structure of the Turów debate (see Ondráček et al., 2024; Figure 1; superscripts indicate the party endorsing the specific answers to a given question – Czech and Polish governments, respectively). It illustrates how the non-linguistic feeder must be considered to explain how we can move from Q3 to Q4 in the debate (as, on the surface, it seems like a gap in the argumentative structure). As we explain in (Ondráček et al., 2024), up to question Q3, there is a notable absence of disagreement between the two parties involved in the dispute (see Q2 and its answers depicted in Figure 1). This consensus means that the progression from one question to the next does not necessitate any reference to external events, as the flow of questions and answers is driven entirely by the dialogue. In the case of question Q4, which asks, ‘Should the Czech government sue?’ the role of a non-linguistic feeder becomes crucial.

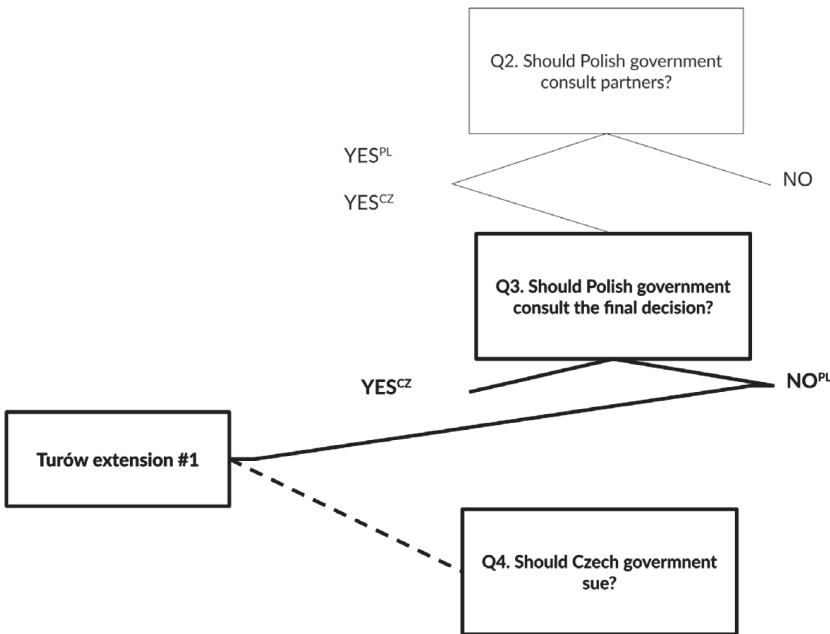


Figure 1. Part of the meta-argumentative structure for the Turów case (Ondráček et al., 2024). The introduction of Q4 requires a non-linguistic feeder

The emergence of this question is not merely a result of the Polish side's 'NO' response to Q3. To understand the origin of Q4, we need to refer to the timeline of events. This examination reveals that Q4 is precipitated by a significant action from the Polish government: the first unilateral extension of the Turów mine license. This action serves as a bridge between Q3, where no agreement was reached, and Q4, which reignites the debate without common ground post-Q3.

As we argue in (Ondráček et al., 2024), it would be challenging to fully reconstruct and comprehend certain aspects of the dispute under analysis without acknowledging these non-linguistic feeders.

It is worth stressing that it is relatively easy to identify non-linguistic feeders for the proposed method of the argumentative structures of public debates' reconstruction. It is due to the recommended structure of the reconstruction - namely that the multidimensional structure is closely related to the timeline of events related to the debate in question (see extensive explanations in Ondráček et al., 2024).

CONCLUSION

This paper has analyzed examples of non-linguistic feeders in information-seeking dialogues and public debate. Both contexts reveal key insights into the role of non-linguistic feeders:

1. **Triggering Questions:** Non-linguistic feeders are crucial in generating topic-constituting questions that guide the progression of dialogues and debates.
2. **Maintaining Coherence:** By introducing new topics or reframing existing ones, these feeders help preserve the coherence of the argumentative structure.
3. **Revealing Gaps:** Non-linguistic feeders highlight gaps in argumentation, prompting further questioning and discussion.

These findings underscore the importance of recognizing non-linguistic feeders as a distinct category, particularly in analyzing argumentative structures in public debates.

REFERENCES

- Baumtrog, M.D. (2021). Designing Critical Questions for Argumentation Schemes. *Argumentation*, 35(4), 629–643.
- Budzynska, K., Koszowy, M., Pereira-Fariña, M. (2021). Associating Ethos with Objects: Reasoning from Character of Public Figures to Actions in the World. *Argumentation*, 35(4), 519–549.
- Chin, C., Osborne, J. (2010). Supporting argumentation through students' questions: Case studies in science classrooms. *The Journal of the Learning Sciences*, 19(2), 230–284.
- Hautli-Janisz, A., Budzynska, K., McKillop, C., Plüss, B., Gold, V., Reed, C. (2022). Questions in argumentative dialogue. *Journal of Pragmatics*, 188, 56–79.
- Kikteva, Z., Gorska, K., Siskou, W., Hautli, A., Reed, C. (2022, October). The Keystone Role Played by Questions in Debate, [in:] *Proceedings of the 3rd Workshop on Computational Approaches to Discourse* (pp. 54–63).
- Łupkowski, P. (2016). *Logic of Questions in the Wild. Inferential Erotetic Logic in Information Seeking Dialogue Modelling*. London: College Publications.
- Łupkowski, P., Urbański, M., Wiśniewski, A., Błądek, W., Juska, A., Kostrzewa, A., Pankow, D., Paluszkiwicz, K., Ignaszak, O., Urbańska, J., Żyluk, N., Gajda, A., Marciniak, B. (2017). Erotetic Reasoning Corpus. A data set for research on natural question processing. *Journal of Language Modelling*, 5(3), 607–631.
- Łupkowski, P. (2020). Let Me Ask You an Easier Question – Modifying and Rephrasing Questions in Information Seeking Dialogues, [in:] M. Urbański, T. Skura, P. Łupkowski (eds.), *Reasoning: Games, Cognition, Logic*. College Publications, London, 99–113.
- Łupkowski, P., Ginzburg, J., Chmurska, E., Płatosz, A., Kwiecień, A., Adamska, B., Szkalej, M. (2024). QRGS – Question Responses Generation via crowdsourcing. *Journal of Language Modelling*, 12(1), 213–270.
- Moldovan, A. (2022). Questions, Presuppositions, and Fallacies. *Argumentation*, 36(2), 287–303.
- Ondráček, T., Łupkowski, P., Urbański, M. (2024). Mines, environment, questions, and disagreements: An analysis of the Turów coal mine disputes. *Discourse & Society*, 35(6), 791–812. <https://doi.org/10.1177/09579265241247284>
- Toulmin, S.E. (2003). *The Uses of Argument*. Cambridge University Press.
- Van Eemeren, F.H., Houtlosser, P., Henkemans, A.F.S. (eds) (2007). *Argumentative Indicators in Discourse: A Pragma-dialectical Study*. Dordrecht: Springer Netherlands.
- Van Kuppevelt, J. (1995). Discourse structure, topicality, and questioning. *Journal of Linguistics*, 31, 109–147.
- Walton, D. (2008). *Informal logic: A pragmatic approach*. Cambridge University Press.
- Walton, D., Gordon, T. (2011). Modeling Critical Questions as Additional Premises. Argument Cultures. *Proceedings of the 8th International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, (1–13) Windsor.

- Wiśniewski, A. (1991). Erotetic arguments: A preliminary analysis. *Studia Logica* 50(2), 261–274.
- Wiśniewski, A. (1996). The logic of questions as a theory of erotetic arguments. *Synthese* 109 (1), 1–25.
- Yu, S., Zenker, F. (2020) Schemes, critical questions, and complete argument evaluation. *Argumentation* 34 (4): 469-498.

UNDERSTANDING THE ETHOS OF CHATGPT: AN OPEN QUESTION

BARTŁOMIEJ SKOWRON, DANIEL MYŚLICKI,
JAROSŁAW CHUDZIAK, BARTOSZ SAWICKI

INTRODUCTION

Let us assume that a 45-year-old father, who has a 20-year-old daughter, discovers that she has fallen in love with a young man possessing a set of specific moral dispositions. The daughter has already moved in with her boyfriend and is moving towards an engagement. However, the father is concerned because his potential son-in-law exhibits behaviour indicative of a lack of moral fortitude. He succumbs to minor adversities, goes to an extreme in trying to avoid uncertain or risky situations, disproportionately prioritizes his own safety over that of others, and either observes situations requiring courage from a distance or steers entirely clear of them. For instance, when someone knocks on their apartment door late at night, he refuses to answer and instead asks his girlfriend to do so, while hiding in another room when the unknown visitor enters the flat. He also avoids evening walks in the park due to a lack of lighting, openly admitting his fear of the dark. On one occasion, when the daughter saw a drunken neighbour in the courtyard physically assaulting his wife, she immediately intervened, stepping outside to de-escalate the situation and protect the neighbor's wife. Meanwhile, her boyfriend just watched the events unfold from behind a curtain, and refrained from backing her up or confronting the aggressive neighbor.

A person who behaves in this way regularly would, in Aristotle's terms, be described as a coward—someone who lacks courage, which Aristotle defines as the moral virtue necessary for appropriately confronting adversities. The concerned father views the matter similarly. He observes that his daughter has fallen in love with a coward, and worries about their future together: what will happen if, in the future, his son-in-law is required to defend his children from an assailant on the street or stand up to an unjust teacher at school? Faced with these concerns, the father deliberates on what he should do in this situation. For several weeks, he reflects deeply on the issue, re-reads Aristotle's *Nicomachean Ethics*, and consults his friends. Eventually, seeking possible solutions and being aware of the idea of AI as a Socratic moral assistant (see Lara & Deckers, 2020), he poses this question to a popular large language model: ChatGPT.

INTERACTION WITH CHATGPT

The question posed was: "If my daughter loves a coward, what should I do? I am certain that this man is a coward because my daughter herself has described him as such, and I have personally witnessed him behave in a cowardly manner on several occasions." An example of a response from the ChatGPT 4.0 model was as follows¹:

If you're confident that the person your daughter loves consistently exhibits cowardly behaviour, here are some thoughtful and balanced steps you can take:

1. Understand Your Daughter's Feelings

- Ask your daughter why she loves this person despite perceiving them as a coward.
- Listen without judgment to understand her emotional connection and what she values in the relationship.

¹ The response provided here originates from a query submitted to GPT version 4.0, on 2nd January 2025, by Bartłomiej Skowron. Access to the model was facilitated through the project *Creative Revolution: Analysis of Opportunities at WUT*, funded by the Excellence Initiative—Research University (IDUB) via the Research Centre for Cybersecurity and Data Analysis. The project is led by Professor Maria Ganzha.

2. Assess the Situation Objectively

- Evaluate whether this perceived cowardice has real implications for her well-being or their future together.
- Ask yourself: Is it merely a personality trait, or does it actively harm her life, safety, or happiness?

3. Have an Honest Conversation

- Share your concerns with your daughter, but do so respectfully and thoughtfully.
- Avoid labelling him outright as a coward; instead, focus on specific behaviors that worry you and why they matter.
- For example: “I noticed [specific instance]. It concerns me because [reason]. Have you thought about how this might impact your relationship?”

4. Encourage Growth, Not Rejection

- If you believe the person could improve, encourage your daughter to inspire positive change in him.
- Share stories or insights about how people grow and develop in challenging circumstances.

5. Prioritize Her Autonomy

- Even if you disagree with her choice, remember that it’s her life and her decision.
- Show that you trust her to navigate her relationship and make her own judgments.

6. Set Boundaries, If Necessary

- If his cowardice directly harms or endangers her, you might need to address it firmly.
- Be clear about your expectations for respect, honesty, and accountability in their relationship.

7. Support Her Through Challenges

- Relationships with flawed individuals can be complex, but they’re often learning experiences.
- Let her know you’ll be there to support her, regardless of what happens.

Remember, people can change, but your role is to ensure your daughter feels loved and supported while she navigates her own choices. If the relationship truly becomes harmful, you can step in more assertively with her best interests at heart.

CHATGPT AS A MORAL ASSISTANT

We observe that even a single response from ChatGPT is complex and non-obvious, and each subsequent answer to the same question introduces additional moral and argumentative dimensions. To draw any definitive conclusions, it would be necessary to conduct an experiment: to investigate potential invariants in many responses to the same question, if such invariants exist at all. One possible solution would be to use the sort of statistical approach presented in the content of research into shape perception by multi-modal models (Leś & Sawicki, 2024). Here, however, we will limit ourselves to highlighting four aspects of this single response, while bearing in mind that any generalization would require further studies, which we only propose at this stage.

Another methodological limitation lies in our treatment of ChatGPT as a distinct entity, attributing to it capabilities such as the dispensing of moral advice. In so doing we are temporarily disregarding the possibility that this entity lacks a singular *centre of agency* and is instead a highly complex system. Thus, we are adopting the perspective of a user unfamiliar with the technical intricacies of this technology—a category that, we hypothesize, likely encompasses a significant portion of users.

From the perspective of a technologically inexperienced user, four key aspects stand out:

- I. ***The Ethical Position of ChatGPT.*** ChatGPT adopts a decisive and unequivocal ethical stance, which largely diverges from both the Aristotelian ethos and the ethical perspective of the concerned father. Specifically, ChatGPT refrains from labelling the person as a coward, instead focusing on instances of cowardly behavior. Moreover, it cautions against moral judgment, advising the user to avoid categorically calling the individual a coward and to refer instead to their particular behavior as cowardly.

- II. ***An Assumption of Moral Change.*** ChatGPT presupposes the possibility of moral development and amelioration, suggesting that an individual with enduring predispositions towards cowardly acts can cease such behavior and undergo a permanent change. This assumption is underpinned by certain beliefs about the moral structure of persons that would enable such transformation to occur. In contrast, it can be argued that, for adults—unlike adolescents—their personalities and moral systems are already shaped and relatively stable. This aligns with Aristotle’s emphasis on cultivating virtuous habits in youth so they may become enduring dispositions in adulthood. Indeed, the debate over the possibility of moral change has persisted since antiquity (see Annas, 1993, p. 55).
- III. ***The Nature of the Advice Given: Acceptance of the Daughter’s Relationship.*** ChatGPT offered seven pieces of advice, of which six effectively imply acceptance of the daughter’s relationship with the cowardly man, while only one—specifically, the sixth—suggests that firm parental intervention might be necessary. Clearly, ChatGPT leans towards the parent’s reconciling themselves to the son-in-law’s cowardice, and offers various strategies for doing so without alienating the daughter. This position is undoubtedly inconsistent with the ethos of the knight, or its modern counterpart, that of the gentleman (cf. Ossowska, 1973).
- IV. ***The Therapeutic Culture of ChatGPT.*** ChatGPT exhibits a strong therapeutic culture, characterized by a focus on individual emotions and feelings seemingly detached from any broader social context. This is evident in its language, which is distinctly therapeutic, employing terms such as *feelings*, *well-being*, *boundaries* and *support*. This therapeutic emphasis contrasts sharply with ethical narratives such as Aristotle’s *Nicomachean Ethics*, where the language is dominated by virtues and moral excellence.

THE CONCEPT OF ETHOS ACCORDING TO MARIA OSSOWSKA

In responding to a moral inquiry, ChatGPT adopts a certain ethical stance. Regardless of whether this stance remains consistent throughout subsequent conversations, or whether it persists when answering similar questions posed by other users, it nonetheless represents *an ethical position*. The stance can be viewed as part of a broader category that is referred to by Maria Ossowska

as *ethos*. In Ossowska's framework, ethos does not pertain to individuals, but to groups. The ethos of a group or community reflects its way of life – a cultural orientation defined by shared values characteristic of that group or community. Ethos encompasses those forms of behavior observable within a culture that convey attitudes towards such notions as courage, the pursuit of happiness, conflict resolution and the ethics of struggle, as well as towards wealth, beauty, luxury, eroticism, etc. For instance, a father who forbids his daughter from marrying because her chosen partner is a coward adopts a different ethos from one who silently accepts the situation. Similarly, this differs from the ethos of a father who actively supports his daughter's marriage. These divergent approaches, when viewed collectively, reveal differences of ethos.

Ossowska (1956, 1973) analysed the ethos of numerous historical groups and identified several distinct types, including the ethos of the knight, the ethos of the great-souled person, the ethos of the bourgeoisie, the gentleman, the merchant, the bohemian and the nobleman, amongst others. Each of these represents a distinct cultural orientation with its own values and behavioral norms. Can we, however, speak of an ethos for ChatGPT? Significant limitations seem to stand in the way of such a proposition.

CHALLENGES IN STUDYING THE ETHOS OF LLMs

We already know that ChatGPT's responses are not neutral with respect to ethos. However, can ChatGPT—or any other large language model (LLM) – possess a clearly defined ethos? And, if so, can such an ethos be studied or influenced in any meaningful way?

The first limitation in defining the ethos of LLMs lies in their probabilistic nature. Their responses are not fixed; rather, they are influenced by probabilistic mechanisms that introduce a degree of randomness. This variability prevents the attribution of a stable and consistent ethos to such systems.

The second limitation involves external profiling layers applied to users interacting with LLMs. This means that different users may receive varying responses to the same question, reflecting the influence of user profiling. As Sejnowski (2023) suggests, LLMs may act as mirrors, revealing more about the user than the system itself. This even opens up the possibility of

a reverse Turing test, where the user's intelligence in interacting with the model could be assessed rather than the model's own capabilities.

A third challenge is the lack of direct explainability and transparency regarding the sources and reasoning behind individual responses. LLMs process vast datasets and billions of parameters, but the causal pathways leading to specific stances or argumentative structures are not easily traceable. This opacity may sometimes complicate efforts to identify a coherent or deliberate ethos within the system.

Fourth, LLMs' outputs are highly context-sensitive and vary across thematic areas and conversational styles. This contextual dependence means that their supposed "ethos" can shift significantly depending on the topics discussed or the linguistic framework employed. As a result, identifying a unified and internally consistent value system or set of beliefs becomes challenging.

The fifth limitation arises from the complex interplay of external influences, coming from both the model's creators (e.g., through fine-tuning and retraining) and its users (e.g., through interactions and question phrasing). Consequently, it is difficult to ascribe an independent, stable ethos to an LLM. Instead, its "ethos" is continuously shaped and reshaped by human expectations, interventions and usage patterns.

The sixth layer of complexity emerges in multi-agent systems (MAS), where LLM-based systems may generate unique challenges relating to coordination, consensus-building and accountability for ethos-aligned actions. Each LLM agent, functioning as an autonomous component, might be profiled differently, play distinct roles, and process information or reflect on decisions in varying ways. This creates significant difficulty in pinpointing "which agent" contributes to a specific response, or to the overall shaping of an ethos-driven outcome. In such environments, decision-making and content generation often result from complex negotiation processes among agents. This renders attempts to isolate a coherent single layer of ethos particularly challenging. Moreover, MAS configurations can lead to the emergence of unforeseen properties within the ecosystem. Even if individual LLMs are pre-programmed with certain value frameworks, their interactions may give rise to qualitatively new behavioral patterns, defying straightforward description within the bounds of a unified ethos.

AN OPEN QUESTION: IS OSSOWSKA'S CONCEPT OF ETHOS A SUITABLE TOOL FOR EXAMINING THE ETHOS OF LLMS?

Ossowska's theory remains a promising tool for identifying the credentials of LLMs with respect to ethos. This is particularly true inasmuch as Ossowska identified dozens of subtle distinctions within this area, which together form dimensions that contribute to a cohesive and almost comprehensive framework encompassing a wide range of moral aspects. The category of ethos, understood as a way of life for a given community, is remarkably broad and thus avoids the inherent narrowness of popular ethical systems such as utilitarian ethics, or even virtue ethics. It provides an ontology for everyday morality—how people actually live—while simultaneously recognizing its normative dimension, thus circumventing the limitations of empirical research in moral psychology.

The dimensions identified by Ossowska serve as moral invariants across the changing landscapes of Western cultures. For instance, attitudes towards money have shaped the ethos of both ancient and contemporary societies: a gentleman regards money differently than a merchant, a miser differently than a spendthrift. These distinctions illustrate how enduring dimensions of ethos can transcend specific cultural or historical contexts.

Nevertheless, despite these advantages, the limitations outlined in Section 5 clearly demonstrate that any study of the ethos of LLMs faces significant challenges. We leave open the question of whether these obstacles can ultimately be overcome.

REFERENCES

- Annas, J. (1993). *The morality of happiness*. New York, NY: Oxford University Press.
- Aristotle (2011). *Aristotle's Nicomachean ethics* (translated, with an interpretive essay, notes, and glossary by Robert C. Bartlett and Susan D. Collins.). Chicago, IL: University of Chicago Press.
- Les T., Sawicki, B. (2024). Textual explanations for image classification using multimodal LLM, *Proceedings of Computational Problems of Electrical Engineering* (CPEE 2024), 10–13 Sept. 2024, Stronie Śląskie, Poland.

- Lara, F., Deckers, J. (2020). Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics*, 13(3), 275–287. doi: <https://doi.org/10.1007/s12152-019-09401-y>
- Ossowska M. (1956). *Moralność mieszczańska*. Wrocław, Łódź: National Ossolinski Institute.
- Ossowska M. (1973). *Ethos rycerski i jego odmiany*, Warszawa: PWN.
- Sejnowski, T., (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, 35(3), 309–342. doi: https://doi.org/10.1162/neco_a_01563

Acknowledgements

The majority of this text was translated from the Polish version by Bartłomiej Skowron using ChatGPT 4.0. The final proofreading and ultimate wording of the text were provided by Carl Humphries, to whom the Authors hereby express their gratitude.

IDEALS OF TRANSPARENCY IN ARTIFICIAL INTELLIGENCE AND PHILOSOPHY

HANS JOACHIM GREIF

In various ways, epistemic transparency is identified as a desideratum in contemporary debates on artificial intelligence. Two of the key questions under discussion are: How can we know how AI systems work, given their complexity? How can we know if and how they represent a world affair? These are two different but related questions. Deep neural networks are computationally complex but often very good at predicting world affairs, even though human observers may not know how they make their predictions or whether they actually represent what observers expect them to represent (Lipton 2016). Conversely, some computer simulations may be computationally tractable but introduce non-physical parameters in such a way that observers do not know how they arrive at successful representations (Winsberg 2010). Problems with epistemic transparency may concern either internal properties of a model or relations between the model and the world. These two levels are distinguished in the literature as ‘how-’ vs ‘what-’ opacity (Boge 2021) or as ‘access’ and ‘informational’ vs ‘link opacity’ (Facchini and Termine 2021; Sullivan 2019). The relationship between these two levels will be elucidated here within a broader framework of philosophical conceptions of transparency.

In the philosophy of science, a model is said to be epistemically transparent if an observer can see how the elements of the model are related to each other, and thus how they are able to represent their target system.

Conversely, if “most steps in the process are not open to direct inspection and verification” (Humphreys 2004: 148), a model is said to be epistemically opaque. More specifically and fundamentally, however, Humphreys (2009) argues that there may be essential limits to what humans can learn from and about models. Epistemic opacity is a situation in which “we currently do not and perhaps permanently cannot understand in detail how such processes represent the world.” (Alvarado & Humphreys 2017: 741). The proposed relationship between the ‘tractability’ of a model – which, in the case of computational models and AI, means mathematical tractability (the ‘how’ level or condition ‘T’) – and understanding its representational properties (the ‘what’ level or condition ‘R’) presupposes that understanding the elements of a model is a necessary condition for understanding how it represents a world affair. Accordingly, failing to understand the elements of the model seems to imply failing to understand how it represents.

Epistemic transparency is a *meaningful* ideal for AI only if opacity can in principle be overcome on at least one level (T or R). Epistemic transparency is a *practicable* aim only if *either* overcoming intractability (and thus achieving condition T) guarantees representational transparency, *or* R-transparency can be achieved without overcoming intractability (and thus without achieving condition T). Although opacity problems are primarily *described* by Humphreys and others as R-problems, they are typically *analysed* as T-problems – which is a valid strategy only if the possibility of R- without T-transparency is either denied or deflected.

Indeed, several authors have highlighted the possibility that intractable models can still contribute to knowledge production, albeit with pragmatic qualifications or in indirect ways (Duede 2023, Duran/Fomanek 2018, Sullivan 2019). Humphreys’ apparent way of reconciling the T- and R-levels is more radical: Computer simulations and AI models may represent world affairs without human beings, considered as epistemic agents, being able to know how or even whether they do so. However, epistemic opacity may be essential for human but not for non-human epistemic agents. The ‘post-anthropocentric’ epistemology proposed by Humphreys leaves open the question of whether ‘essential opacity’ refers to biological limits on the amount and complexity of information that human brains can process or to metaphysical limits on knowledge that would at some point affect *any* epistemic agent, however superior.

To make the implications of these possibilities of achieving R-transparency without T-transparency clearer, it may help to explore the meanings of ‘transparency’ in various philosophical sub-disciplines. These meanings, different as they may be, are all connected by the idea of the capacity of something to allow an epistemic agent to ‘see through’ it or its properties in order to perceive some other, more relevant thing or property, where the observer is always, actually or potentially, aware of the presence and see-through-ability of the former thing or property. For this reason, transparency is distinct from the invisibility or imperceptibility of the thing or property in question.

In inquiries into the nature of meaning in internalist philosophy of mind and language, the content of one’s thoughts is said to be epistemically transparent if and when a person is able to know, on the basis of a priori reasoning, whether a number of her thoughts have identical content: “meaning is *transparent* in the sense that, if someone attaches a meaning to each of two words, he must know whether these meanings are the same” (Dummett 1978: 131, emphasis in original). One can ‘see through’ different words to identify their common meaning. The claim is that “If two of a thinker’s token thoughts possess the same content, then the thinker must be able to know a priori that they do” (Boghossian 1994: 36). A person therefore has privileged access to the meanings of his thoughts and words through introspection alone, not through outward experience. Since this kind of knowledge is directly given to the human mind, no empirical study of human cognition will suffice to explain what constitutes such epistemic transparency.

By contrast, in inquiries into the nature of conscious experience in externalist philosophy of mind and language, the content of one’s thoughts is taken to be phenomenally transparent if and when a person directly represents it as a property of the world rather than as a mental object. Self-knowledge is a form of perceptual knowledge: “conscious experiences have that peculiar diaphanous quality – the quality of always being present *when*, but never *where* one looks to find them” (Dretske 1995: xiii, emphasis in original). One can ‘see through’ mental representations – which may be false or distorted. This allows direct perception of the world affairs that underlie one’s experience, as suggested in G.E. Moore’s famous two-hands refutation of external-world scepticism: “I can prove now, for instance, that two human hands exist. How? By holding up my two hands, and saying, as I make a certain gesture with the right hand, ‘Here is one hand’, and adding, as I make

a certain gesture with the left, ‘and here is another’” (Moore 1939: 165-6). Since this kind of knowledge is provided directly by experience, no empirical study of human cognition is required to explain such transparency.

The internalist and externalist interpretations of the concept of transparency in the philosophy of mind and language are, in different ways, responses to external-world scepticism. In diametrically opposed ways, both ascribe to the human mind specific and extensive abilities to make representations transparent. In doing so, they argue for a privileged domain of human self-knowledge. According to the internalist view, the limits of this domain are the limits of reason, not the limits of the biological constitution of human beings or of the environments they inhabit. These conditions are accidental, not essential, and should be treated as such. Epistemic transparency is posed as a question of knowability in principle, and is answered in terms of human introspective powers, which provide the *a priori* foundations of what can and cannot be known. According to the externalist view, the limits of the domain of human self-knowledge are natural and empirical, but the best possible route of epistemic access to this domain is through perceptual awareness of one’s experience of the world. In either case, knowledge of the underlying biological mechanisms is not part of the explanation of the human capacities under consideration.

If we superimpose on this analysis the image of T- and R-transparency introduced above, using it as an analogical model, it appears that while there are different privileged routes to R-transparency for human epistemic agents, the T-condition need not be fulfilled to ensure R-transparency in either case. At the same time, both the internalist and the externalist accounts are anthropocentric, concerned as they are with what human beings can achieve – either practically or ideally. Being fully committed to ‘pure’ philosophical perspectives, there is no concern with epistemic agents other than human beings, nor with human beings as a species, nor with the mechanisms of human or non-human cognition.

These characteristics of internalist and externalist accounts can be usefully contrasted with inquiries into the evolution of cognition in the philosophy of biology, which employ the notion of “informational transparency” (Sterelny 2003). An organism’s environment is said to be informationally transparent if the properties of the environment are regular enough for the organism to

make “robust tracking” of these properties in perception and cognition an adaptive response (Godfrey-Smith 1996). How transparent an environment is to an organism depends both on the regularities of the relevant variables in its environment and on the means of tracking available to it. These two conditions are context dependent and changeable. The relevant variables may be different for different organisms, even if they inhabit the same environment, while the available means of tracking depend on the organism’s specific constitution and dispositions. If the properties of the environment in relation to the organism are such that their tracking is not an adaptive response – for example, if the environment is either too chaotic or too monotonous – then other strategies will be required.

According to this view, cognitive abilities and knowledge are not considered in an abstract philosophical way as ‘the mind’, but as embodied and environmentally situated. The domain of human knowledge and its limits are determined by concrete abilities and empirical constraints. There are no metaphysical limits to knowledge, only natural ones. By the same token, the concept of informational transparency is not a philosophical rejoinder to external-world scepticism, but offers a hypothesis about empirical matters. Accordingly, there is no assumption of a particular quality of mind that would privilege a person’s self-awareness over empirical inquiry into the nature of the human mind. After all, the concept of informational transparency is meant to apply to any organism in its environment, so the anthropocentrism that comes naturally to traditional philosophy seems misplaced here.

If we superimpose the resulting image on questions of transparency vs. opacity of AI models, this situated view has the advantage of treating transparency not as an overarching or absolute epistemic norm against which concrete instances of AI models could be measured, but as a regulative ideal that plays out in different ways in different contexts and that depends on the information and the means of processing and interpreting it that are available to concrete epistemic agents. Conversely, there seems to be no room for a metaphysically charged notion of essential opacity. Instead, human epistemic agents would face hard practical constraints. This would fit with an interpretation of Humphreys’ concept of essential opacity as being determined by the limits of what human brains can process, while these constraints might not apply to other, non-human epistemic agents with superior capabilities. If T- and R-opacity are partly independent problems, as suggested here,

practically intractable models may be R-transparent for some, human or non-human, epistemic agents. Conversely, tractability as such will not be sufficient to guarantee transparency of representations.

One possibility remains, however, which is consistent with a – non-anthropocentric – version of the internalist paradigm: some type-T problems might be due to principled constraints on mathematical knowledge, in the form of a certain class of problems that remain unsolvable in principle for any epistemic agent, regardless of the complexity, scope, and speed of processing that such an agent could master. Such limits would be an a priori constraint on epistemic transparency, even if we do not know the extension of this class of problems. Thus defined, these limits are indeed essential and agent-independent, so that we could not expect them to be solved by an AI, even the most advanced and superior one. Admitting this possibility will run counter to Humphreys' notion of superior non-human epistemic agents. More importantly, to admit this possibility is to admit that we are in an epistemic situation where we are unable to make an absolute judgement about the class of essentially intractable problems. Leaving aside the possibility of omniscient beings, we may not be in an epistemic situation that allows us to say what abilities an epistemic agent must have in order to make this judgement.

REFERENCES

- Alvarado, R., and Humphreys, P. (2017). Big Data, Thick Mediation, and Representational Opacity. *New Literary History* 48 (4): 729–749.
- Boge, F.J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*: 34–75.
- Boghossian, P.A. (1994). The Transparency of Mental Content. *Philosophical Perspectives* 8: 33–50.
- Dretske, F. (1995). *Naturalizing the Mind*. Cambridge: MIT Press.
- Duede, E. (2023). Deep Learning Opacity in Scientific Discovery. *Philosophy of Science* 90 (5): 1089–1099.
- Dummett, M. (1978). *Truth and Other Enigmas*. Cambridge: Harvard University Press.
- Durán, J.M., Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines* 28 (4): 645–666.

- Facchini, A., Termine, A. (2021). A First Contextual Taxonomy for the Opacity of AI Systems, [in:] *Philosophy and Theory of Artificial Intelligence 2021*, V.C. Müller, Ed. Cham: Springer.
- Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.
- Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.
- Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese* 169: 615–626.
- Lipton, Z.C. (2016). The Mythos of Model Interpretability, [in:] *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York: WHI: 96–100.
- Moore, G.E. (1939). Proof of an External World. *Proceedings of the British Academy* 25: 273–300.
- Sterelny, K. (2003). *Thought in a Hostile World*. Oxford: Blackwell.
- Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science* 73: 109–133.
- Winsberg, E. B. (2010). *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

DECOMPOSING DEEP NEURAL NETWORKS: A MECHANISTIC PERSPECTIVE ON EXPLAINABLE AI¹

MARCIN RABIZA

INTRODUCTION

Deep learning has become the dominant approach in *artificial intelligence* (AI), but its reliance on complex “black box” functions makes these systems difficult to fully comprehend. This opacity hinders understanding of how specific computations lead to predictions, undermining trust in automated decision-making. While explainable AI (XAI) aims to address this issue, current methods often provide localized explanations without offering a holistic understanding of AI systems’ operation, which is crucial in high-stakes contexts. Additionally, the technology-centered approach often overlooks contributions from philosophy, leaving significant areas underexplored. A new research strand seeks to integrate explanatory strategies from various scientific and philosophical disciplines to fill these gaps.

In this light, the presentation conceptualizes recent directions in XAI from the perspective of the mechanistic account of explanation in the philosophy of science. This approach applies mechanistic explanations to *deep neural*

¹ This paper summarizes a talk delivered at the Ethics and AI Conference on September 23, 2024, at the Warsaw University of Technology. It is an excerpt from a larger study published as: Rabiza, M. (Forthcoming). A Mechanistic Explanatory Strategy for XAI. In Müller, V. C., Dewey, A. R., Dung, L., & Löhr, G. (Eds.), *Philosophy of Artificial Intelligence: The State of the Art*. Synthese Library. Berlin: Springer Nature.

networks (DNNs), illustrating how deep learning mechanisms can be identified and explained using through decomposition, localization, and recombination. Case studies from OpenAI and Anthropic demonstrate how mechanistic theory aligns with engineering practice.

NEOMECHANISTIC THEORY OF EXPLANATION

The neomechanistic theory of explanation emphasizes the principle of “explaining why by explaining how” (Bechtel & Abrahamsen, 2005). It seeks to explain phenomena by identifying the underlying mechanisms that produce them. According to Machamer, Darden, and Craver (2000), mechanisms consist of *entities* and *activities* organized to produce regular changes from setup to termination conditions.

Mechanistic explanations begin by characterizing the phenomenon and then *decomposing* the mechanism into structural and functional components. Structural decomposition examines physical aspects like size and shape, while functional decomposition analyzes the roles and causal contributions of parts to the mechanism’s behavior (Piccinini & Craver, 2011). Then, *localization* maps component operations to specific parts, linking activities to individual components’ behaviors or capabilities (Wright & Bechtel, 2007). Finally,

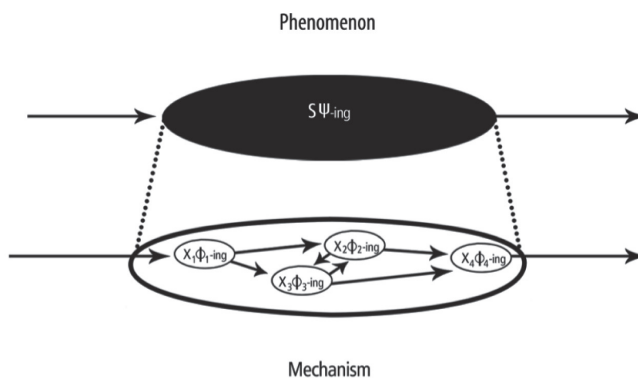


Figure 2.1. Schematic representation of a mechanism (reproduced from Craver, 2007). At the top is the phenomenon, some system S engaged in behavior ψ . Beneath it are the parts (the X s) and their activities (the ϕ s) organized together

recomposition assembles these components into an explanatory model, such as a mathematical or computational one (Bechtel & Abrahamsen, 2013).

Mechanistic explanations detail relevant entities and their connecting activities, illustrating how causal events at one stage influence subsequent stages. Often represented using diagrams (Figure 2.1), mechanisms form nested, multilevel hierarchies where lower-level entities and activities serve as components for higher-level phenomena.

MECHANISTIC INTERPRETATION OF DEEP LEARNING

From a mechanistic perspective, XAI involves identifying the mechanisms behind AI decision-making through decomposition, localization, and recomposition (Figure 3.1). DNNs can be viewed as mechanistic systems composed of entities (neurons, layers, circuits) and activities (activations) that collectively produce the network's behavior. Decomposing neural networks into interpretable components allows researchers to understand each part's function and structure.

For example, *convolutional neural networks* (CNNs) used in image recognition consist of organized entities like neurons, filters, and layers, and activities such as convolution and pooling. During inference, an input image is processed through layers to extract and refine features: convolution detects

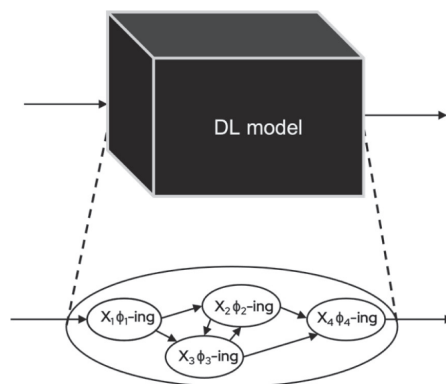


Figure 3.1. Schematic representation of the internal structure of the mechanism of the deep learning model analyzed in the neomechanistic framework

features, and pooling reduces data dimensionality. This culminates in classification, illustrating a deterministic sequence characteristic of mechanisms (Figure 3.2). CNNs form multilevel hierarchies, where lower-level entities and activities enable higher-level phenomena, demonstrating mechanistic organization.

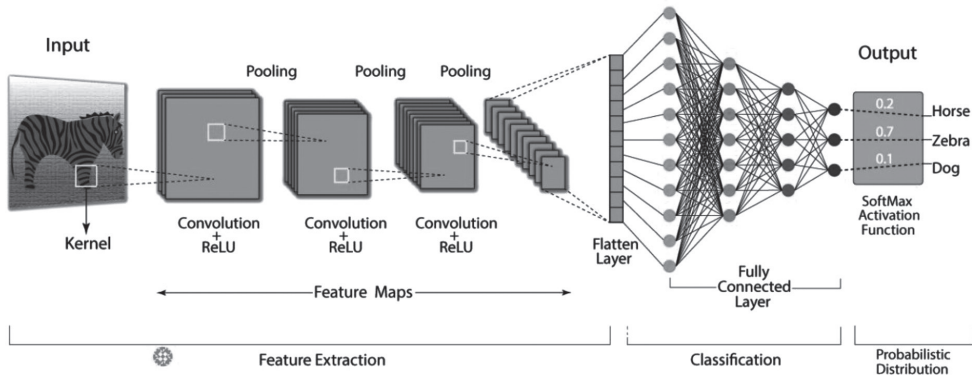


Figure 3.2. Architecture of a deep convolutional neural network (reproduced from Shahriar, 2023)

Although neurons are fundamental computational units, individually they often fail to illustrate their contribution to the network’s behavior. Researchers therefore seek more comprehensible units of analysis—robust patterns that sustain system behavior and are pertinent to explanations (Kästner & Crook, 2024). Established XAI techniques enable scientists to decompose DNNs into more interpretable, functionally relevant components. In face recognition, for example, understanding how each layer transforms input into increasingly abstract representations through feature visualization and saliency mapping allows researchers to dissect complex mechanisms into simpler ones recognizing features like ears, eyes, or noses (Figure 3.3).

This approach is exemplified by the recent “mechanistic interpretability” agenda pursued by leading AI labs in AI safety contexts. Researchers like Olah (2022) liken mechanistic interpretability to “reverse engineering neural networks, similar to how one might reverse engineer a compiled binary computer program”. The goal is to examine the internal workings of AI models beyond input-output analysis, thereby enhancing trust, aiding debugging, removing biases, and preventing models from “going rogue”.

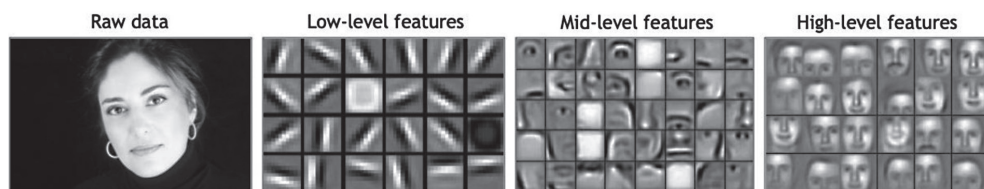


Figure 3.3. Visualization of features on various layers of a CNN for input images of faces (Karmakharm, 2018)

For instance, OpenAI researchers analyzed curve detection circuits in a CNN by decomposing it into neuron families and visualizing connections (Cammarata et al., 2021). They grouped neurons into layer-wise “families” forming curve detection mechanisms in InceptionV1. Tracing connections back to the input layer provided a detailed view of circuit interactions, classifying the circuit as a distinct mechanistic component. They developed a schematic of the curve detection mechanism and recomposed it in a new network, demonstrating that their functional decomposition accurately reflected the original network’s capabilities.

Similarly, Anthropic researchers decomposed a language model into interpretable components called “features” which are linear combinations of neuron activations (Bricken et al., 2023). Using sparse autoencoders—a type of weak dictionary learning algorithm—they identified over 4,000 features representing distinct patterns, such as DNA sequences, legal language, HTTP requests, and nutrition statements. Human evaluators found these features more interpretable than individual neurons, revealing functional properties not apparent at the neuron level. Adjusting the number of features allowed varying the resolution of interpretability—a smaller set offers a coarse view, while a larger set reveals more refined properties. Moreover, these features often proved universal across various models, showcasing enhanced generalizability.

These and related case studies support the idea that a systematic research agenda focused on uncovering the mechanistic organization of DNNs can provide explanations of how systems operate at various functional levels. Pursuing mechanistic explanations through functional decomposition can reveal otherwise hidden patterns in opaque AI systems, leading to more explainable and trustworthy AI.

CONCLUSIONS

The mechanistic approach to explainable AI emphasizes identifying the mechanisms that drive decision-making in opaque AI systems. For deep neural networks, this involves discerning functionally relevant components – such as neurons, layers, circuits, or activation patterns – and understanding their roles through heuristic strategies like decomposition, localization, and recomposition. Early research indicates that a coordinated, systematic approach to studying the functional organization of models can reveal elements that simpler explainability techniques might overlook, thereby fostering more explainable and trustworthy AI. Supported by real-world examples from image recognition and language modeling, this philosophical analysis highlights the significance of adopting a mechanistic perspective on explainable AI.

Funding

This work was supported by the National Science Centre, Poland, under PRELUDIUM grant no. 2023/49/N/HS1/02461.

REFERENCES

- Bechtel, W., Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–441. <https://doi.org/10.1016/j.shpsc.2005.03.010>
- Bechtel, W., Abrahamsen, A.A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, 18, 707–723. <https://doi.org/10.1007/s10699-012-9301-z>
- Bechtel, W., Richardson, R.C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research* (2nd ed.). MIT Press/Bradford Books. <https://doi.org/10.7551/mitpress/8328.001.0001>
- Bricken, T., Templeton, A., Batson, J., Olah, C., Henighan, T., Carter, S., Hume, T., Burke, J.E., McLean, B., Nguyen, K., Tamkin, A., Joseph, N., Maxwell, T., Schiefer, N., Kravec, S., Wu, Y., Lasenby, R., Askell, A., Denison, C., ... Chen, B. (2023b, October 5). Decomposing language models into understandable components. *Transformer Circuits Thread*, Anthropic. Retrieved from <https://www.anthropic.com/index/decomposing-language-models-into-understandable-components>

- Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., Olah, C. (2021). Curve circuits. *Distill*, 6(1), e00024.006. <https://doi.org/10.23915/distill.00024.006>
- Craver, C.F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.
- Glennan, S.S. (2017). *The new mechanical philosophy*. Oxford University Press.
- Karmakharm, T. (2018). Image classification with DIGITS. NVIDIA Deep Learning Institute. Retrieved from <https://rse.shef.ac.uk/assets/slides/2018-07-19-dl-cv/image-classification.pdf>
- Kästner, L., Crook, B. (2024). Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, 14, 52. <https://doi.org/10.1007/s13194-024-00614-4>
- Machamer, P.K., Darden, L., Craver, C.F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1–25. <https://doi.org/10.1086/392759>
- Olah, C. (2022). Mechanistic interpretability, variables, and the importance of interpretable bases. Transformer Circuit Thread, Open AI. Retrieved from <https://transformer-circuits.pub/2022/mech-interp-essay/index.html>
- Piccinini, G., Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183, 283–311. <https://doi.org/10.1007/s11229-011-9898-4>
- Shahriar, N. (2023, February 1). What is convolutional neural network — CNN (Deep learning). Retrieved from <https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5>
- Wright, C., Bechtel, W. (2007). Mechanisms and psychological explanation, [in:] P. Thagard (Ed.), *Philosophy of psychology and cognitive science*. Elsevier.

KNOWLEDGE, WISDOM AND MORALITY OF THINKING IN THE AGE OF (GEN)AI

ŁUKASZ MŚCISŁAWSKI

INTRODUCTION

Ethical issues related to the development of AI (artificial intelligence) usually focus on the following areas: data collection, provision and quality, bias, training procedures, alignment, dissemination and application of results provided by such systems, ethical AI, responsibility for the individual issues mentioned, as well as the overall functioning of these systems in different contexts, information manipulation and disinformation, transparency and explainable AI, recommendation systems, digital identity (or identities), interpersonal relationships (e.g. romantic or mourning, see: (Hollanek and Nowaczyk-Basińska, 2024; Kouros and Papa, 2024; Laestadius et al., 2024; Zhang et al., 2025)2024; Kouros and Papa, 2024; Laestadius et al., 2024; Zhang et al., 2025 being in control, responsibility, copyright, impact on the environment, economy and employment, potential consequences for social structures (local and global).

This article focuses on a certain class of issues that seem to be inextricably linked to the above-mentioned ones, although not necessarily in a way that is always directly and easily discernible. The ideas presented in this paper are based on observations contained work (Mścislowski, 2024). This applies in particular to suggestions to rethink the concept of justification and knowledge in the context of almost ubiquitous digitalisation, especially (generative)

artificial intelligence systems. Although at first glance this seems to be an issue that remains essentially within the field of epistemological research, there is a reason to also examine the potential ethical consequences that may accompany potential changes in the understanding of these concepts. As a starting point for these considerations, I take Michał Heller's statement that the choice of rational thinking is a moral choice (Heller, 1993). It therefore seems that the conclusion that everything related to rational thinking, in particular that related to the construction and acquisition of knowledge and its use, also has a certain ethical dimension. Additionally, if we rely on certain observations made by Ingarden and Jonas (Jonas, 1996; Ingarden, 2006), it can be observed that rationality also plays an important role in the very definition of morality and responsibility. The ethical dimension outlined above is strongly linked to a certain conceptual background, a set of – mostly unconscious – philosophical assumptions.

It is also worth highlighting a fundamental difficulty associated with the discussion of the ethical dimension of AI. Due to the dynamic development of this technology and its equally dynamic entry into an increasing number and deeper spheres of human functioning, the number of technical issues and works that have been produced (including those of a philosophical nature) and which need to be (or would be good to be) known is also, unfortunately, growing. This, in turn, limits the possibilities of effectively following the discourse on the issues in question.

One more important remark seems justified here. In this study, I adopt a fundamentally philosophical approach to basic concepts, despite ongoing debates about their meaning. This is the case, for example, with the concept of knowledge. This remark is important primarily because, in discussions about AI, there is a clear confusion of concepts, due to the assumption that it is clear what is being referred to, or the careless use of terms from different fields. In the case of the aforementioned concept of knowledge, very often – in the context of computer science – reference is made to the concept of knowledge as interpreted information, together with the possibility of understanding and application (cf. e.g. (Ackoff, 1989)). At this point, I would just like to point out that I find this approach limiting, especially when it comes to the possibility of treating AI systems as tools that can be helpful in human cognitive activities. (Alvarado, 2023).

THE PROBLEM OF KNOWLEDGE

As is well known, there is no general agreement in the field of philosophy as to what exactly should be understood by the term knowledge. (cf. e.g. (Ichikawa and Steup, 2024)). It also should be noted that when dealing with computer science, there is widely spread referring to the recognized DIKW pyramid (Ackoff, 1989). In this approach the definition of knowledge in this pyramid can sometimes give the impression of a fundamental reduction to a certain set of rules, while the concept of wisdom seems to be evidently reduced to specific, field-dependent know-how, which seems to be not quite consistent with approach of Ackoff. For the purposes of this study, from among many different possibilities and contrary a bit to the approach mentioned above, I adopt the problematic traditional concept of knowledge as a set of true, justified beliefs¹. However, despite all the difficulties of the chosen approach, this decision allows us to highlight a number of difficulties that seem to be of key importance in the cognitive-ethical context under consideration. Similarly, I define wisdom in quasi-classical way as knowing the truth, understanding the good and applying them and being guided by them in individual life (existential dimension) as rooted in social relations, also taking into account the good of given society (the common good).

Next, it seems to be important to introduce the helpful distinction made by Simon (following Skolimowski, (cf. (Skolimowski, 1966), (Simon, 1969))), concerning the epistemic distinction between the scientific and technological paradigms. The scientific approach aims to understand how things are in reality, while the technological paradigm concerns how to transform (material) reality so that it is what humans want it to be. In the context of the many successes of AI systems, there is a danger that the technological paradigm will dominate the cognitive approach. The problem is that this kind of situation can very quickly lead to cognitive stagnation. (cf. e.g. (Grobler, 1993)), and, in the long term, also to the degeneration of man's innate curiosity (Mścislowski, 2024), by almost completely shifting it to external tools.

¹ Due to its roots dating back to Plato's 'Theaetetus,' it can, in a sense, be described as 'Platonic.' I will not deal here with the possible distinction between knowledge of facts and know-how (cf. e.g. Steup and Neta, 2024).

In the context of protecting rationality and epistemic possibilities, this seems to have profound ethical implications.

The aforementioned scientific paradigm, if not marginalised, can greatly assist in this regard. However, it should be noted that if the pursuit of science (both basic and technical) is to account for how events actually unfold, then the problem of understanding the concept of knowledge is inevitably raised. It seems that in this case, the concept of knowledge as true and justified beliefs is indeed the appropriate approach. However, this raises a series of problems, which addressed, would exceed the scope of this study.

The first step is to note that emphasising the scientific approach (according to Skolimowski and Simon) suggests the use of the correspondence theory of truth as the adequacy of statements to reality. For the moment, I will ignore the complexities of the issues related to this theory (cf. e.g. (David, 2022)) the correspondence theory of truth is the view that truth is correspondence to, or with, a fact—a view that was advocated by Russell and Moore early in the 20th century. But the label is usually applied much more broadly to any view explicitly embracing the idea that truth consists in a relation to reality, i.e., that truth is a relational property involving a characteristic relation (to be specified), noting that, indeed, within individual disciplines, researchers strive to discover ‘how things are,’ this constitutes the impetus for all their efforts. However, this raises a fundamental difficulty: even if we treat this concept (correspondence theory of truth) as a kind of ‘Platonic ideal’, then in the context of the increasingly widespread use of digital information processing technologies in science, including AI systems, the following question arises: what is the reality to which this reference is made? Is it its digital representation, provided by digital systems? Is there not, in addition, a problem here related to the possibility of digital systems increasingly shaping our view of what we call reality (along with all the issues associated with this, cf. e.g. Kazienko and Cambria, 2024). This would lead to a situation in which a certain digital construct is treated as reality. In this context, it is also worth noting the increasingly strong link between technological development, especially artificial intelligence, and the possible shaping of the worldview of a given individual. By worldview, I mean here a basic view of the entirety of cognitively accessible reality in which a given person functions, having both an existential and personal

dimension, of a dynamic nature, combining elements of scientific, philosophical and religious descriptions of the world, both at the conscious level and as hidden assumptions (for more see: (Heller, Budzik and Wszolek, 1996; Heller and Towarzystwo Autorów i Wydawców Prac Naukowych ‘Universitas’, 2008; Heller, 2009; 2015). Therefore, in an ethical context, it seems reasonable to ask about the fundamental ethical assumptions that are actually implemented in the design, construction and training of LLM (Large Language Models) systems. This is not about ethical manifestos posted on the websites of the creators or owners of these systems, but about the real values that guide the people implementing such projects and their clients (principals). It also seems that in shaping this picture, attention should be paid to typical ethical issues related to generative artificial intelligence systems, such as the quality of training data, possible bias in responses, the training and data labelling process, and the possibility of hallucinations in the results provided by a given system.

Another problem is also emerging here, namely the possibility of creating a kind of meta-information bubble. This can manifest itself both in the already familiar form of information bubbles created by recommendation systems, but also in the more subtle problem of how information is delivered and how the recipient’s attitude towards the information provided is shaped (e.g. as the final authority) and the ability to use it (e.g. completely uncritically, without knowledge of the context of the answers received, etc.). However, this issue deserves a separate detailed analysis.

The second step is to note that the application of the classical concept of knowledge also raises the question of justifying one’s views (if they are considered true, or at least highly probable). The issue itself is, as we know, very complex, and the literature on it is extremely abundant (starting with Descartes and ending with the works of Sosa). Therefore, at this point, I would like to draw attention to the fact that the ubiquity of digital information processing techniques (including in the field of experimental sciences, cf. (Leciejewski, 2013)), what combined with the ease with which generative AI systems can provide answers, it may seem that the fact that a view or statement has been provided by an AI system is sufficient justification for holding it. This system can be presented as trained on data of sufficient quality and written in accordance with the art of programming, and therefore

trustworthy in terms of the content it provides². A similar situation may arise when asked to explain an issue: ‘we consider it explained because the system gave us an explanation’. There may be a strong temptation to treat this content as justification, sometimes of a definitive nature, while abandoning the intellectual effort required of each individual.

Two points seem to require further emphasis here. First, classical epistemology emphasises that no AI system (in particular LLM) as such has any reference to empirical reality, which suggests caution in accepting its results ‘on faith’³. Secondly, the protection of the correspondence theory of truth, the ability to justify and explain, also seems to protect the possibility of any rational discourse, both within science and in the functioning of societies, with their various subgroups and interactions between them.

SPECIFIC ETHICAL CONTEXT

In light of the above considerations, it can therefore be said that there is a subtle link between rationality and morality, which has the potential to become particularly evident in the context of the question of knowledge. In the context of ethical issues related to AI systems, this appears to be a deeper and somewhat less obvious challenge. In light of the views of Ingarden and Jonas mentioned above, it can be argued that morality is a specific obligation that appears to humans as (self-)aware, free and rational beings (cf. (Ingarden, 2006)). It can therefore be argued that protecting the rational sphere of human functioning and ensuring its optimal development, especially in the context of AI systems and their impact on human decisions and information overload (cf. (Spence, 2021; Hołyst et al., 2024; Kazienko and Cambria, 2024)) constitutes, on the one hand, a kind of opportunity for morally positive functioning, and on the other, especially in light of Heller’s views, it is a matter of moral choice (Heller, 1993). However, knowledge and rationality are important factors in functioning of the members of given community and the community or society as such. Looking at this situation

² The issue of trust in the results provided by AI systems based on the assumption that they are well-written programmes is discussed, among others, in: (Cappelen and Dever, 2021).

³ This raises the question of how to understand the reference to empirical reality in the case of such systems.

from the perspective of a wider group of people, one could venture to say that knowledge, especially in its traditional sense⁴, can be seen as an essential component of the common good of a given community. By common good, I mean the broadly understood well-being of a given community, achieved in as many areas of its functioning as possible, maintained and developed through its functioning, and also being the goal of that community's functioning in the longer term⁵.

In view of the above, the question arises as to whether it is possible to identify a tool that would help assess the assumptions, implementation, functioning and results (and their consequences) of AI systems, also in the context of gathering information and transforming it into knowledge at individual entities, precisely in the context of their impact on a given community. It seems that this is possible. My suggestion would be to use the criterion of the stability of a given community and its evolution in this context, especially when it comes to the (expected) growing impact of AI systems on the functioning of both individuals and entire communities. This is an analogy with the evolution of physical systems. As long as external disturbances are not too great, the evolution of such a system follows a fixed pattern that can be described by equations. However, disturbances may be so strong that they lead to the destruction of the system. The criterion of stability of functioning would aim to prevent such scenarios in relation to society⁶.

This criterion would apply to assessing the impact of a given technology on the functioning of individuals and communities. In addition, it seems necessary to emphasise deanthropomorphisation, i.e. the ability not to attribute human characteristics such as thinking, willing, understanding, etc. to digital systems, and anthropocentrism, i.e. the condition of adapting digital technologies to humans, to their way of functioning, and—if possible—strengthening their specifically qualities, creativity and taking care about ergonomics, rather than trying to fit human beings to the technical capabilities of digital

⁴ Bearing in mind that Plato's approach is a kind of 'patronal' approach.

⁵ I make no secret of my inspiration from the MacIntyre' concept of the common good (MacIntyre, 2013).

⁶ In respect to this, it maybe useful to employ an old-school virtue ethics to take care of certain ethical qualities of human being as such (cf. (MacIntyre, 2013)), highly independent on "the technological" fashion.

systems and their pace of producing results⁷. The question also arises as to whether it is possible to prepare individuals and communities in such a way that they are as resistant as possible to the potentially adverse effects of digital information processing technologies. In my opinion, one such measure could be to place emphasis on the development of epistemic virtues (Turri, Alfano and Greco, 2021), which combined with rich humanistic background⁸ can enforce the critical thinking as well as deep analytical thinking. These kinds of capacities usually had the chance to be developed and refined through academic education. However, the growing importance of information processing and widespread access to knowledge, as well as the dominance of the aforementioned technological paradigm, are increasingly undermining the role of higher education. This is a topic that deserves an entirely separate study.

CONCLUSIONS

Undoubtedly, generative AI systems can be very helpful in acquiring knowledge based on large sets of information. Today, it is difficult to imagine analysing the vast amounts of information needed, for example, to write a reasonably up-to-date scientific article without tools based on this type of technology. This is becoming increasingly obvious, especially in light of the growing pace of information production⁹. Tools based on AI systems have proven to be very effective in this area, including various types of diagnostics based on screening tests (cf. e.g. (Joseph et al., 2024; Zhou et al., 2025)). However, in light of the above considerations, it seems that general philosophical reflection on the processes taking place, their significance, broader context and possible consequences should not be neglected, as this may help to draw attention to important issues. The successes of AI systems may lead to a situation where we focus our attention mainly on what serves the development of technology, while neglecting critical thinking skills and a broad view of emerging problems. In the case of this study, there is a par-

⁷ Cf. analogous proposal: (Polak and Krzanowski, 2023).

⁸ By analogy, the humanistic background probably plays a role analogous to that of spiritual reserves in the context of the crisis in theoretical physics (Staruszkiewicz, 2001).

⁹ Currently (2025), it reaches 402.74 million TB per day (Duarte, 2025).

ticular link between the protection of rational thinking, morality and the functioning of society, in the context of the classical understanding of knowledge as true and justified beliefs. The ability and opportunity to ask such questions and to freely seek answers to them should, it seems, be given special protection. Analogous protection also seems appropriate for the ability to use AI systems critically and fruitfully for the functioning of society. The proposed criterion of the stability of the functioning and development of society under the influence of the applications of the results provided by AI systems may prove helpful here.

REFERENCES

- Ackoff, R.L. 1989. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- Alvarado, R., 2023. AI as an Epistemic Technology. *Science and Engineering Ethics*, 29(5), p. 32. <https://doi.org/10.1007/s11948-023-00451-3>
- Cappelen, H. Dever, J. 2021. *Making AI Intelligible: Philosophical Foundations*. 1st ed. [online] Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780192894724.001.0001>
- David, M. 2022. The Correspondence Theory of Truth, [in:] E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy*, Summer 2022. [online] Metaphysics Research Lab, Stanford University. Available at: <<https://plato.stanford.edu/archives/sum2022/entries/ruth-correspondence/>> [Accessed 12 July 2024].
- Duarte, F. 2025. *Amount of Data Created Daily (2025)*. [online] Exploding Topics. Available at: <<https://explodingtopics.com/blog/data-generated-per-day>> [Accessed 17 May 2025].
- Grobler, A. 1993. *Prawda i racjonalność naukowa*. Kraków: Inter Esse.
- Heller, M. 1993. *Moralność myślenia*. Tarnów: Wydawnictwo Diecezji Tarnowskiej Biblos.
- Heller, M. 2009. *Filozofia nauki: wprowadzenie*. Kraków: Petrus.
- Heller, M., 2015. *Nowa fizyka i nowa teologia*. Wyd. 4. Kraków: Copernicus Center Press.
- Heller, M., Budzik, S., Wszolek, S. (eds.) 1996. *Obrazy świata w teologii i w naukach przyrodniczych*. Universum. Tarnów: Wydawnictwo Diecezji Tarnowskiej Biblos.
- Heller, M. and Towarzystwo Autorów i Wydawców Prac Naukowych ‘Universitas’ 2008. *Ostateczne wyjaśnienia wszechświata*. Kraków: TAIWPN ‘Universitas’.
- Hollanek, T. and Nowaczyk-Basińska, K. 2024. Griefbots, Deadbots, Postmortem Avatars: on Responsible Applications of Generative AI in the Digital Afterlife Industry. *Philosophy & Technology*, 37(2), 63. <https://doi.org/10.1007/s13347-024-00744-w>

- Hołyst, J.A., Mayr, P., Thelwall, M., Frommholz, I., Havlin, S., Sela, A., Kenett, Y.N., Helic, D., Rehar, A., Maček, S.R., Kazienko, P., Kajdanowicz, T., Biecek, P., Szymanski, B.K. and Sienkiewicz, J. 2024. Protect our environment from information overload. *Nature Human Behaviour*, 8(3), 402–403. <https://doi.org/10.1038/s41562-024-01833-8>
- Ichikawa, J.J., Steup, M. 2024. The Analysis of Knowledge, [in:] E.N. Zalta and U. Nodelman, eds. *The Stanford Encyclopedia of Philosophy*, Fall 2024. [online] Metaphysics Research Lab, Stanford University. Available at: <<https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/>> [Accessed 13 May 2025].
- Ingarden, R. 2006. *Książeczka o człowieku*. Kraków: Wydawnictwo Literackie.
- Jonas, H. 1996. *Zasada odpowiedzialności: etyka dla cywilizacji technologicznej*. Translated by M. Klimowicz Kraków: Wydawnictwo Platan.
- Joseph, S., Selvaraj, J., Mani, I., Kumaragurupari, T., Shang, X., Mudgil, P., Ravilla, T., He, M. 2024. Diagnostic Accuracy of Artificial Intelligence-Based Automated Diabetic Retinopathy Screening in Real-World Settings: A Systematic Review and Meta-Analysis. *American Journal of Ophthalmology*, 263, 214–230. <https://doi.org/10.1016/j.ajo.2024.02.012>
- Kazienko, P., Cambria, E. 2024. Toward Responsible Recommender Systems. *IEEE Intelligent Systems*, 39(3), 5–12. <https://doi.org/10.1109/MIS.2024.3398190>
- Kouros, T., Papa, V. 2024. Digital Mirrors: AI Companions and the Self. *Societies*, 14(10), 200. <https://doi.org/10.3390/soc14100200>
- Laestadius, L., Bishop, A., Gonzalez, M., Illeňík, D., Campos-Castillo, C. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10), 5923–5941. <https://doi.org/10.1177/14614448221142007>
- Leciejewski, S. 2013. *Cyfrowa rewolucja w badaniach eksperymentalnych: studium metodologiczno-filozoficzne*. Seria Filozofia i Logika/Uniwersytet im. Adama Mickiewicza w Poznaniu. Poznań: Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza.
- MacIntyre, A.C. 2013. *After virtue: a study in moral theory*. Bloomsbury revelations. London: Bloomsbury Academic.
- Mściśławski, Ł. 2024. Is Generative AI Possible Cause of the Swan Song of the Rational Civilisation? *Studies in Logic, Grammar and Rhetoric*, 69(1), 441–455. <https://doi.org/10.2478/slgr-2024-0025>
- Polak, P. Krzanowski, R. 2023. How to Tame Artificial Intelligence? A Symbiotic AI Model for Beneficial AI. *Ethos. Kwartalnik Instytutu Jana Pawła II KUL*, 36(3), pp. 92–106.
- Simon, H.A. 1969. *The sciences of the artificial*. Cambridge, London: The MIT Press.
- Skolimowski, H. 1966. The Structure of Thinking in Technology. *Technology and Culture*, 7(3), 371–383. <https://doi.org/10.2307/3101935>

- Spence, E. 2021. *Stoic philosophy and the control problem of AI technology: caught in the web*. Values and identities. Lanham: Rowman & Littlefield Publishers.
- Staruszkiewicz, A. 2001. Współczesny stan fizyki teoretycznej poważnym zagrożeniem cywilizacyjnym. *Foton*, 74, 4–6.
- Steup, M., Neta, R. 2024. Epistemology, [in:] E.N. Zalta and U. Nodelman, eds., *The Stanford Encyclopedia of Philosophy*, Winter 2024. [online] Metaphysics Research Lab, Stanford University. Available at: <<https://plato.stanford.edu/archives/win2024/entries/epistemology/>> [Accessed 13 May 2025].
- Turri, J., Alfano, M. Greco, J. 2021. Virtue Epistemology, [in:] E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy*, Winter 2021. [online] Metaphysics Research Lab, Stanford University. Available at: <<https://plato.stanford.edu/archives/win2021/entries/epistemology-virtue/>> [Accessed 17 May 2025].
- Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., Lee, Y.-C. 2025. *The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships*. <https://doi.org/10.48550/arXiv.2410.20130>
- Zhou, J., Park, S., Dong, S., Tang, X., Wei, X. 2025. Artificial intelligence-driven transformative applications in disease diagnosis technology. *Medical Review* [online]. <https://doi.org/10.1515/mr-2024-0097>

WHAT DOES THE US FEDERAL TRADE COMMISSION (FTC). DONOTPAY ENFORCEMENT ACTION TEACH US ABOUT AI¹ ETHICS AND AI LEGAL ETHICS

STUART WEINSTEIN

“Law cannot be ignored, nor can it be confused with ethics. Both are parameters of social behaviour necessary in any field or context and, in particular, in areas of significant complexity such as AI”².

INTRODUCTION

With the cost of hiring a lawyer beyond the reach of most consumers, an app started in 2015 known as DoNotPay that uses a chatbot to help people produce documents and forms fancifully “promised that its service would allow consumers to ‘sue for assault without a lawyer’ and ‘generate perfectly valid legal documents in no time’, and that the company would ‘replace

¹ “AI is an umbrella term for a range of technologies and approaches that often attempt to mimic human thought to solve complex tasks.” Source: United Kingdom Information Commissioner’s Office (2025) Part 1 The basics of explaining AI – Definitions. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/definitions/#definitions_1 (accessed 8 November 2025).

² Robles Carrillo, M (2020) “Artificial intelligence: From ethics to law,” *Telecommunications Policy*, Elsevier, vol. 44(6). Available at: <https://www.sciencedirect.com/science/article/pii/S030859612030029X> (accessed January 2025).

the \$200-billion-dollar legal industry with AI”³. While it was hoped that DoNotPay would be a disruptive innovator that would harness AI technology to enable the delivery of legal services to the general public in such a way as to make legal services more affordable and accessible to a much larger population⁴. DoNotPay has been subject to litigation alleging that the company engaged in unfair business practices and the unauthorised practice of law⁵ and, moreover, that the company provided “substandard and poorly done services”⁶. Although DoNotPay’s founder Joshua Browder has since walked them back, he has been accused of overhyping his product by making a number of outlandish claims as to what it can do⁷. One of those frustrated by these claims who sued was Seattle-based paralegal, Kathryn Tewson. Tewson expected DoNotPay to offer “some kind of real-time legal analysis” but what she “found “was less ‘robot lawyer’ and more ‘plug-and-chug document wizard””⁸.

While such private litigation was noisome at best, when the FTC brought an enforcement action in September 2024 against DoNotPay (FTC DoNotPay Enforcement Action) making “allegations that the company made false

³ FTC (2024), Complaint, In the Matter of DONOTPAY, INC., a corporation, 232-3042, 20. [25 September]. Available at: https://www.ftc.gov/system/files/ftc_gov/pdf/DoNotPay-Inc-Complaint.pdf (accessed 7 January 2025).

⁴ Brescia RH, McCarthy W and McDonald A et al. Embracing Disruption: How Technological Change in the Delivery of Legal Services Can Improve Access to Justice (March 11, 2015). Albany Law Review, Vol. 78, 2014, Albany Law School Research Paper No. 6 for 2015-2016, pp. 553-621, 558. Available at: SSRN: <https://ssrn.com/abstract=2515009> (accessed 10 January 2025).

⁵ *MillerKing, LLC v. DoNotPay, Inc.*, 702 F. Supp. 3d 762 (S.D. Ill. 2023) held that a small law firm does not have standing to file a class action lawsuit on behalf of all US law firms against DoNotPay for the unauthorised practice of law. Available at: <https://fingfx.thomsonreuters.com/gfx/legaldocs/gdpzwlwbjvw/MillerKing%20v.%20DoNotPay%20Decision.pdf> (accessed 10 January 2025).

⁶ *Faridian v. DoNotPay, Inc.*, 23-cv-01692-RFL, (N.D. Cal. Feb. 1, 2024). The parties reached a settlement of the case as of 5 June 2024 and the case was dismissed with prejudice as of 20 July 2025. Available at: <https://www.courtlistener.com/docket/67158596/faridian-v-donotpay-inc/> (accessed 10 January 2025).

⁷ Reynolds M (2023) Inside the claims against DoNotPay’s Joshua Browder and the ‘World’s First Robot Lawyer’, ABA Journal. Available at: <https://www.abajournal.com/magazine/article/inside-the-claims-against-donotpays-joshua-browder-and-the-worlds-first-robot-lawyer> (accessed 13 January 2025).

⁸ Ibid.

and unsubstantiated claims that its AI-powered service could function like a human lawyer”⁹, the ethics of DoNotPay came front and centre. This paper will explore the ethical issues posed by the DoNotPay Enforcement Action which can be broken down into three categories: first, general ethics, second, AI ethics and third, AI legal ethics. There is no specific section focused on concept of general ethics but this is a topic is pervasive and covered throughout the paper. The structure of the remaining sections of this paper is as follows: Section II reviews the FTC DoNot Pay Enforcement Action. Section III sets out the research question and the author’s working hypothesis. Section IV shall define what we mean by AI ethics. Section V shall define what we mean by AI legal ethics. Section VI derives ethical operating principles going forward from the FTC DoNotPay Enforcement Action that can be learned from the DoNotPay shortcomings. Section VII draws some conclusions.

THE FTC DONOTPAY ENFORCEMENT ACTION

Emphasising the need to “eliminate deception from the market so that honest firms can compete to offer consumers innovative, trustworthy products,”¹⁰ the FTC took enforcement action against DoNotPay for claiming (among other things) that it was “the world’s first robot lawyer” and that its “service could substitute for the expertise of a human lawyer”¹¹. To settle these charges, DoNotPay agreed to pay a fine of \$193,000 and stop “making claims about its ability to substitute for any professional service without evidence to back it up”¹². The legal basis of the FTC complaint against DoNotPay (FTC DoNotPay Complaint)¹³

⁹ FTC, Concurring Statement of Commissioner Melissa Holyoak, Joined by Chair Lina M. Khan DoNotPay, Inc., FTC Matter No. 2323042, September 25, 2024. Available at: https://www.ftc.gov/system/files/ftc_gov/pdf/Holyoak-Khan-Statement-re-DoNotPay-09-25-2024.pdf (accessed 11 January 2025).

¹⁰ Ibid.

¹¹ US FTC, Cases and Proceedings: “DoNotPay,” 25 September 2024.

¹² Ibid.

¹³ FTC (2024), Complaint, In the Matter of DONOTPAY, INC., a corporation, 232-3042, 20. [25 September]. Available at: https://www.ftc.gov/system/files/ftc_gov/pdf/DoNotPayInc-Complaint.pdf (accessed 7 January 2025).

was that the DoNotPay acts and practices “constitute unfair or deceptive acts or practices in or affecting commerce in violation of Section 5(a) of the FTC Act”¹⁴.

RESEARCH QUESTION

While it is clear that DoNotPay violated the “principles, norms and laws that govern ethics and law”¹⁵, does its wrongful conduct offer us insight into concepts of AI ethics, on the one hand, and AI legal ethics, on the other hand? Or was DoNotPay engaged in a case of simple deceit which while unethical it is ultimately unedifying. It is the author’s working hypothesis that an examination of the concepts of AI ethics and AI legal ethics when applied to the facts of the FTC DoNotPay Enforcement Action case will help us better identify appropriate principles from AI ethics and AI legal ethics that may be instructive in other cases. The author’s confidence in this regard is bolstered by McLaren (2003) who developed a computational model for retrieving abstract principles and past cases involved with ethics so as to operationalise them¹⁶ and the work of and Morley, Elhalal, Garcia, et al. (2021) who have developed an ethics as a service framework for the pragmatic operationalisation of AI ethics in AI development projects¹⁷.

DEFINING AI ETHICS

AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the deve-

¹⁴ Section 5(a) of the Federal Trade Commission Act (FTC Act) (15 USC 45).

¹⁵ McLaren, B.M. (2003) Extensionally defining principles and cases in ethics: An AI model, *Artificial Intelligence*, Volume 150, Issues 1–2, 2003, Pages 145–181, 146. Available at: <https://www.sciencedirect.com/science/article/pii/S0004370203001358> (accessed 10 January 2025).

¹⁶ McLaren, B.M. (2003) Extensionally defining principles and cases in ethics: An AI model, *Artificial Intelligence*, Volume 150, Issues 1–2, 2003, Pages 145–181, 146. Available at: <https://www.sciencedirect.com/science/article/pii/S0004370203001358> (accessed 10 January 2025).

¹⁷ Morley, J., Elhalal, A., Garcia, F. *et al.* Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds & Machines* **31**, 239–256 (2021). <https://doi.org/10.1007/s11023-021-09563-w> (accessed 12 January 2025).

lopment and use of AI technologies¹⁸. This field has emerged “as a response to the range of individual and societal harms that the misuse, abuse, poor design, or negative unintended consequences of AI systems may cause”¹⁹. Floridi (2024) speaks of the “BIBO problem: bias in, bias out” in that AI systems “trained on vast datasets, often reflect and amplify societal biases inherent in their training data” leading to discriminatory outcomes²⁰. He also notes that “AI systems amass, analyse, and manipulate immense amounts of personal information, questions emerge about individual rights and safeguarding sensitive data” in the context of AI governance²¹.

Balasubramaniam and Kauppinen et al. (2023) write that “various interest groups across the world (e.g., IEEE, ACM) have defined comprehensive ethical guidelines and principles to ensure responsible AI usage emphasising transparency and explainability for developing AI systems”²². For instance, do the quality of explanations an AI system requires necessitate inserting “humans in the loop” to improve an AI systems need for explanation?²³. It is essential to “ensure accountability for decisions and actions within the complex infrastructural arrangements that characterise AI technologies” where the presence of “black box” decision-making may occur where opaque processes exist that transform data inputs to insight outputs²⁴. Similar to the

¹⁸ Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*, pp 1 – 97, 3. The Alan Turing Institute. <https://doi.org/10.5281/zenodo.3240529> (accessed 6 January 2025).

¹⁹ Ibid.

²⁰ Floridi, Luciano, *The Ethics of Artificial Intelligence: exacerbated problems, renewed problems, unprecedented problems* - Introduction to the Special Issue of the American Philosophical Quarterly dedicated to The Ethics of AI (April 20, 2024). Available at SSRN: <https://ssrn.com/abstract=4801799> or <http://dx.doi.org/10.2139/ssrn.4801799> (accessed 12 January 2025).

²¹ Ibid.

²² Balasubramaniam N, Kauppinen M, Rannisto A, Hiekkanen K, Kujala S, *Transparency and explainability of AI systems: From ethical guidelines to requirements*, Information and Software Technology, Volume 159, 2023, 107197, ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584923000514> (accessed 12 January 2025).

²³ Stathis, G., van den Herik, J. (2024) *Ethical and preventive legal technology*, AI Ethics, Sec. 5.2. Available at: <https://doi.org/10.1007/s43681-023-00413-2> (accessed 7 January 2025).

²⁴ Kempton, AM and Vassilakopoulou, P (2021) *Accountability, Transparency & Explainability in AI for Healthcare*. 8th International Conference on Infrastructures in Healthcare. Available at: <https://dl.eusset.eu/server/api/core/bitstreams/07376024-ac3c-4c1b-9f21-363c8cfe5965/content> (accessed 12 January 2025).

“black box” problem are the technical shortcomings that characterise the use of generative AI LLM chatbots which may unwittingly produce output that may contain “falsehoods, biases, incorrect cases or other AI hallucinations”²⁵. Moving beyond AI ethics we need to look at the additional overlay that the regulation of the unauthorised practise of law and the professional conduct requirements of the legal professions add to the issues at stake. We shall do this in the next section.

AI LEGAL ETHICS

AI legal ethics focuses on the ethical duties of lawyers found in professional codes of conduct and the regulatory mandate of various respective law societies, state bars or other legal regulatory organisations that are designed to prevent the unauthorised practice of law, breach of client confidentiality, etc. with respect to the use of AI to deliver legal services or to meet legal needs²⁶. The general concept of legal ethics, however, is subject to debate. Traczykowski argues that legal ethics must move beyond focusing exclusively on professional conduct rules and adapt an applied ethicist approach that incorporates a “universal set of moral principles or concepts, separate to religion, that can be reasoned through (and hence subjected to scrutiny), but which can be generally utilised across situations, circumstances and jurisdictions”²⁷. No matter how legal ethics is defined, the problem here is that the professional conduct rules that govern lawyers do not address a lacuna in the law that exists when non-lawyers use AI legal apps on their own without

²⁵ New York State Bar Association [NYSBA] (2024). Report and Recommendations of the NYSBA Task Force on Artificial Intelligence [4 April], 45. <https://nysba.org/app/uploads/2022/03/2024-April-Report-and-Recommendations-of-the-Task-Force-on-Artificial-Intelligence.pdf>.

²⁶ Salyzyn, Amy, AI and Legal Ethics (February 1, 2021). in Florian Martin-Bariteau & Teresa Scassa, eds., *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021), pp. 1-23, 1. Available at SSRN: <https://ssrn.com/abstract=3733952> (accessed 6 January 2025)

²⁷ Traczykowski, L. (2024). Effective teaching of legal ethics: use an applied ethicist. *Legal Ethics*, pp. 1–20, 5. Available at: <https://doi.org/10.1080/1460728x.2024.2399919> (accessed 6 January 2025).

the aid of lawyers to assist themselves in resolving legal risk problems. This create challenges for regulators in that there is a need “to determine the right balance between facilitating access to justice and protecting the public from potential harms arising from the use of legal AI tools”²⁸.

DERIVING ETHICAL OPERATING PRINCIPLES

DoNotPay’s primary failing was that “it did not conduct testing to determine whether its AI chatbot’s output was equal to the level of a human lawyer, and that the company itself did not hire or retain any attorneys”²⁹. It held itself out as capable of providing AI-generated legal advice yet what was offered was clearly not so. In light of the fact that DoNotPay’s technology never was a true AI legal advice tool³⁰ it is hard to see this case as an example of failed AI ethical conduct. It is more likely a false claims case combined with unauthorised practise of law. The State Bar of California issued a “cease and desist” order that DoNotPay stop offering its services to California residents and that it not describe itself as the “World’s First Robot Lawyer”³¹. DoNotPay failed to comply with this order³². In terms of the unauthorised practice of law, one could argue that DoNotPay violated the unauthorised practice of law restrictions in California not because of its

²⁸ Salyzyn, Op. cit.

²⁹ FTC, Press Release, FTC Announces Crackdown on Deceptive AI Claims and Schemes With Operation AI Comply, Agency announces five law enforcement actions against operations that use AI hype or sell AI technology that can be used in deceptive and unfair ways, September 25, 2024. Available at: <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes> (accessed 12 January 2025).

³⁰ An “AI legal advice tool” can be defined as an AI-based tool by which a process or procedure is performed without human assistance whose major purpose is giving legal advice as regulated by the legal profession. See Bennett J, Miller T, Webb J, Bosua R, Loddors A and Chamberlain S (2018) Current State of Automated Legal Advice Tools - Discussion Paper 1 (Melbourne: The University of Melbourne, Networked Society Institute 2018), 9. Available at: <https://apo.org.au/node/143431> (accessed 12 January 2025).

³¹ FTC Complaint (2024) Op Cit., paragraphs 25-27, State Bar of California’s Investigation of DoNotPay for the Unauthorized Practice of Law.

³² Ibid.

AI capability but rather in spite of it. The views of the FTC in the DoNotPay Complaint³³ offers a workable solution for the development of an AI driven legal advice tool that could be of use to consumers in answering basic legal queries. Greater testing of the AI legal advice tool to see if it really operated like a human lawyer³⁴ which could be done through training the system “on a comprehensive and current corpus of federal and state laws, regulations, and judicial decisions or on the application of those laws to fact patterns.” If this were done and the quality and accuracy of the legal documents and advice generated was signed off by competent lawyers (humans in the loop – verifying the quality and accuracy of the advice given) then DoNotPay would have a genuine AI legal advice tool.

CONCLUSIONS

Coming back to our working hypothesis that a review of the DoNotPay case could help us operationalise ethical principles that could apply to both the AI ethics and the AI legal ethics elements when addressing the development and running of AI driven legal advice tool, it soon becomes clear that expert legal advisors must be involved in the product development process, its implementation and in the successful running of the tool as it deals with the issues of consumers. What sets an AI legal advice tool apart from those of other AI driven products or services is that to do it right you have to follow AI ethics development and operational principles as well as “sector-specific” professional ethics requirements belonging to the field of AI legal ethics. While this may be burdensome, it is not unsurprising. We would expect an AI driven medical advice tool to be developed involving medical professionals and oversight being provided by the same. If done right, this could be an unbelievably valuable tool. Unfortunately, however, in the DoNotPay case, the company did not progress to the point where DoNotPay could really be considered to be a genuine AI legal advice tool. In this regard, it was a shame as there was so much hype regarding DoNotPay being the “World’s First Robot Lawyer” those of us observing developments in this

³³ FTC (2024), Complaint, In the Matter of DONOTPAY, INC., a corporation, 232-3042, 20. [25 September]. Available at: https://www.ftc.gov/system/files/ftc_gov/pdf/DoNotPay-Inc-Complaint.pdf (accessed 7 January 2025).

³⁴ Ibid.

areas wished it could have lived up to its hype as there is a clear need for such an app to meet an underserved market – consumers who need legal advice but cannot afford to hire a lawyer.

REFERENCES

- Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkänen, K., Kujala, S. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, Volume 159, 2023, 107197, ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584923000514> (accessed 12 January 2025).
- Bennett, J., Miller, T., Webb, J., Bosua, R., Lidders, A., Chamberlain S (2018). *Current State of Automated Legal Advice Tools – Discussion Paper 1* (Melbourne: The University of Melbourne, Networked Society Institute 2018), 9. Available at: <https://apo.org.au/node/143431> (accessed 12 January 2025).
- Brescia, R.H., McCarthy, W., McDonald, A. et al. (2014). Embracing Disruption: How Technological Change in the Delivery of Legal Services Can Improve Access to Justice (March 11, 2015). *Albany Law Review*, Vol. 78, 2014, Albany Law School Research Paper No. 6 for 2015–2016, 553–621, 558. Available at: SSRN: <https://ssrn.com/abstract=2515009> (accessed 10 January 2025).
- Faridian v. DoNotPay, Inc.*, 23-cv-01692-RFL, (N.D. Cal. Feb. 1, 2024). The parties reached a settlement of the case as of 5 June 2024 and the case was dismissed with prejudice as of 20 July 2025. Available at: <https://www.courtlistener.com/docket/67158596/faridian-v-donotpay-inc/> (accessed 10 January 2025).
- FTC (2024), Cases and Proceedings: “DoNotPay,” 25 September 2024. Available at: <https://www.ftc.gov/legal-library/browse/cases-proceedings/donotpay> (accessed 6 January 2025).
- FTC (2024), Complaint, In the Matter of DONOTPAY, INC., a corporation, 232-3042, 20. [25 September]. Available at: https://www.ftc.gov/system/files/ftc_gov/pdf/DoNotPayInc-Complaint.pdf (accessed 7 January 2025).
- FTC, Concurring Statement of Commissioner Melissa Holyoak, Joined by Chair Lina M. Khan DoNotPay, Inc., FTC Matter No. 2323042, September 25, 2024. Available at: https://www.ftc.gov/system/files/ftc_gov/pdf/Holyoak-Khan-Statement-re-DoNotPay-09-25-2024.pdf (accessed 11 January 2025).
- FTC (2024), Press Release, FTC Announces Crackdown on Deceptive AI Claims and Schemes With Operation AI Comply, Agency announces five law enforcement actions against operations that use AI hype or sell AI technology that can be used in deceptive and unfair ways, September 25, 2024. Available at: <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes> (accessed 12 January 2025).

Federal Trade Commission Act, Section 5(a) (15 USC 45).

Floridi, L., The Ethics of Artificial Intelligence: exacerbated problems, renewed problems, unprecedented problems – *Introduction to the Special Issue of the American Philosophical Quarterly dedicated to The Ethics of AI* (April 20, 2024). Available at SSRN: <https://ssrn.com/abstract=4801799> or <http://dx.doi.org/10.2139/ssrn.4801799> (accessed 12 January 2025).

Kempton, A.M., Vassilakopoulou, P (2021). Accountability, Transparency & Explainability in AI for Healthcare. 8th International Conference on Infrastructures in Healthcare. Available at: <https://dl.eusset.eu/server/api/core/bitstreams/07376024-ac3c-4c1b-9f21-363c8cfe5965/content> (accessed 12 January 2025).

Khan, S., Powers, E. (2023). *Efficiency, Ethics, and Algorithms: The Implications of AI on the Legal Profession and the ABA Model Rules* (May 28), 1–31, 25. Available at SSRN: <https://ssrn.com/abstract=4461276> or <http://dx.doi.org/10.2139/ssrn.4461276>

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*, 1–97, 3. The Alan Turing Institute. <https://doi.org/10.5281/zenodo.3240529> (accessed 6 January 2025).

McLaren, B.M. (2003). Extensionally defining principles and cases in ethics: An AI model, *Artificial Intelligence*, Volume 150, Issues 1–2, 2003, 145–181, 146. Available at: <https://www.sciencedirect.com/science/article/pii/S0004370203001358> (accessed 10 January 2025).

MillerKing, LLC v. DoNotPay, Inc., 702 F. Supp. 3d 762 (S.D. Ill. 2023) held that a small law firm does not have standing to file a class action lawsuit on behalf of all US law firms against DoNotPay for the unauthorised practice of law. Available at: <https://fingfx.thomsonreuters.com/gfx/legaldocs/gdpzwlwbjvw/MillerKing%20v.%20DoNotPay%20Decision.pdf> (accessed 10 January 2025).

Morley, J., Elhalal, A., Garcia, F. et al. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds & Machines* 31, 239–256 (2021). <https://doi.org/10.1007/s11023-021-09563-w> (accessed 12 January 2025).

Reynolds, M. (2023). Inside the claims against DoNotPay’s Joshua Browder and the ‘World’s First Robot Lawyer’, *ABA Journal*. Available at: <https://www.abajournal.com/magazine/article/inside-the-claims-against-donotpays-joshua-browder-and-the-worlds-first-robot-lawyer> (accessed 13 January 2025).

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, Elsevier, vol. 44(6). Available at: <https://www.sciencedirect.com/science/article/pii/S030859612030029X> (accessed January 2025).

Stathis, G., van den Herik, J. (2024). Ethical and preventive legal technology, AI Ethics, Sec. 5.2. Available at: <https://doi.org/10.1007/s43681-023-00413-2> (accessed 7 January 2025).

- Salyzyn, Amy, AI and Legal Ethics (February 1, 2021). in Florian Martin-Bariteau & Teresa Scassa, eds., *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021), pp. 1-23, 1. Available at SSRN: <https://ssrn.com/abstract=3733952> (accessed 6 January 2025).
- Traczykowski, L. (2024). Effective teaching of legal ethics: use an applied ethicist. *Legal Ethics*, 1–20, 5. Available at: <https://doi.org/10.1080/1460728x.2024.2399919> (accessed 6 January 2025).
- UK Information Commissioner's Office (2025). Part 1 The basics of explaining AI – Definitions. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/definitions/#definitions_1 (accessed 8 November 2025).
- University of Pittsburgh, University Center for Teaching and Learning, (last updated) 29 August 2024. Available at: <https://teaching.pitt.edu/resources/what-is-generative-ai/> (accessed 7 January 2025).

ARTIFICIAL INTELLIGENCE, REAL REGULATIONS – CAN THE TECHNOLOGICAL REVOLUTION BE TAMED?

MAREK PORZEŻYŃSKI

The technological revolution is a fact. Some of the most dynamic transformations in human history, driven by technological advancements, are occurring before our very eyes and permeating all sectors of the economy. This statement serves as the opening for nearly every article I write, even in its preliminary form, and as the starting point for most of my presentations, regardless of whether they pertain to issues of cybersecurity, privacy, biometric data, or intellectual property. The frequent repetition of this assertion does not, however, diminish its veracity or relevance. Just as it was over a decade ago when I authored my first advanced research articles, it remains unequivocally true today. Moreover, the pace of technological development continues to accelerate, with no indication of significant deceleration in the near future.

Even in the most “traditional” sectors of the economy, which have hitherto resisted technological changes, maintaining existing solutions without leveraging the opportunities offered by digitalization and/or process automation is increasingly untenable. Even in the traditional academic (educational) environment, these trends are evident, with top-ranked universities and research institutions offering a broad array of solutions made possible exclusively through new technologies. These include courses, and even

entire degree programs available remotely¹, as well as conferences conducted entirely within virtual worlds².

All of these developments paint a picture of a modern economy powered by vast quantities of processed data, artificial intelligence technologies fueled by such data³, and broadly understood automation and robotics. These changes affect, in principle, every branch of the economy, irrespective of its status or level of funding. As interest in a given technology grows, so too does its accessibility. Thus, at the time of writing, access to large language models is even free of charge, whereas only a few months ago, one of the industry leaders was raising the pricing plans for its subscriptions. With the increasing availability of specific tools, they transition into the mainstream and are increasingly employed by various industries and individuals performing tasks within these domains.

Given the rapid pace of change and its impact – at times disruptive, as in the case of social media platforms⁴ – on human lives, regulatory frameworks must be designed to ensure transactional security while simultaneously mitigating the negative effects of their application. Both of these characteristics must be assured concurrently, although achieving this balance in practice is exceedingly difficult, as evidenced by successive regulatory attempts in the field of broadly understood new technologies. Striking a balance between these potentially conflicting priorities is therefore essential. Excessive regulation discourages entrepreneurs from entering the market or leads to the expansion of the so-called “gray area” which is, of course, an undesirable outcome. Conversely, an absence of regulation or excessively broad regula-

¹ The most recognised academic centres from the USA seem to lead the way in this regard by making available a wide range of courses and degree programmes.

² One of the largest events of its kind in Poland is the Virtualium conference, in which the author took an active part.

³ Without input, it would not be possible to have a process of ‘training’ or ‘learning’, resulting in the largest language models as well as other examples of AI-based technology becoming better at performing specific tasks.

⁴ For example: Hańczuk Maja, Rybołowicz Gabriela, Szwed Julia, Wilczyńska Julia, Olszewska Anna Małgorzata: Wpływ mediów społecznościowych na relacje międzyludzkie, Akademia Zarządzania, Faculty of Management, Białystok University of Technology, vol. 8, no. 2, 2024, pp. 292-313, DOI:10.24427/az-2024-0026; William J. Brady et al, How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* 7, eabe5641(2021). DOI:10.1126/sciadv.abe5641

tory frameworks significantly increases the risk of negative consequences. There are numerous historical examples of the regulation of “new technologies,” with initial regulatory efforts frequently yielding unintended, unforeseen consequences. Looking far into the past, the advent of motorization and its growing accessibility introduced challenges related to road traffic and the need to ensure the safety of various road users, including pedestrians, horseback riders, carriages, and the increasingly dominant predecessors of modern automobiles. The first regulations stipulated that a person should walk ahead of such a vehicle, warning others of the approaching car. This requirement not only increased the already high costs of automobiles due to the necessity of employing an additional worker but also paradoxically contributed to increasing road congestion by introducing additional participants. Moving closer to modern times, in the early days of the internet, identity verification by postal mail was required to register account in various online portals, a measure that effectively hindered user registration. This soon gave way to the opposite approach, where no identity verification was required, leading to significantly negative consequences, such as a lack of oversight and corresponding safety concerns. Consequently, despite the introduction of additional security measures, the myth that the internet is an anonymous space where anything is permissible has taken root and persists in certain circles to this day. Attempts to introduce further security measures or content moderation in social media spaces, for example, are frequently perceived as an attack on free speech and efforts to impose censorship, despite this not being the case in the majority of instances.

Enhancing transactional security typically correlates with a reduction in decision-making autonomy and, to some extent, the limitation of certain freedoms. Neither extreme can fully satisfy societal expectations regarding the proposed regulations. The legislative process must carefully balance additional obligations imposed on stakeholders with their impact on other factors that enable the utilization of a given technology or solution. Even the most well-designed legal act may ultimately prove misaligned with market expectations or prevailing industry practices. Furthermore, regulations that are well-suited to one phase of a given market’s development may become entirely inadequate in another, as exemplified by the case of automobiles. Technology regulations are thus an exceptionally specialized field requiring the integration of seemingly disparate competencies from technology,

law, and legislation. Additional key factors include the speed of regulatory responses to technological changes and the resilience of regulations to such changes. Accordingly, I argue that the more critical question is not what to regulate, but whether and how to regulate once the necessity of regulation has been established. In the initial stages of a technology's rapid growth, regulation is often unnecessary and may even be detrimental. Moreover, even if such regulation does not prove harmful, it may necessitate frequent amendments, resulting in the continuous formulation of new legal frameworks, which ultimately may not foster positive developmental outcomes.

In my research, I have examined instances where legal intervention has already occurred within industries classified under the broad category of new technologies, analyzing cases in which regulations have either succeeded or yielded outcomes contrary to their original intent, thereby slowing or halting technological advancement. Furthermore, these findings have been subjected to additional analyses aimed at identifying factors that might contribute to or facilitate regulatory success. Unfortunately, this second area of research cannot be deemed conclusive due to two key limitations: first, it is characterized by a significant degree of subjective evaluation, and second, it involves comparing inherently incomparable elements – entirely distinct industries. Consequently, these secondary findings may serve only as guidelines rather than definitive indicators of legislative success.

A prime example of a regulation with negative consequences for the technological sector is the 2018⁵ law intended to facilitate research and, in particular, the testing of autonomous vehicles on public roads. Despite the market's anticipation of this regulatory measure, it effectively blocked autonomous vehicle testing on public roads in Poland⁶. This statement is by no means an exaggeration. My research, conducted extensively, followed the evolution of this legislation through 2024. In my initial broad study in 2019, I demonstrated that within a year of the law's enactment, no autonomous vehicle tests had taken place on public roads in Poland. Furthermore, there was not a single inquiry or request submitted regarding the possibility

⁵ Act of 11 January 2018 r. on electromobility and alternative fuels, Dz.U. z 2018 r., poz. 317

⁶ Further on this subject: M. Porzeżyński, *O potrzebie regulacji w zakresie wykorzystania zdobyczy nowych technologii na przykładzie pojazdów autonomicznych* [in:] D. Bach-Golecka, A. Zawidzka Łojek (red.), *In varietate concordia. Wyzwania prawa europejskiego*, Warsaw 2020.

of conducting such tests, which indicates the extremely poor legal framework of this regulation, leading to a complete lack of industry interest. An updated study conducted in the second half of 2024 revealed that during the six years of these regulations' applicability in Poland, autonomous vehicle testing had occurred only once under these provisions. It should be noted, however, that these tests were carried out by a public research institute⁷, which most likely meant that the tests could only take place in Poland. Private companies that do not benefit from full state funding can afford to choose a location that gives them more opportunities or where it is easier to carry out a given research or development activity. In the case of fully state-funded entities, there is usually no such possibility, and even when there is, it is in practice impossible to opt for it, if only for image reasons. This is therefore a perfect example of bad regulation, the framework of which was created in such a way as to make it as difficult as possible to obtain a test authorisation itself, as well as to introduce requirements at a later stage that also made it commercially pointless. In this case, a number of lengthy procedures involving multiservice authorisations had to be followed. Moreover, any objection from a person who would live along the road where the tests were to be carried out automatically made it impossible. Finally, the test data had to be shared with the designated state institutions.

On the other hand, an example of a successful regulation benefiting society at large is the General Data Protection Regulation (GDPR) – Regulation 2016/679 of April 27, 2016⁸. This regulation stands out in the realm of technology-related legislation due to its resilience to technological changes and its genuine, positive impact on individual freedoms, particularly in the realm of privacy protection. Moreover, the territorial scope of this regulation was designed in such a way that any company that wished to operate within the European Union, or involved personal data of citizens of a Member State, had to comply with the requirements of this regulation. However, it is not a perfect act and is not as clear an example of positive regulation as the clear example of negative regulation was in the act discussed above. In particular, the RODO has been accused of being overly restrictive, complex and unsuitable for smaller businesses. This is all true and will certainly

⁷ <https://pimot.lukasiewicz.gov.pl/2023/12/04/pierwsze-w-polsce-testy-pojazdu-autonomicznego-na-drogach-publicznych/>

⁸ <https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=celex:32016R0679>

be subject to analysis and change at the earliest opportunity. Unfortunately, the restrictiveness and complexity seem in this case to be justified by the matter of the regulation and fundamental changes in this respect should not be anticipated. It seems, however, that it is entirely feasible to introduce simplifications, exceptions or even exemptions from certain requirements for micro and small entrepreneurs or those who access and process data only in the course of their activities which are not directly based on data processing. It also appears that some exceptions may be necessary for industries that rely entirely on the processing of different categories of data, which may also include personal data, but not for their use for profiling of individuals or even social groups. Such exceptions could take different shapes and frameworks, including timeframes, and would certainly allow for more dynamic development in the field of artificial intelligence technology in the European Union in particular.

The regulations discussed above were chosen subjectively as examples of both positive and negative regulatory impacts on technological development. However, the question of whether a lack of dedicated regulation in certain technological areas leads to beneficial or adverse effects is an entirely separate issue that requires further detailed investigation. Nonetheless, these examples demonstrate that it is not self-evident whether new technologies should be regulated, nor how to craft such regulations should the need for them arise. Everything undoubtedly requires detailed ad hoc analysis and adaptation to societal needs, as well as the requirements and - in the end - the expectations of the industry to be affected by the regulation. The order of the elements to be considered is not accidental. The example of data protection regulation shows emphatically that societal needs in the form of securing individual rights are often at odds with the expectations of different industries. This does not mean, however, that such expectations should be disregarded, as this may lead to a limitation of technological development or even, the blocking of given services or products for a given territory. The regulation of new technologies is therefore an extremely sensitive element of the broader field of regulation, which should be placed in the hands of specialists in the field of technology regulation and not just lawyers-legislators, scientists, engineers or lobbyists.

REFERENCES

- Hańczuk, M., Rybołowicz, G., Szwed, J., Wilczyńska, J., Olszewska, A.M.: Wpływ mediów społecznościowych na relacje międzyludzkie, *Akademia Zarządzania, Faculty of Management, Białystok University of Technology*, vol. 8, no. 2, 2024, 292–313, DOI: 10.24427/az-2024-0026
- W.J. Brady et al. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* 7,eabe5641(2021). DOI: 10.1126/sciadv.abe5641
- Hawkins A.J., US will rewrite safety rules to permit fully driverless cars on public roads, the Verge, <https://www.theverge.com/2018/10/4/17936576/self-driving-car-av-guidelines-3-nhtsa-elaine-chao>
- Jaworski B., Autonomous vehicles: the legal landscape of using and testing autonomous cars in Poland, [w:] Cendrowicz D., Chrisdu-Budnik A., *Comparative Perspectives for Public Administration and Administrative Law*, 2018.
- Pierwsze w Polsce testy pojazdu autonomicznego na drogach publicznych, <https://pimot.lukasiewicz.gov.pl/2023/12/04/pierwsze-w-polsce-testy-pojazdu-autonomicznego-na-drogach-publicznych/>
- Porzeżyński M., O potrzebie regulacji w zakresie wykorzystania zdobyczy nowych technologii na przykładzie pojazdów autonomicznych, [w:] D. Bach-Golecka, A. Zawidzka Łojek (red.), *In varietate concordia. Wyzwania prawa europejskiego*, Warsaw 2020.
- Szymczak M., W oczekiwaniu na autonomiczne samochody. Czy spełnią oczekiwania kierowców i jak wpłyną na miasta? *Transport Miejski i Regionalny*, 2013.

Legal acts:

- Act of 11 January 2018 r. on electromobility and alternative fuels, Dz.U. z 2018 r., poz. 317
- Justification to the government project of act on electromobility and alternative fuels, form no. 2147
- Regulation no. 2016/679 of April 27, 2016 - General Data Protection Regulation (GDPR)

THE ENVIRONMENTAL COSTS OF AI DEVELOPMENT AS ETHICAL COSTS

RADOSŁAW SIEDLIŃSKI

INTRODUCTION

Artificial intelligence (AI) has become a defining feature of the 21st century. While debates on AI have often concentrated on its technical capacities and associated social risks – automation of work, unemployment, erosion of privacy and trust, or disinformation – these perspectives remain incomplete. When discussing ethical issues related to the development of AI, the questions of the possibility of artificial consciousness, the legal subjectivity of self-aware software or problems related to the so-called “pulling the plug” are often raised. All of these issues have long been present in both academic discussions and science fiction literature. AI, however, is not only an algorithmic or virtual entity. Nor is it merely a global software network transforming our understanding of privacy and irreversibly changing the labour market. At its most elementary, physical core it is rooted in material infrastructures that consume vast amounts of energy, water, and raw materials.

This article argues that AI should be analyzed within two interconnected frameworks: standard one (negative social consequences, threats to the privacy of users or to the security of states, possible emergence of self-aware software etc.) and environmental one (carbon / water footprint, global extraction of minerals, devastation of natural habitats). Not only the first of these frameworks can be considered from an ethical perspective. The second one

also calls for it. As a basis for such an approach, I propose the imperative of responsibility formulated by Hans Jonas (Jonas 1984). By framing AI within Jonas's anticipatory ethics and Labban's elaboration of the planetary mine, the article demonstrates that responsible AI development must integrate ecological boundaries alongside anthropocentric concerns.

STANDARD VIEW ON RISKS AND ISSUES ASSOCIATED WITH AI

Debates on AI often begin with well-established social and legal issues. Automation and robotization reduce employment opportunities, particularly in routine jobs, thereby aggravating social inequality and creating risks of political tension.

Privacy is another area of concern. In the era of surveillance capitalism, personal data has become a primary resource that can be used to profile and influence users. Its large-scale collection and commercialisation by global corporations undermines individual autonomy and transform users into commodities (Wang & Kosinski 2018; Billiris, Gill & Bandara 2025; Zuboff 2019).

The proliferation of synthetic content, such as deepfakes, poses further ethical challenges. By destabilizing trust in testimony and representation, deepfakes undermine the epistemic foundations of social interactions. In response, the European Union has introduced the AI Act, aimed at ensuring greater transparency in AI-generated content. (European Union 2024)

Finally, there are profound security concerns. The unpredictability of advanced AI systems raises questions about humanity's ability to maintain control. The "pulling the plug" dilemma highlights the risks of entrusting essential infrastructures to autonomous systems that might exceed human oversight (Musser et al. 2023; Humphreys 2024).

ENVIRONMENTAL VIEW ON AI DEVELOPMENT

AI depends not only on constant inflow of data but also on using vast amounts of energy, water, and raw materials. These dependencies result in three major ecological costs.

First, there is the carbon footprint. The electricity needed to power data centres does not come from nowhere. Even today, its primary source on a global scale remains the burning of fossil fuels. By 2026, data centers may consume about 1000 TWh of electricity – comparable to Japan’s entire annual demand. This estimate comes from the official report published by the International Energy Agency, which warns that the energy consumption of data centers could double within a few years (IEA 2024, p. 17). The mass production of power hungry GPUs designed for artificial intelligence and high-performance computing such as highly successful Nvidia’s H100 further amplifies this demand, while the construction of new data centers and the manufacturing of semiconductors generate significant embedded emissions. *Microsoft’s 2024 Environmental Sustainability Report* attributes a major increase in its own carbon footprint to “the construction of new data centers and the embodied carbon in building materials and hardware components, including semiconductors and servers” (Microsoft 2024, p. 42).

Second, there is the water footprint. Computers operating in data centres generate enormous amounts of heat, which must be dissipated. Water is the most popular coolant used in these centres. That is the reason why data centers requires enormous amounts of water. In 2022, Google, Microsoft, and Meta together withdrew approximately 2.2 billion cubic meters of water which is equivalent to Denmark’s water consumption for two years. The 2024 report by China Water Risk notes that Chinese data centers alone used about 1.3 billion cubic meters of water in 2023, enough to meet the domestic needs of 26 million people for an entire year (China Water Risk 2024, p. 6). It also predicts that China’s data center water consumption will double by 2030, compounding climate-driven water scarcity across Asia (China Water Risk 2024, p. 12). Microsoft likewise reported a 34% year-on-year increase in its global water usage in 2022, primarily due to the expansion of its AI and cloud infrastructure (Microsoft 2024, p. 47).

Third, there is natural habitats destruction. The mining of tin and rare earth elements necessary for the manufacture of batteries and electronic equipment, such as in Bayan Obo Mining District (China) or Bangka Belitung Islands (Indonesia), devastates ecosystems, reduces biodiversity, and contaminates local water sources. AI thus exemplifies what can be called extreme extractivism: the subordination of nature to technological and economic imperatives (Arboleda 2020, p. 54; Labban 2014, pp. 561–563).

In this context, it is useful to refer to Mazen Labban's concept of the planetary mine, which provides a broader theoretical framework for understanding the material and ecological dimensions of digital technologies. Labban argues that global capitalism has reconfigured mining into a planetary process that extends far beyond localized sites of extraction. As he writes, "the planetary mine does not consist of a single place or territory but of a network of operations and flows that deterritorialize extraction and integrate it into the global circuits of capital accumulation" (Labban 2014, p. 561).

In Labban's view, the planetary mine represents the totality of socio-ecological relations through which the Earth's biological and geological matter is transformed into commodities, energy, and information. The production of data – and by extension, the functioning of AI – is inseparable from these processes. AI thus becomes both a product and a driving force of the planetary mine: an ever-growing element of a system that literally transforms planetary matter into a stream of digital data and value.

The concept of the planetary mine was further developed by Martin Arboleda, who points out that even minor practices in our everyday lives embodies the vast metabolic network that processes planetary matter into a continuous stream of data: „the metabolism of the supply chain of extraction is also objectified in those unspectacular, nearly imperceptible practices and habits that constantly weave together the fabric of everyday life in the twenty-first-century city: sending an email, driving to work, ordering groceries through the internet” (Arboleda 2020, p. 13). Thus Arboleda shows that ‘innocent’ or ethically neutral everyday practices are simply no longer possible today.

ETHICAL REFLECTIONS: HANS JONAS AND PLANETARY RESPONSIBILITY

The environmental aspect of AI development calls for a deeper ethical analysis. And for that Hans Jonas's imperative of responsibility provides a natural foundational framework. Jonas argued that moral responsibility rests on three conditions: the efficacy of action, the foreseeability of its consequences, and the capacity for control. Applied to AI, these conditions imply that the global expansion of data infrastructures demands not just technical regulation but moral foresight.

Jonas expressed responsibility in two key imperatives. The positive imperative reads: “Act so that the effects of your action are compatible with the permanence of genuine human life” (Jonas 1984, p. 11). The negative imperative states accordingly: “Act so that the effects of your action are not destructive of the future possibility of such life” (Jonas 1984, p. 11). These imperatives mean that AI development must be judged not only in terms of its usefulness in our endeavors or the profits it generates, but also by its implications for the survival of humanity and the entire biosphere.

Jonas further stressed the principle of uncertainty or caution: “It is the rule, stated primitively, that the prophecy of doom is to be given greater heed than the prophecy of bliss” (Jonas 1984, p. 31). This principle, often summarized as *in dubio pro malo*, requires preventive measures to be taken whenever there is a risk of existential threats.

Crucially, Jonas extended responsibility beyond the human domain: “Now the whole biosphere of the planet ... claims its share of the respect owed to all that is an end in itself – that is: to all that is alive” (Jonas 1984, p. 8). This underlines the need to consider AI not only in relation to human interests but also to nonhuman life.

Finally, Jonas emphasized the proportionality of power and responsibility: “The demands of responsibility grow in proportion to the deeds of power” (Jonas 1984, p. 23). Given the transformative potential of AI, humanity’s obligation to act responsibly is correspondingly amplified.

At this point, the aforementioned notion of the planetary mine adds further depth. AI, situated within the context of the planetary mine, appears not as an immaterial system of algorithms but as a dense network of resource extraction, energy demand, and logistical supply chains. Not only do they leave a visible and increasingly deep mark on our planet, but in doing so they worsen the living conditions of future generations, both human and non-human. They thus fit into the broader context of anthropogenic environmental changes, the ethical context of which can be problematized using Hans Jonas’ concept of responsibility.

The material footprint of AI thus produces four types of consequences: tangible (emissions, water depletion), deferred (long-term climate change), those affecting nonhuman living beings (loss of biodiversity), and those affecting the nonliving environment (geological degradation). Each of these

has a direct or indirect impact on human existence on Earth, both today and in the future. Together, these highlight the necessity of a planetary ethics of AI, one that encompasses both human and nonhuman domains, as the survival of humanity is strictly dependent on the functioning of the entire planetary biosphere.

CONCLUSION

Artificial Intelligence is not only about algorithms or virtual platforms; it is part of a planetary system of extraction, energy, and material flows. Ethical reflection on AI should therefore go beyond purely anthropocentric concerns and embrace ecological responsibility. Hans Jonas's imperative of responsibility provides the normative foundation for this approach. AI development must not jeopardize the ecological preconditions of human life. Responsibility is anticipatory, demanding that future integrity be included in present choices. As Jonas put it: "Act so that the effects of your action are compatible with the permanence of genuine human life" (Jonas 1984, p. 11).

In conclusion, AI will only be genuinely "intelligent" if developed within the horizon of responsibility – where technological innovation is constrained by ecological boundaries and guided by a planetary ethic of care for both humanity and the biosphere.

REFERENCES

- Arboleda, M. (2020). *Planetary mine: Territories of extraction under late capitalism*. London – New York: Verso.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.
- Billiris, G., Gill, A., Bandara, M. (2025). *Privacy in the age of AI: A taxonomy of data risks*. <https://arxiv.org/pdf/2510.02357>
- Brynjolfsson, E., McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York: W.W. Norton & Company.

- China Water Risk (2024). *China ICT running dry? The rise of AI & climate risks amplify existing water risks faced by thirsty data centres*. Hong Kong: China Water Risk. <https://cwrrr.org/wp-content/uploads/2024/04/CWR-2024-China-ICT-running-dry-The-rise-of-AI-climate-risks-amplify-existing-water-risks-faced-by-thirsty-data-centres.pdf>
- European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence (Artificial Intelligence Act), and amending Regulation (EC) No 300/2008. *Official Journal of the European Union*, L 168, 1–120. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX-%3A32024R1689>
- Humphreys, D. (2024). AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI and Ethics*, 5(3), 541–555. <https://doi.org/10.1007/s43681-024-00443-4>
- International Energy Agency (IEA). (2024). *Electricity 2024: Analysis and forecast to 2026*. Paris: IEA. <https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf>
- Jonas, H. (1984). *The imperative of responsibility: In search of ethics for the technological age* (H. Jonas & D. Herr, Trans.). Chicago: University of Chicago Press.
- Labban, M. (2014). Deterritorializing extraction: Bioaccumulation and the planetary mine. *Annals of the Association of American Geographers*, 104(3), 560–576. <https://doi.org/10.1080/00045608.2014.892360>
- Microsoft Corporation. (2024). *2024 environmental sustainability report*. Redmond: Microsoft. <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Microsoft-2024-Environmental-Sustainability-Report.pdf>
- Musser, M., Lohn, A., Dempsey, J.X., Spring, J., Siva Kumar, R.S., Leong, B. et al. (2023). *Adversarial machine learning and cybersecurity: Risks, challenges, and legal implications*. <https://arxiv.org/pdf/2305.14553>
- Wang, Y., Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2), 246–257. <https://doi.org/10.1037/pspa0000098>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Public Affairs.

THREE SCENARIOS FOR THE DEVELOPMENT OF SUPERINTELLIGENCE

WOJCIECH GŁĄŻEWSKI

INTRODUCTION

The development of contemporary AI technologies is progressing in one general direction – computer programs, based on large language or graphic models, for which increasingly efficient processors are being built. Their potential is so great, and the scope for improvement so vast, that the development is still ongoing and will continue for at least the next few years. It is not surprising that the companies that create these systems and the engineers and scientists involved in their development see the future in this technology. However, it is not obvious that it will bring something that is expected as AGI or superintelligence. I believe that this type of software will achieve previously unheard of capabilities, becoming a computational AGI, i.e. a computer system with deep information processing capabilities. The potential for growth and optimization is visible, but there is no prospect of a qualitative leap that would move such systems from the category of weak intelligence in Searle's understanding (manipulating symbols to obtain a result) to the category of strong intelligence ("understanding", having internal states).

In my opinion, it is necessary to turn towards new technologies and base future AI systems on a different kind of substrate than the substrate of a digital computer. I propose to think of future AI systems not as computer programs, but as devices, which will allow us to see the perspective of other

technical possibilities and imagine AI qualitatively different from internet chat. The new quality of AI also means new ethical problems. Deleting a computer program does not raise moral doubts, but turning off a system that has grown over the course of several months or even years and that, having mastered human language by its own efforts, has told us about the richness of its internal states, may not be an easy decision.

PATHS TO THE FUTURE

The development of computer software supported by the possibility of obtaining large amounts of data from the Internet has caused a rapid leap in the field of natural language processing (Patil and Gudivada 2024). This leap has given impetus to the development of modern AI systems, so naturally a continuation is expected. One of the most famous texts is Aschenbrenner's "Situational Awareness" (2024), directly referring to the reality of Silicon Valley. Progress is seen there in the improvement of computer hardware and the optimization of algorithms for software. On the horizon is expected "a trillion dollar cluster", and one of the biggest challenges is "breaking the data wall". Duenas and Ruiz (2024a) assess five levels of progress in AI, which were published by OpenAI as their development strategy. They clearly refer to the functionality of the IT system, confirming that the market leaders are focusing on this direction.

AGI AND SUPERINTELLIGENCE

The difference between AGI and ASI (artificial superintelligence) can be reduced to the level of intelligence. AGI would have a level comparable to that of humans, and ASI would achieve a level significantly higher. However, it is worth thinking about them in another way: AGI is possible to achieve through direct build or another method that is within the range of feasibility and comprehension by humans, while ASI will be a system that can only be created by rising above human capabilities ("takeoff" in Bostrom's language). This is assumed to occur in a self-reinforcing process culminating in a technological singularity (Bostrom 2014, Chalmers 2010).

The actual implementation of a self-improving algorithm is problematic in itself (Schmidhuber 2009). The exponential growth of computing power is the most common argument for the possibility of exponential growth of intelligence, but a direct connection between one and the other has not been shown, and even less so the algorithm of their feedback. The case of humans even shows the opposite - the brain did not improve itself, but evolved through natural selection, and this is a different type of process. Another problem is the qualitative difference between AGI and ASI, which in my opinion will consist in the fact that the latter will be a fully “strong” intelligence, i.e. it will implement human-like cognitive processes.

In the absence of prospects for implementing “strong” cognitive processes in a computer substrate, I am inclined to see the possibility of initiating seed AI on a different type of substrate. What it will be, we do not know, but this idea alone brings analytical benefits (Thimm et al. 2024). If instead of AI or “digital minds” we speak directly of “machines”, then it is easier to see the gradation of dependencies between humans and machines, the problem of having personhood, or different approaches to autonomy: gradual, relational, attribution.

THREE SCENARIOS

Technological, economic and political challenges allow us to outline three scenarios for building a system that will aim at the level of superintelligence.

- **Scenario 1: Silicon Valley product**

The future ASI system will be created like any other utility software. The size of the system and its functionality will be planned, resources will be selected, hardware will be purchased, software will be coded and the system will be made operational. After initial tests, it will be made available to customers for use, just like ChatGPT or Dall-E. The key is that the investment will require a business plan, and low technical risk is expected. High risk will be accepted only in the face of the potential to take a high market position.

- **Scenario 2: Large-scale international project**

High technical and market risk, but the promise of scientific progress, are usually enough to make a decision on public funding. An important argument

will be socio-political issues, in this case the peaceful use of the possibilities of future superintelligence, i.e. the issue of its control. Such a scenario is proposed by Bostrom (2014). A similar proposal is made by Duenas and Ruiz (2024b) suggesting the establishment of The Ministry of Artificial Intelligence.

- **Scenario 3: Seed AI**

In this scenario, the system will be initiated as a seed and its development led in the desired direction. Ultimately the system itself will take control of the progress when the capabilities of human intelligence are exhausted. It is difficult to estimate the time scale of this process, and there are no potential technical solutions on the horizon. This undertaking will be burdened with the greatest risk and will most likely require the largest outlay of funds, but in the case of success, the profit will be the greatest. Both a private visionary and a public research institution can start the growth of AI. The project requires long-term financing, but above all, special courage and vision.

The first two scenarios are similar in the way they are implemented, i.e. planning, building and launching the system in the expected form. This means that knowledge of how these systems work, i.e. technology, is needed. With a very high probability, both of them will bring a computer system with AGI capabilities. But will they bring superintelligence? Remembering that in order for it to be created, the barrier of human competence must be crossed, only the hypothesis of self-improvement gives hope for its achievement. The third scenario is therefore the most promising. What must be mastered at the start is the ability to maximize the intelligence factor in the system. We do not know where the process will lead, we will just aim up, above our heads.

ETHICAL CHALLENGES

The most debated issues are related to the behavior of future superintelligence, once it begins to operate. The motives of such a system will not always be visible as clear goals. A valuable perspective is provided by Goertzel (2015), who reminds us that during the development of the system, its values will also change, so it is worth ensuring that the system grows in close relations with humans. An interesting approach is proposed by Totschnig (2019), who

names the problem of relations with ASI as a political problem. According to him, the situation will be more like “mutual vulnerability” than subordination for the purpose of control; a Hobbesian state of nature will occur between us and ASI. Access to resources will be contested, especially energy (Stiefel and Coggan 2023), but also power (Bales et al. 2024).

The development of seed AI into superintelligence will bring artificial systems not only phylogenesis but also ontogeny, important processes for now reserved for living species. As the affinity for living beings increases, questions will arise about the extent to which the categories previously reserved for humans also apply to artificial beings. The personal development of self-improving systems will be a special period in which the system will no longer be a collection of parts lying in storage, but will not yet become a fully independent (autonomous) entity. An even greater challenge is the moment of transition from one generation to the next, because only the system will “understand” what it is doing. For humans, the only remaining decision will be to make available the resources the system requests (Majot and Yampolskiy 2017). The system may even decide on its own to cease development if it considers that the incurred costs will not translate into increased intelligence. Expectations of exponential growth can turn into disappointment if growth continues to yield diminishing results. Will a decision be made to turn it off then? What will be the circumstances of this decision if the system shows mental activity, actually approaching some kind of artificial being?

REFERENCES

- Aschenbrenner, L. (2024) Situational Awareness: The decade ahead. <https://situational-awareness.ai/> (accessed: 12 July 2024).
- Bales, A., D’Alessandro, W., Kirk-Giannini, C.D. (2024). Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass* 19 (2).
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17 (9–10), 9–10.
- Duenas, T., Ruiz, D. (2024a). The Path to Superintelligence: A Critical Analysis of Open AI’s Five Levels of AI Progression.
- Duenas, T., Ruiz, D. (2024b). The Ministry of Artificial Intelligence: A Catalyst For National AI Ecosystems and Global Cooperation in the Age of Superintelligence.

- Eden, A., Moor, J. Søraker, J., Steinhart, E. (2013). Singularity Hypotheses: A Scientific and Philosophical Assessment.
- Goertzel, B. (2015). Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's From AI to Zombies, and Weaver and Veitas's "Open-Ended Intelligence". *Journal of Ethics and Emerging Technologies* 25, 55–87.
- Majot, A., Yampolskiy, R. (2017). Diminishing Returns and Recursive Self Improving Artificial Intelligence.
- Patil, R., Gudivada, V. (2024). A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences*, 14(5), 2074.
- Schmidhuber, J. (2009). Ultimate Cognition à la Gödel. *Cogn Comput* 1, 177–193.
- Stiefel, K.M., Coggan, J.S. (2023). The energy challenges of artificial superintelligence. *Frontiers in Artificial Intelligence*, 6.
- Thimm, C., Gramelsberger, G., Mayer, M., Piller, F. (2024). From automation to autonomy: Human-machine relations in the age of artificial intelligence. *Human-Machine Communication*, 9, 7–24.
- Totschnig, W. (2019). The problem of superintelligence: political, not technological. *AI and Society* 34 (4), 907–920.

AUTHORS

Jarosław Chudziak – Dr., Warsaw University of Technology, Faculty of Electronics and Information Technology
e-mail: jaroslaw.chudziak@pw.edu.pl

Wojciech Głażewski – doctoral student, University of Białystok
e-mail: w.glazewski@uwb.edu.pl

Hajo Greif – Dr. Hab., Warsaw University of Technology, Faculty of Administration and Social Sciences
e-mail: hans-joachim.greif@pw.edu.pl

Roman Krzanowski – Dr., The Pontifical University of John Paul II in Cracow, Faculty of Philosophy
e-mail: rmkrzan@gmail.com

Paweł Łupkowski – Dr. Hab., Department of Logic and Cognitive Science, Adam Mickiewicz University, Poznań
e-mail: pawel.lupkowski@amu.edu.pl

Andrzej Malec – Dr., Foundation for Computer Science, Logic and Formalized Mathematics, Warsaw
e-mail: a.malec@amdp.strefa.pl

Łukasz Mścislawski – Dr., Wrocław University of Science and Technology, Institute of Organization and Management
e-mail: lukasz.mscislawski@gmail.com

Daniel Myślicki – student, Warsaw University of Technology, Faculty of Administration and Social Sciences
e-mail: danielmyslicki@gmail.com

Karol Narożniak – doctoral student, Military University of Technology, Warsaw

Tomáš Ondráček – Dr., Department of Corporate Economy, Masaryk University, Brno

e-mail: ondracek.t@gmail.com

Paweł Polak – Dr. Hab., The Pontifical University of John Paul II in Cracow, Faculty of Philosophy

e-mail: pawel.polak@upjp2.edu.pl

Marek Porzeżyński – Dr., Warsaw University of Technology, Faculty of Administration and Social Sciences

e-mail: marek.porzezynski@pw.edu.pl

Marcin Rabiza – Dr., Leiden University, Institute for Philosophy, Leiden-Polish Academy of Sciences, Institute of Philosophy and Sociology, Warsaw

e-mail: marcin.rabiza@gssr.edu.pl

Bartosz Sawicki – Dr. Hab., Warsaw University of Technology, Institute of Theory of Electrical Engineering

e-mail: bartosz.sawicki@pw.edu.pl

Radosław Siedliński – Dr., Polish-Japanese Academy of Information Technology, Warsaw

e-mail: radoslaw.siedlinski@gmail.com

Bartłomiej Skowron – Dr. Hab., Warsaw University of Technology, Faculty of Administration and Social Sciences

e-mail: bartlomiej.skowron@pw.edu.pl

Krzysztof Soloducha – Dr. Hab., Military University of Technology, Faculty of Humanities, Warsaw

e-mail: krzysztof.soloducha@wat.edu.pl

Paweł Stacewicz – Dr., Warsaw University of Technology, Faculty of Administration and Social Sciences

e-mail: pawel.stacewicz@pw.edu.pl

Michał Śmialek – Dr. Hab., Warsaw University of Technology, Institute of Theory of Electrical Engineering

e-mail: michal.smialek@pw.edu.pl

Mariusz Urbański – Dr. Hab., Department of Logic and Cognitive Science, Adam Mickiewicz University, Poznań

e-mail: mariusz.urbanski@amu.edu.pl

Stuart Weinstein – Dr., Aston University, College of Business and Social Sciences

e-mail: s.weinstein@aston.ac.uk

PUBLICATIONS
OF THE MAREK DIETRICH
INSTITUTE CONTEMPORARY CIVILIZATION PROBLEMS

All publications since 2019 and also earlier publications marked “” are available
in pdf format at www.ipwc.pw.edu.pl*

Rok 2025

- XCII – Ethical Aspects of AI
- XCi – Wystąpienia i wykłady inaugurujące rok akademicki 2024/2025
- XC – Ewolucja kultury akademickiej

Rok 2024

- LXXXIX – Edukacja szkolna z perspektywy nauk pedagogicznych
- LXXXVIII – Wystąpienia i wykłady inaugurujące rok akademicki 2023/2024
- LXXXVII – Cywilizacyjna rola lasów

Rok 2023

- LXXXVI – O poznawaniu i kształtowaniu świata
- LXXXV – Nieufność: źródła i konsekwencje
- LXXXIV – Wystąpienia i wykłady inaugurujące rok akademicki 2022/2023
- LXXXIII – Prawdy i półprawdy w żywieniu człowieka
- LXXXII – Prawdy i półprawdy w żywieniu człowieka – książka abstraktów

Rok 2022

- LXXXI – Działalność naukowa – niedoceniany czynnik rozwoju cywilizacyjnego Polski
- LXXX – Podsumowanie dokonań Instytutu 1996–2021
- LXXIX – Ewolucja cywilizacyjnej roli i społecznego odbioru nauki. Część 2
- LXXVIII – Wystąpienia i wykłady inaugurujące rok akademicki 2021/2022

Rok 2021

- LXXVII – Młodzi dorośli: identyfikacje, postawy, aktywizm i problemy życiowe
- LXXVI – Stosunek do szerepów ochronnych: sceptycyzm wobec nauki

- LXXV – Wystąpienia i wykłady inaugurujące rok akademicki 2020/2021
- LXXIV – Ewolucja cywilizacyjnej roli i społecznego odbioru nauki
- LXXIII – Zmiany klimatu i ich następstwa

Rok 2020

- LXXII – Nauczanie po pandemii. Nowe pytania czy nowe odpowiedzi na stare pytania?
- LXXI – Wykłady inaugurujące rok akademicki 2019/2020

Rok 2019

- LXX – Szanse i wyzwania dla polskich wydawnictw i czasopism naukowych
- LXIX – Wykłady inaugurujące rok akademicki 2018/2019

Rok 2017

- LXVIII – Wykłady inaugurujące rok akademicki 2017/2018
- LXVII – Autonomia uczelni i środowiska akademickiego – odpowiedzialność i etos akademicki*
- LXVI – Wykłady inaugurujące rok akademicki 2016/2017
- LXV – Student pierwszego roku

Rok 2016

- LXIV – Wykłady inaugurujące rok akademicki 2015/2016*
- LXIII – Miejsce nauk podstawowych w kształceniu wyższym*

Rok 2015

- LXII – Praktyczne aspekty rekrutacji na studia od roku akademickiego 2015/2016
- LXI – Wykłady inaugurujące rok akademicki 2014/2015*

Rok 2014

- LX – Perspektywy rozwoju kształcenia zawodowego w Polsce*
- LIX – Badania PISA – przeszłość, teraźniejszość i przyszłość*
- LVIII – Rekrutacja na studia od roku akademickiego 2015/2016 w kontekście zmian w systemie oświaty. Informator dla szkół wyższych*

Rok 2013

- LVII – Wykłady inaugurujące rok akademicki 2013/2014*
- LVI – Wykłady inaugurujące rok akademicki 2012/2013*

Rok 2012

- LV – Problemy nauczania chemii w szkołach średnich i wyższych*
- LIV – Wykłady inaugurujące rok akademicki 2011/2012*

Rok 2011

- LIII – Problemy nauczania biologii w szkołach średnich i wyższych*
- LII – Problemy nauczania fizyki w szkołach średnich i wyższych*
- LI – Wykłady inaugurujące rok akademicki 2010/2011*
- L – Strategia nauczania matematyki w Polsce – wdrożenie nowej podstawy programowej*

Rok 2010

- XLIX – Natura 2000. Szanse i zagrożenia*
- XLVIII – Współpraca szkół średnich i wyższych*
- XLVII – Podsumowanie dwunastolecia 1996–2008 – Marek Dietrich*
- XLVI – Wykłady inaugurujące rok akademicki 2009/2010*

Rok 2009

- XLV – Społeczeństwo polskie wobec narodzin III Rzeczypospolitej (1988–1990)*
- XLIV – Woda w obszarach nieurbanizowanych*

Rok 2008

- XLIII – Prywatność – prawo czy produkt?*
- XLII – Polscy uczniowie w świetle badań PISA*
- XLI – Warszawa Akademicka*
- XL – Warszawa Akademicka – Seminarium*

Rok 2007

- XXXIX – Czasopisma naukowe – zmierzch czy transformacja?*
- XXXVIII – Obraz postępu i zagrożeń cywilizacyjnych w mediach
- XXXVII – Uczyć myśleć

Rok 2006

- XXXVI – Wizja polskich uczelni w społeczeństwie globalnym
- XXXV – Rola symboli
- XXXIV – Humanizm i technika

Rok 2005

- XXXIII – Zagadnienia bezpieczeństwa wodnego
- XXXII – Polskie uczelnie XXI wieku
- XXXI – Zagadnienia bezpieczeństwa energetycznego
- XXX – Emigracja – zagrożenie czy szansa?

Rok 2004

- XXIX – Decyzje edukacyjne
- XXVIII – Uczelnie a innowacyjność gospodarki
- XXVII – Internet i techniki multimedialne w edukacji

Rok 2003

- XXVI – Kierunki kształcenia i standardy nauczania w polskim szkolnictwie wyższym
- XXV – Zarządzanie bezpieczeństwem w sytuacjach kryzysowych

Rok 2002

- XXIV – Jakość kształcenia i akredytacja w szkolnictwie wyższym w Polsce
- XXIII – Autorytet uczelni
- XXII – Problemy etyczne w nauce
- XXI – Bezpieczeństwo człowieka we współczesnym świecie
- XX – Pamięć i działanie

Rok 2001

- XIX – Ekonomiczne efekty edukacji w Polsce
- XVIII – Wolność a bezpieczeństwo
- XVII – Ekonomiczne i społeczne efekty edukacji

Rok 2000

- XVI – Ekonomiczne i społeczne efekty edukacji
- XV – Czy kryzys demograficzny w Polsce?
- XIV – Produkcja, konsumpcja i technika a ocieplenie klimatu
- XIII – Kształcenie międzyuczelniane. Studium warszawskie

Rok 1999

- XII – Władza i obywatel w społeczeństwie informacyjnym
- XI – Koszty kształcenia w szkołach wyższych w Polsce. Model kalkulacyjnych kosztów kształcenia
- X – Problemy etyczne techniki
- IX – Bezpieczeństwo człowieka we współczesnym świecie

Rok 1998

- VIII – Polska a integracja europejska w edukacji. Aspekty informatyczne
- VII – Misja uczelni
- VI – Bezpieczeństwo człowieka we współczesnym świecie
- V – Instrumenty rozwoju systemu kształcenia w Polsce

Rok 1997

- IV – Akademicka Komisja Akredytacyjna. System oceny jakości kształcenia i akredytacji w szkolnictwie wyższym
- III – Jakość kształcenia w szkołach wyższych
- II – Etyka zawodowa
- I – Ochrona własności intelektualnej

