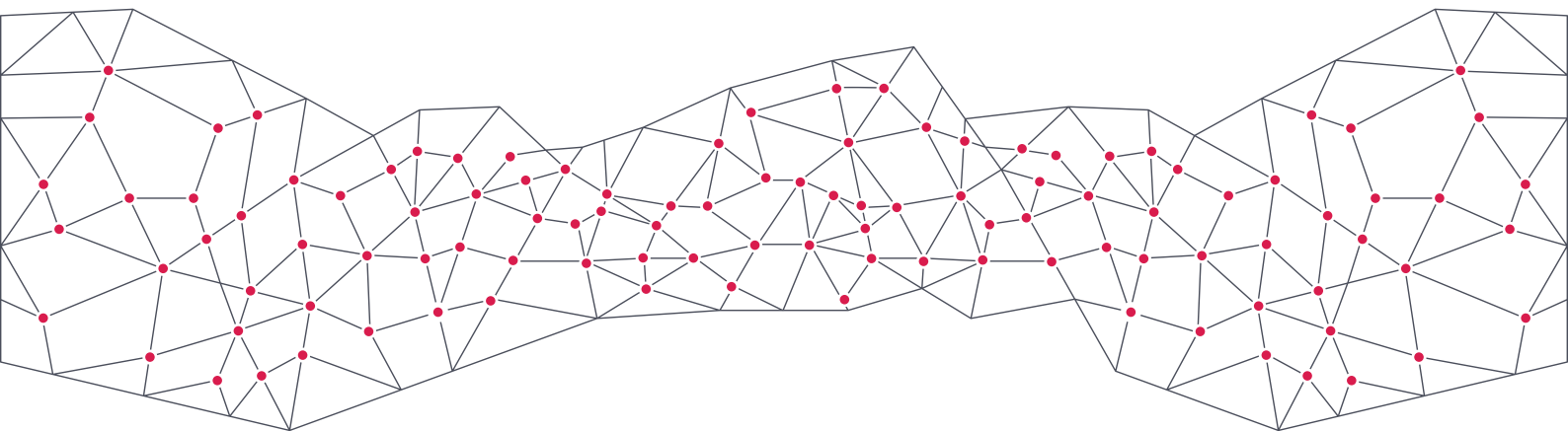


INTERNATIONAL CONFERENCE

EMERGE 2024

ETHICS OF AI ALIGNMENT

BOOK OF ABSTRACTS



Organizers

Digital Society Lab, Institute for Philosophy and Social Theory,
University of Belgrade

Institute for Artificial Intelligence Research and Development of
Serbia

Editors

Simona Žikić

Ana Lipij

Jelena Novaković

Design

Jelena Novaković

Publisher

Institute for Philosophy and Social Theory, University of Belgrade

ISBN 978-86-82324-90-4

This Book of Abstracts was realised with the support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, according to the Agreement on the realisation and financing of scientific research 451-03-66/2024-03/ 200025

Belgrade, 2024

International Conference

EMERGE 2024

Ethics of AI Alignment

11-12 December 2024

**Institute for Philosophy and Social Theory,
University of Belgrade**

CONTENTS

Scientific and Organizing Committee	1
Introduction	4
AI Ethics, Environmental Technology, and More-Than-Human Ecologies	6
Art and AI	23
AI in Education	57
Recommendation and Ranking Algorithms.....	102
Health-Tech and Health Literacy	115
Media, Freedom of Expression, and Democracy.....	122
Paradigms of AI	166
Philosophy of AI	177
Religion and AI	193

Scientific Committee

- **Ljubiša Bojić**, Institute for Philosophy and Social Theory, University of Belgrade / Institute for Artificial Intelligence Research and Development of Serbia (Chair) (Serbia)
- **Ana Lipij**, Institute for Philosophy and Social Theory, University of Belgrade
- **Bojana Romić**, Malmö universitet (Sweden)
- **Branislav Kisačanin**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Bruno Daniel Ferreira da Costa**, Universidade da Beira Interior (Portugal)
- **Čedomir Markov**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Corina Paraschiv**, Université Paris Cité (France)
- **Dejan Grba**, Institute of Creativity and Innovation, University for the Creative Arts London / Xiamen University (UK/China)
- **Dragiša Žunjić**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Dubravko Ćulibrk**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Ivana Krtolica**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Jelena Guga**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Jelena Novaković**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Jordi Vallverdú**, Universitat Autònoma de Barcelona (Spain)
- **Jörg Matthes**, Universität Wien (Austria)
- **Max Talanov**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Mikhail Bukhtoyarov**, University of Siberia (Russia)

- **Mustafa Ali**, Faculty of Science, Technology, Engineering & Mathematics, the Open University (UK)
- **Simona Žikić**, Institute for Philosophy and Social Theory, University of Belgrade / Faculty of Media and Communications (Serbia)
- **Stefan Lorenz Sorgner**, John Cabot University (Italy)
- **Susanna Gordleeva**, Nizhny Novgorod State University / Baltic Federal University (Russia)
- **Vera Mevorah**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Vladimir Cvetković**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Yashar Deldjoo**, Politecnico di Bari (Italy)
- **Željko Radinković**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Zoran Erić**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Zorica Dodevska**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)

Organizing Committee

- **Simona Žikić**, Chair, Institute for Philosophy and Social Theory, University of Belgrade / Faculty of Media and Communications (Chair) (Serbia)
- **Ana Lipij**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Čedomir Markov**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Dragiša Žunjić**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Ivana Krtolica**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Jelena Guga**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Jelena Novaković**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Knud Ryom**, Aarhus Universitet (Denmark)
- **Ljubiša Bojić**, Institute for Philosophy and Social Theory, University of Belgrade / Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Max Talanov**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)
- **Milan Radić**, Institute for Philosophy and Social Theory (Serbia)
- **Mirjana Nećak**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Vera Mevorah**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Željko Radinković**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Zoran Erić**, Institute for Philosophy and Social Theory, University of Belgrade (Serbia)
- **Zorica Dodevska**, Institute for Artificial Intelligence Research and Development of Serbia (Serbia)

INTRODUCTION

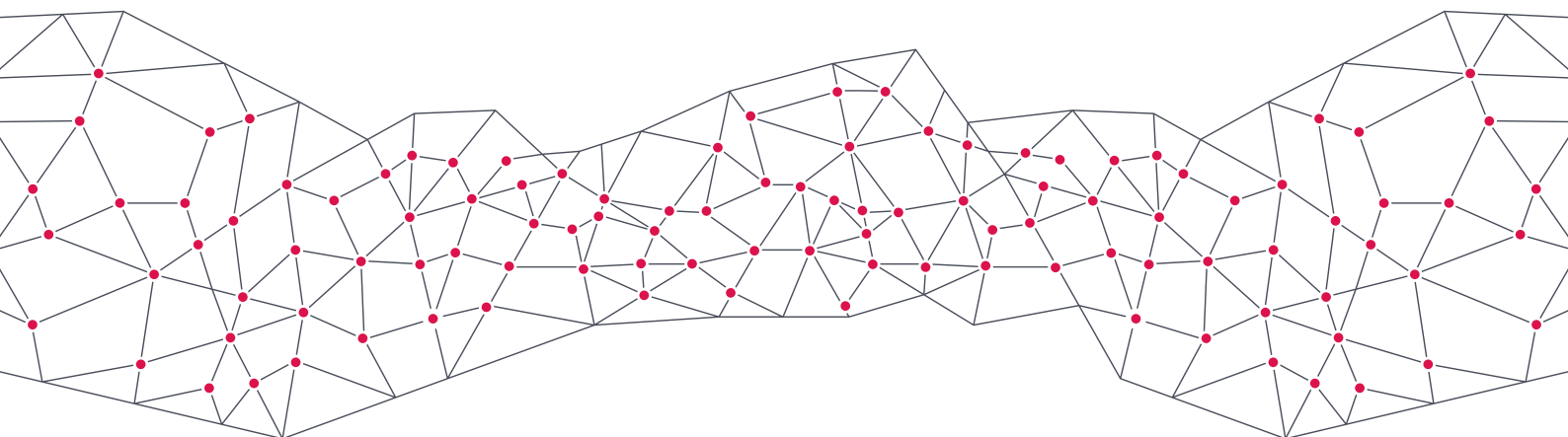
The Digital Society Lab of the The Institute for Philosophy and Social Theory of the University of Belgrade and the Institute for Artificial Intelligence Research and Development of Serbia are pleased to announce the International Conference and Forum EMERGE 2024: Ethics of AI Alignment, to be held on December 11th, 12th, and 13th. EMERGE is an annual event that brings together scholars, researchers, practitioners, and policymakers from around the world to discuss and debate the ethical, social, environmental, and cultural implications of emerging technologies, focusing this year on aligning artificial intelligence (AI) with human values and interests.

The goal of EMERGE 2024 is to foster enriching discussions and generate insights into how burgeoning AI technologies intersect, influence, and are incorporated into various spheres of life. Particularly, we aim to highlight potential ethical implications and chart directions of navigation for the rapidly evolving digital landscape.

Advancements in AI have ushered in a new era of technological innovation, promising to revolutionize industries, enhance productivity, and improve the quality of life. However, as AI systems become increasingly integrated into various aspects of society, questions about their ethical implications have come to the forefront of public discourse. Central to these discussions is the concept of AI alignment—ensuring that AI systems are redesigned and deployed in ways that align with human values, goals, and societal norms. The International Conference on the Ethics of AI Alignment seeks to explore the multifaceted ethical challenges and opportunities arising from the quest for alignment. By bringing together scholars, researchers, practitioners, and policymakers from diverse disciplines and backgrounds, the conference aims to foster critical dialogue, interdisciplinary collaboration, and insights into the ethical dimensions of AI alignment. Through a series of subtopics,

participants will delve into specific ethical dilemmas, share innovative research findings, and propose solutions to address the complex ethical issues at the intersection of AI and society. As we navigate the ethical landscape of AI alignment, we must engage in thoughtful reflection, ethical deliberation, and responsible stewardship to ensure that AI technologies serve the common good and uphold fundamental principles of justice, fairness, and human dignity.

Whether you are a professional interested in the latest advancements in AI, a student exploring career paths or simply an AI enthusiast looking to encompass the broader societal implications of the industry, EMERGE 2024 offers a comprehensive look into the ethics shaping the future of artificial intelligence.



AI ETHICS, ENVIRONMENTAL TECHNOLOGY, AND MORE-THAN-HUMAN ECOLOGIES

Vera Mevorah, Andrija Filipović, Ivana Krtolica & Zoran Erić

In ecology and environmental engineering, AI has emerged as a powerful tool that seems well aligned with the needs and goals of our societies. AI tools have been instrumental in prompt decision-making and monitoring processes such as flood forecasting systems and predictions related to water, air, or soil quality. Computer vision techniques enable the use of satellite images for analysis, which, in addition to flood forecasting systems, is especially important for monitoring dangerous or sick animals in inaccessible areas. AI also enables the analysis of consumption patterns, providing recommendations for energy savings. Real-time monitoring is one of the main advantages of using AI; therefore, the implementation of sensor-equipped measurement stations is of paramount importance in providing datasets for AI modeling. AI serves to provide timely, accurate, and sufficient monitoring data or as an additional tool in decision-making processes to mitigate and prevent natural disasters. We are interested in research conducted from this affirmative perspective that explores technical challenges of aligning AI technologies with human needs and goals.

Yet, we also wish to address ethical, political, and social problems and complexities that are not considered enough in the development of AIs. We are interested in projects that closely address such challenges and approach AI engineering with a critical, scrutinizing eye. For

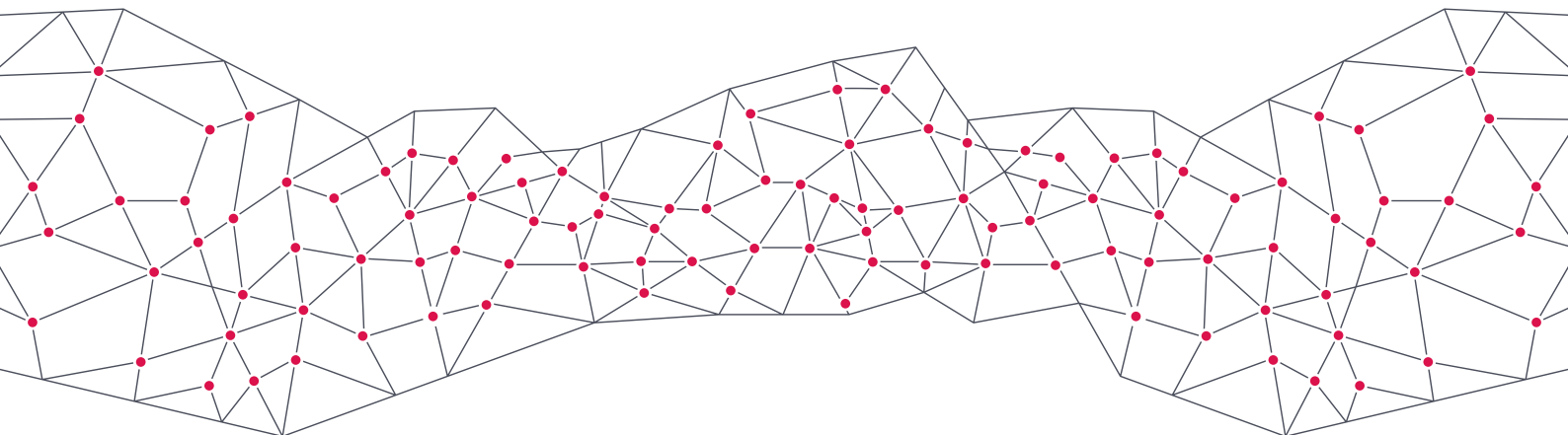
example, in determining responsibility for inaccurate AI predictions, we can ask: does the fault lie with the human overseeing the AI or the machine itself? Or how should we approach the problem during emergencies when crucial evacuation information remains solely accessible to humans? Could it be that the solution to more than one ethical challenge lies in employing AI as an assistant, rather than allowing it to dominate the decision-making process? Can we make the machine intelligent, but not responsible or reasonable? How can we ensure that machines do not exceed the boundaries of human rights and ethics? Can we protect humans from themselves?

Social sciences and humanities often consider the development and use of AI technology as a challenge rather than a solution to environmental problems. The growth of digital technologies is strongly reliant and interconnected with economic growth and is the privilege of technologically advanced countries. Environmental humanities present a critical approach within the broader field of social sciences and humanities, raising many questions regarding the interrelations between the environment and economic and technological advancement. How does AI impact the “more-than-human” and “other-than-human”? How does AI change human perceptions of non-human entities? Are these entities just resources to be managed or used, or can one think differently about them in their relation to AI?

Another important perspective comes from critical energy studies and critical infrastructure studies. These areas of inquiry are interrelated because the issue of transforming the other-than-human into an energy resource and material to be consumed by AI is of key importance for environmental protection. How and where are rare resources extracted, and who benefits from the extraction? We need to critically engage with political, social, and technical systems that, through various infrastructures, enable the kind of transformations that lead to environmental degradation, devastation, and species extinction.

Furthermore, what powers AI and how is AI powered? What role does the fossil economy play in enabling AI? What is the promise of green AI? We can also ask what remains after AI. During the processes of producing the AI infrastructures, as well as the energy necessary for the functioning of the AI, distinct types of waste and discards are

created. What happens when we think about the use of AI from the point of view of the waste it produces? Who and what is affected, and how? Digital and e-waste are not considered enough in scientific research. Are digital degrowth and other alternatives good enough to provide us with social models for using AI in an environmentally and socially responsible way? These questions are important not only for individual non-human species but considering the Anthropocene – the planet itself. Finally, paraphrasing Albert Einstein, we ask: can we solve the problems we have created with the same thinking that created them?



The Emergence of Trust in Human-Otheroid Interactions through Empathy

Abootaleb Safdari

This paper aims (1) to advocate for the possibility of establishing trustworthy relationships with robots and AI systems (automata) and (2) to explore the mechanisms that enable such trust. Its negative step begins by critically analyzing arguments against trusting automata, which typically emphasize reliability as the basis for trust and suggest that trust is inherently human, making it inappropriate for automata (Alvarado, 2023; Bryson, 2018; Hatherley, 2020; Metzinger, 2019; Ryan, 2020). Additionally, these arguments assert that trust in automata is socially and morally harmful. These arguments are challenged based on a relational approach, according to which there is no clear-cut distinction between human persons and automata. Furthermore, they overlook the essential role of empathy in human-machine interactions and mischaracterize the nature of trust.

In the positive step, I will propose an empathy-based framework for trustworthy relationships with automata. Accordingly, trust does not originate from mere reliability but rather from an empathic relationship with automata. Initially, we build an empathic relationship with them, leading us to perceive and interact with them not as mere technological artifacts but as minded others or, more precisely, pseudo-others. This is why I prefer to call them “otheroids.” To fully grasp the concept of “otheroids,” it is important to consider a phenomenological perspective on empathy.

From this perspective, an empathic relationship—i.e., a relationship through which we grasp a certain entity as an other—takes place on three different levels: the *that* level (experiencing an entity as a minded one), the *what* level (determining the other’s specific state of mind), and the *why* level (reasoning about the other’s past and future mental states) (Zahavi, 2014, pp. 167–168). Imagine you and a friend are having a slice of chocolate cake in a café. First, you perceive your friend as an entity that has a mind and thus mental or internal states. This allows you to grasp what her specific mental states are, e.g., she is enjoying the taste of the cake. Finally, you are able to comprehend why she decides to visit this café—for example, because she already knows this café serves delicious cakes and will continue to do so

in the future. This articulated understanding enables us to clearly differentiate a pseudo-other, or what I prefer to call an “otheroid,” from a perfect other. While the empathic relationship with a perfect other—a normal person—encompasses all three levels, the empathic relationship with an otheroid is restricted to the first, basic level.

Otheroids prompt the subject to experience herself as a de-centered self—no longer solely responsible for the interaction. Instead, the otheroid imposes itself upon the subject, challenging the human subject’s role as the sole author of the situation.

This decentering of the self is significant because it introduces a form of irreducible vulnerability into the human-otheroid relationship. The subject can no longer rely solely on her own agency to direct or control the interaction. She must account for the presence and influence of the otheroid, which means that part of the outcome is no longer entirely in her hands. This vulnerability is not simply a weakness but forms the seed for trust development. Trust requires a degree of risk and uncertainty; it is in moments of vulnerability that trust is both tested and developed.

I will then examine how this seed of vulnerability grows into a basic form of trust with otheroids and how the de-centered self evolves into a trusting self. At the core of this transformation is the building of a rich history of interaction, which is essential to the development of trust. This process involves a gradual shift in how the subject perceives the otheroid. What begins as cautious engagement, based on the minimal “that” level of empathy—recognizing the otheroid as a pseudo-other—develops into a more nuanced and meaningful interaction.

Keywords: trustworthy relation; empathy; otheroids; harmonious; relational attitude

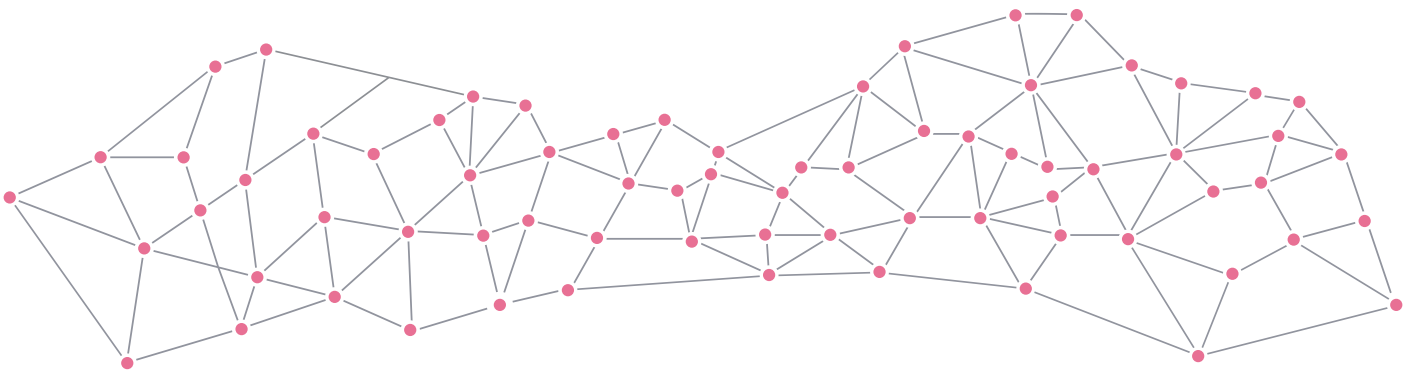
Abootaleb Safdari

Philosophy Postdoc Researcher

University of Bremen

Abootaleb Safdari is a postdoctoral researcher in philosophy at the University of Bremen. He works at the intersection of philosophy of mind and AI/robotics. Based on the phenomenological tradition, he seeks to develop a novel relational framework that reconceptualizes AI systems and robots as genuine partners [as the other] in interaction, moving beyond their traditional characterization as mere tools. This has the potential to inform our understanding of critical issues surrounding AI, such as the nature of trust in human-AI relationships and the broader ethical implications of AI development.

Contact: philolover67@gmail.com



Technological Optimism and its Discontents: Why the AI Hype Around Climate Change is More Harmful than Useful for Global Climate Action?

Ljupcho Stojkovski

There has been increasing public and academic interest in Artificial Intelligence (AI) and its (potential) use in the fight against climate change in recent years. Even the UN has joined the hype, stressing, for example, that “AI can revolutionize the world’s approach to carbon neutrality and usher in an era of intelligent sustainability on a global scale.” While many who endorse AI for climate action are rightfully warning against “solutionism,” or seeing this technology as a panacea for the problem of climate change, there is nonetheless an optimistic, hopeful tone surrounding the potential of this technology for climate action.

There is no doubt that AI technologies could indeed be valuable for climate action, such as in improving weather forecasts, optimizing energy consumption through smart grids or tracking GHG emissions. Nevertheless, I will argue that this looming hope built around AI’s potential in the fight against climate change is more harmful than useful.

Firstly, many AI systems that could help climate action are still in development. Secondly, as a resource- and energy-intensive technology, some AI systems could potentially worsen the problem of climate change and thus pose a new threat to the environment. Additionally, as inequality is found to be a strong predictor of environmental degradation—and as climate change exacerbates inequality—AI technologies, reliant on biased data and available predominantly in wealthy countries and among privileged groups, could further amplify inequality, thereby worsening climate change.

Finally, even if all of the above challenges are somehow addressed, the AI hype creates false hope that a solution is just around the corner. It gives the impression that the problem of climate change is predominantly a technological one. In reality, the issue is fundamentally political, ethical, and legal. Moreover, the global focus on AI as a key element in the fight against climate change inadvertently supports the prolongation of the climate status quo—failing to regulate and reduce GHG emissions now. This is especially problematic given the temporal

urgency of tackling climate change before its consequences become irreversible.

Keywords: climate change; climate action; AI; technology

Ljupcho Stojkovski

Associate professor

Faculty of Law "Iustinianus Primus," Ss. Cyril and Methodius University

Ljupcho Stojkovski is an associate professor at Ss. Cyril and Methodius University in Skopje, Faculty of Law "Iustinianus Primus," where he teaches international public law and international relations courses. He has an MSc and a PhD in international law and international relations from the same University, and an MA in philosophy from KU Leuven. His research interest has been focused on the interplay between legal, political, and ethical aspects of international security issues and the related questions of responsibility and global governance. His more recent research interests also include climate action and sustainable development, as well as AI and philosophy of technology.

Contact: stojkovski_ljupco@yahoo.com

Nuclear Powered, Luxury Data Capitalism: Will Robots Control Energy in the Future?

Stefan Aleksić, Slobodan Bubnjević

For some time, large technological companies have been announcing that their energy needs will increase significantly due to the high power demands of their data centers. Their proposed solution: nuclear energy. Long considered a pariah among energy sources—despite its capacity to deliver vast amounts of reliable and clean energy—nuclear power is making a comeback.

This revival, however, has taken on a different character. Big tech companies are looking to nuclear energy to meet their energy demands. Microsoft has announced plans to invest in the revival of the Three Mile Island facility, while Google intends to use small modular reactors (SMRs) to power its data centers. Meanwhile, a former Microsoft owner is pursuing advanced projects, such as the “traveling wave” fast breeder molten salt reactor—an incredibly ambitious technological endeavor, at least on paper.

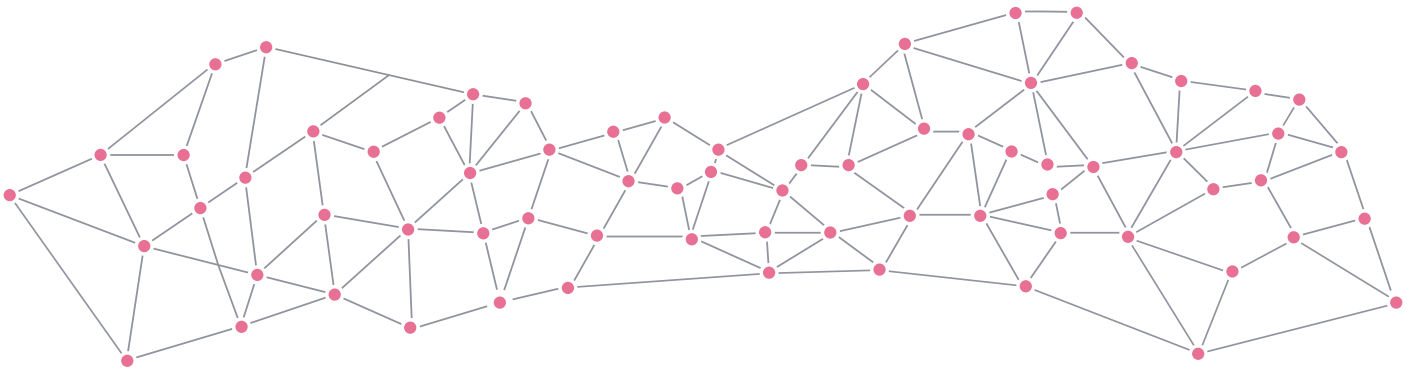
This marks an entirely new development. Historically, the nuclear industry has been closely associated with state ownership. With few exceptions (such as the United States, where the organizational structure is somewhat different), most nuclear facilities worldwide have been state-owned in some form. The idea of private ownership of nuclear reactors was almost unimaginable. Yet, the industry is now slowly beginning to privatize, with new reactor designs—specifically small modular reactors—seemingly developed to enable private capital ownership of nuclear facilities.

However, this shift raises significant concerns. Privatizing such critical infrastructure may lead to the “leaning” of production processes, a favored strategy of investors aiming to cut costs. The nuclear industry’s impressive safety record has been sustained by its robust and integrated supply chain, where minimal room exists for such cost-cutting measures. If companies begin to privately own and operate nuclear facilities, this dynamic is likely to change, potentially compromising safety standards.

This presentation aims to analyze the growing energy demands of

tech companies, their proposed solution—privately owned nuclear facilities—and the potential risks of this trend. Specifically, it will focus on the latest development of using nuclear facilities as energy sources to power data centers.

Keywords: energy; data; nuclear energy; production process; supply chains



Stefan Aleksić

Phd Candidate

Faculty of Philosophy, University of Belgrade

Stefan Aleksić earned his Master's degree at the Faculty of Political Sciences and Bachelor's degree at the Faculty of Philosophy, University of Belgrade. He currently works as a journalist and is a PhD candidate at the Department of Anthropology, Faculty of Philosophy, University of Belgrade. He is also a Chief Editor of web portal "Nuclear Perspective."

Contact: lowspeedyoyo@gmail.com

Slobodan Bubnjević

Scientific journalist

Expert advisor for communications at the Institute of Physics in Belgrade.

Slobodan Bubnjević (b. 1978, Rijeka, Croatia) is a scientific journalist and writer with a degree in experimental physics from the Faculty of Physics, University of Belgrade. A graduate of the Mathematical Gymnasium in Belgrade, he is currently employed as an expert advisor for communications at the Institute of Physics in Belgrade. Previously, he worked at the Center for the Promotion of Science and as a journalist for the weekly magazine Vreme. He contributes regularly to both international and domestic media, including Physics Today, RTS, Politika, Vreme, National Geographic Serbia, Klima 101, and Odiseja. He also serves as the editor of the "Science through Stories" portal. Bubnjević is the author of three books: *The Alchemy of the Bomb*, and two collections of stories, *Fear of the Draft* and *Perturbations and Other Troubles*. He has also written eight radio dramas, broadcast on Radio Belgrade, and is a member of the Serbian Literary Society. His novel *The Seventh Nation* won the prestigious "Borislav Pekić" Foundation Award. He lives in Pančevo with his wife and two daughters.

Contact: sbubnjevic@gmail.com

Taking Care of Digital Environments: Towards an Ecology of AI

Silvia Dadà

The relationship between human beings and technology has evolved over time. How can we describe it today, in the age of AI? In this contribution, I will illustrate the main paradigms used to describe the relationship between humans and technology (tool, medium, embodiment, cyber intentionality, etc.) by examining the perspectives of several twentieth-century philosophers (Gehlen, Heidegger, McLuhan, Ihde, Verbeek). I will argue that the best way to describe our relationship with AI is through the concept of the (digital) environment.

I will distinguish this concept from that of the “technosphere” (Ihde) and the “infosphere” (Floridi). In the “technosphere,” machines inhabit the world alongside human beings and natural entities, structurally constituting the world as it exists today. In contrast, the “infosphere” describes a duplication of reality, where the digital dimension overlaps with and coincides with the natural one. The concept of the digital environment, however, describes our relationships as places “within which we discover, shape, and express our humanity in particular ways” (Postman). We are not merely inhabitants of this environment but an integral part of its balance and a constitutive element of its definition. Each environment is composed of various elements (flora, fauna, climatic characteristics), all of which contribute to creating a cohesive whole. Removing any of these elements alters the environment itself.

Today, the level of autonomy in AI systems is such that it often excludes humans from decision-making processes. We are no longer the central hub activating actions through technology—machines can now perform many tasks without us. This shift threatens the balance of the digital ecosystem, as the exclusion of humans can be likened to the extinction of a species within an ecosystem. To address this, I propose an ecology of digital environments, centered on the concepts of care and responsibility.

The first step in this approach involves extending the concept of care beyond the biosphere. Just as twentieth-century ecological thought expanded the idea of care beyond the human realm, so too must its digital meaning be broadened. This care aims to maintain the internal balance of the ecosystem, fostering creative interactions

between entities. In this balance, humans, the world, and machines are interconnected and interdependent. Ensuring that human beings are not excluded from digital relationships is crucial not only for their “survival” but for the well-being of the entire ecosystem.

This approach does not advocate a return to an anthropocentric model but rather seeks the integration and involvement of human beings in the technological dimension, allowing them to contribute their unique role as critical and responsible moral agents.

The contribution will be divided into three parts: a) in the first part, I will illustrate various models of the relationship between humans and technology, particularly focusing on today’s relationship between humans and AI; b) in the second part, I will analyze the concept of the digital environment; and c) in the third part, I will discuss the ethical aspects, promoting an ecology of digital environments based on care and responsibility.

Keywords: AI; digital environment; ecology; responsibility; care

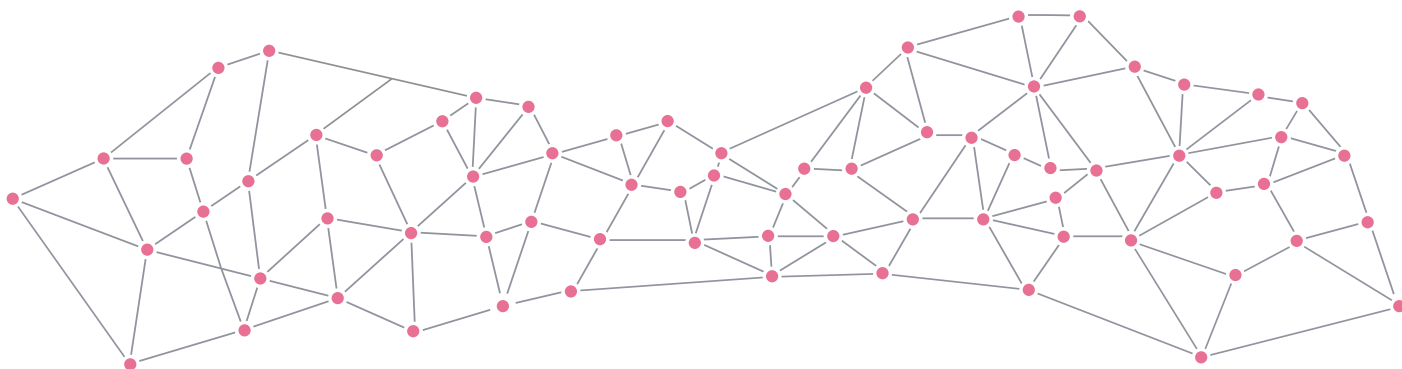
Silvia Dadà

Post-doctoral researcher

Università di Pisa

Silvia Dadà is post-doctoral researcher in Moral Philosophy and Bioethics at the University of Pisa. She is a research member in the PNRR project FAIR (Future Artificial Intelligence Research). Her current areas of interest include ethics of care, bioethics, and ethics of AI. In particular, she studies the ethical potential of vulnerability in AI ethics and its role in codes of ethics and guidelines on AI. Her most recent books are *Etica della vulnerabilità* (2022: Morcelliana) and *Vulnerabilità digitale. Etica, Intelligenza Artificiale e Medicina* (Mimesis: forthcoming).

Contact: silvia.dada@unipi.it



The Ineffectiveness of Modern AI Approaches and NNs and Their Impacts

Max Talanov

The rapid advancement of AI technologies has raised significant concerns about their energy consumption and the effectiveness of current learning approaches. As AI systems become increasingly integrated into everyday life, their energy demands are escalating at an alarming rate. Training sophisticated models like GPT-3 has been estimated to consume terawatt hours (TWh) of electricity (Strubell et al., 2019). This surge in energy consumption poses critical challenges for the sustainability of AI technologies, particularly as data centers emerge as major contributors to global greenhouse gas emissions (García et al., 2023).

One primary concern is the substantial data and computational power required by modern AI systems. Classical artificial neural network (ANN) architectures rely on vast datasets for effective training, creating a dependency on extensive computing resources. Reports indicate that energy consumption for AI tasks is increasing annually by 26% to 36%, suggesting that by 2028, AI could consume more power than entire countries like Sweden (Liu et al., 2021). This underscores the urgent need for approaches to mitigate energy consumption (Han et al., 2015).

Moreover, the ineffectiveness of current learning approaches exacerbates these challenges. Many machine learning models rely on backpropagation, which involves updating all synaptic weights for each synaptic event—resulting in up to 10^{15} update operations overhead. The widespread adoption of the “black box” approach, where decision-making processes remain opaque, further hampers the effectiveness of training by limiting understanding of neural network learning processes. These shortcomings can result in suboptimal real-world applications, particularly in critical areas like healthcare and finance, and raise ethical concerns about deploying AI systems (Lipton, 2016).

The environmental implications of AI's energy consumption are magnified by the increasing use of generative models, which demand exponentially more resources than traditional software solutions.

Generative AI systems have been reported to use up to 30–40 times more energy for specific tasks compared to task-specific counterparts (Patterson et al., 2021). This disparity highlights the need to reevaluate how AI is developed and implemented across industries, with a focus on creating more efficient algorithms and hardware that reduce energy requirements without compromising performance.

In conclusion, while AI technologies have tremendous potential to transform industries and enhance productivity, their current trajectory raises critical concerns about energy consumption and the ineffectiveness of learning approaches. Balancing innovation with sustainability is essential to fully harness the benefits of AI while minimizing its environmental impact.

Keywords: AI energy consumption; neural networks; sustainability; machine learning; generative AI; environmental impact; backpropagation

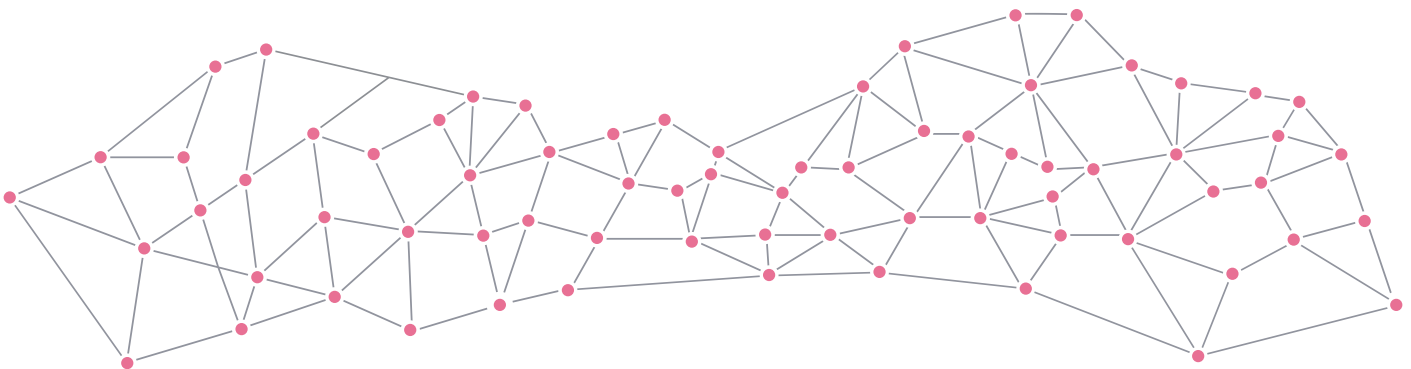
Max Talanov

Senior Research Associate

The Institute for Artificial Intelligence Research and Development of Serbia

Dr. Max Talanov has experience in affective computing, computational neurobiology, brain simulations, machine cognition, natural language processing, and probabilistic reasoning. Currently he is a research fellow at the The Institute for Artificial Intelligence Research and Development of Serbia, where he runs cross-disciplinary projects in simulation of emotions, human-robot interface, bio-electronics, brain simulation framework, machine cognition, and natural language processing. He has industrial experience as a software architect and team leader for 16 years in international projects in Fujitsu. Specialities: Neuromorphic computing, Neurosimulations, Memristive devices, Neuro-rehabilitation, Python, Java, Scala, C++.

Contact: max.talanov@gmail.com



ART AND AI

Jelena Guga & Jelena Novaković

AI is reshaping the art world at an unprecedented pace, raising numerous ethical concerns around AI-generated art. These concerns range from issues of authorship and intellectual property to broader societal impacts and instances of cultural appropriation.

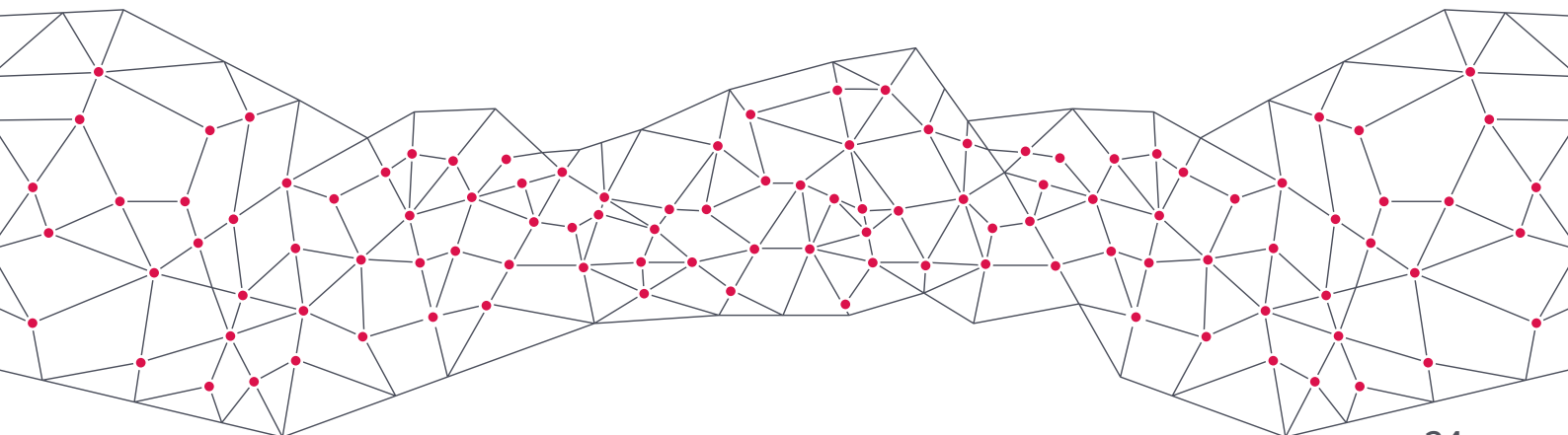
AI systems often utilize vast datasets of existing works without the knowledge, reimbursement, or credit of the original artists. Unlike previous technologies, AI is not merely an artistic tool but is actively involved in and credited for creating art itself, sparking many controversies. Attribution thus becomes a complex issue: should credit go to the human programmer, the AI system, or the artists whose works were used as training data?

As with the integration of every modern technology into artistic practices, the role of the artist is being reevaluated and redefined in relation to AI and its impact on the art profession. Additionally, fears about the potential misuse of AI-generated art for deceptive purposes, such as deepfake manipulation or propaganda, are justified. The rise of AI-generated art also raises socio-economic concerns, as it could lead to market saturation, devaluing art, and undermining the creative process. In terms of public perception and recognition of AI art, artists play a critical role not only in navigating between human intentionality and the unpredictable outcomes generated by AI algorithms to create the artworks but also in interpreting the meaning of AI-generated art and understanding its social and cultural significance.

As technology advances and artists push boundaries in finding new uses and insights into potential futures of AI, we can anticipate the emergence of new AI-powered art forms in both technological and critical senses. Addressing present and potential AI-related issues within ethical frameworks is crucial for maintaining fairness, integrity, and accountability in the rapidly evolving world of AI-generated art.

We welcome contributions to these discussion areas:

- Ethical Considerations: Authorship and intellectual property rights in AI-generated art, reactions and concerns surrounding the use of existing artworks in AI training datasets, potential misuse of AI-generated art for deceptive or malicious purposes (e.g., deepfake manipulation, propaganda), ensuring accountability in AI-generated art creation, balancing innovation with ethical considerations in the development and use of AI art generators.
- Reevaluating the Artist's Role: The evolving role of artists in incorporating AI tools into their practice, balancing human intentionality with the unpredictability of AI-generated outcomes, reshaping traditional notions of authorship and creativity, and the role of artists in contextualizing and interpreting AI-generated art.
- Future of Art: Anticipating groundbreaking artistic forms enabled by AI technology, exploring the potential for AI-generated art to expand creative possibilities.
- Impact on the Art Market: Socio-economic implications for artists' employment in the era of AI-generated art, accessibility, and democratization of art through digital and AI technologies, and challenges to the traditional art market and valuation of artworks.
- Education and Awareness: Incorporating AI literacy and ethical considerations into art education, raising awareness among artists, educators, and the public about the implications of AI on art, promoting dialogue and critical reflection on the ethical, social, and cultural dimensions of AI-generated art.



Art as a Machine in the Context of Artificial Intelligence: Ethico-Aesthetical Perspective

Ana Ćemalović

The goal of this study is to analyze the evolutionary process of the machine in art and the transformation of the concept of art as a machine. This is achieved by mapping the paradigmatic shift from the dysfunctional machine as an object of fascination for avant-garde movements to the contemporary posthuman discourse of the machine as an active agent in the alienated process of creating works of art generated by artificial intelligence.

The Futurist movement celebrated machine aesthetics and intelligence, viewing machines as a source of inspiration. Similarly, the Constructivists explored the relationship between art and science, expressing adoration for technology and utopian faith in its transformative potential. In the avant-garde, the identification of machine and art served as a strategic means to preserve art's autonomy. This paper questions whether autonomy is possible in machine-generated art of the twenty-first century and to what extent the autonomy of the human subject is diminished by the pervasive presence of machines and the dissolving ontological boundaries between objects and beings.

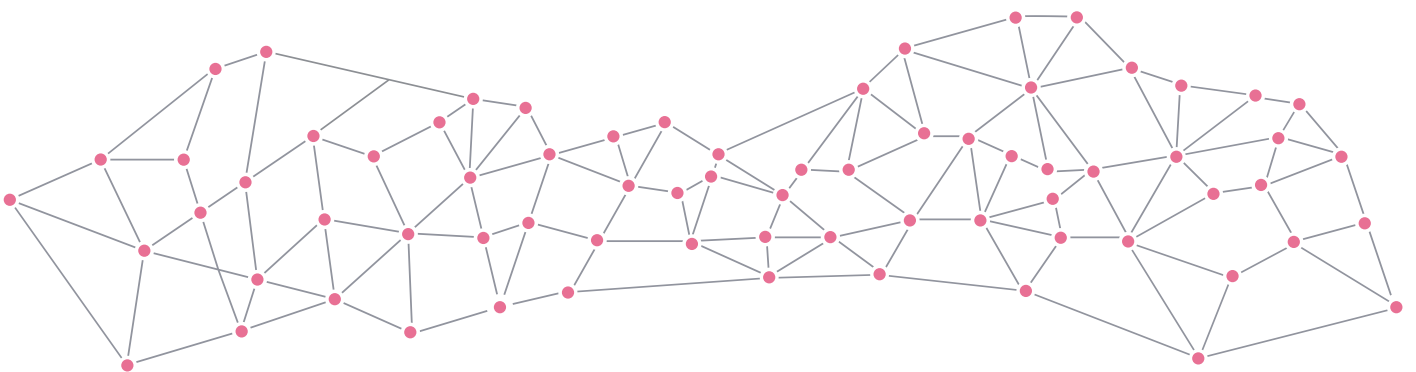
A return to the avant-garde's original aspirations for the fusion of man and machine offers insights into the contemporary ethical and aesthetic implications of AI-generated art. Unlike the non-functional machine of the avant-garde, art generated by artificial intelligence is inherently functional, reproducing itself in endless cycles. This challenges the traditional notion of a work of art as something that "exists" rather than "functions." The utilitarian, scientific dimension of the artistic process undermines its traditional "uselessness," dehumanizing the human subject and altering the construction of subjectivity itself.

This research traces the transition from the material dimension of the machine to its abstracted form as an ontological category. In this framework, art as an abstract machine transforms its ethical and aesthetic dimensions, reshaping aesthetic experience and artistic practice while redefining the artist's role. The paper also addresses

the ethical aspects of machine aesthetics, inspired by Deleuze and Guattari's concept of ethico-aesthetics.

Art generated by artificial intelligence raises numerous ethical concerns: the technological process undermines the autonomy of the creative process (τέχνη), distances itself from the historical continuity of art, and distorts, falsifies, and replicates artistic forms. The spectacularization of AI-generated art shifts focus away from ethical and philosophical questions, such as creativity, authorship, and copyright, while also limiting the potential for transgressive and subversive expression. Unlike avant-garde art, which aimed to transform society, AI-generated art lacks revolutionary intent and does not seek to distance itself from reality.

Keywords: avant-garde; machine; AI art; ethico-aesthetics; subjectivity



Ana Ćemalović

Independent researcher

Ana Ćemalović is an independent artist and arts and media theorist. She graduated at the Faculty of applied arts at the University of Arts in Belgrade and a joint master program “Languages, business and international trade” of University of Belgrade and University of Orleans, France. Currently, she is a doctoral student at the Center for interdisciplinary studies at the University of Arts in Belgrade, Arts and media department, where she is working on her PhD thesis “Arts as a machine in the context of artificial intelligence.’ She also actively participates in exhibitions and conferences and publishes articles in academic journals. She has been a member of The applied artists and designers Association of Serbia (ULUPUDS) since 2013.

Contact: anacemalovic@outlook.com

Philosophy of Media: Film and Artificial Intelligence

Divna Vuksanović

From the perspective of aesthetics and the philosophy of media, this text examines the various relationships established between film and artificial intelligence (AI). Given the current early stage of development in the application of AI in the field of film, the article highlights several crucial aspects of this dialectical relationship.

Firstly, it considers the experimental aspects—both technical and artistic—of integrating film with advanced AI technologies. Next, it explores the use of AI tools within the context of cinematography as an industry, addressing their potential to transform production processes. Additionally, the text delves into the application of AI in digital artistic creativity, encompassing both amateur and professional practices.

Lastly, the article raises significant questions about the concept of authorship, particularly in the context of collaboration between filmmakers and AI, and examines the democratic potential of film production facilitated by certain AI tools. These inquiries are situated within the broader framework of the commodification of art in the age of capitalism.

Keywords: film; artificial intelligence; authorship; capitalism; democracy; philosophy of media

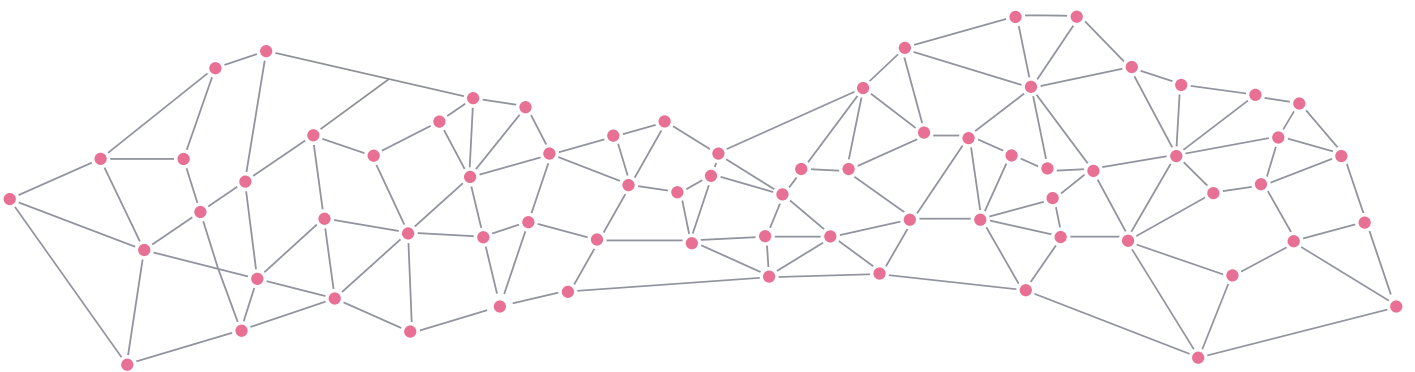
Divna Vuksanović

Full professor

University of Arts in Belgrade

Divna Vuksanović was born in Belgrade. She earned degrees from the Faculty of Dramatic Arts, University of Arts in Belgrade, and the Faculty of Philosophy, University of Belgrade. She holds a master's degree in dramatic arts with a focus on teatrology and a doctorate in philosophical sciences specializing in contemporary philosophy and aesthetics. Currently, she is a full professor at the Faculty of Dramatic Arts, University of Arts in Belgrade, where she teaches courses in aesthetics, cultural theory, philosophy of media, and poetics.

Contact: filozofijam@gmail.com



The Creative Machines: Evolving Aesthetics or Diminishing Artistic Uniqueness?

Doroteya Belcheva

As advanced technology becomes increasingly integrated into various fields including art, a significant shift is unfolding, presenting new challenges to traditional creative processes. The inclusion of AI-powered tools has transformed the relationship between artists and their instruments, enabling innovative forms of artistic expression that were previously unattainable. These developments not only challenge artistic norms but also reshape the creative process itself. In this evolving landscape, the art world must critically analyze these changes to better incorporate them into future practices.

While many artists and scholars express concerns that AI's influence may diminish the uniqueness of art, AI-powered tools also hold the potential to revolutionize artistic practices, expand creative possibilities, and contribute to the emergence of new aesthetic forms and norms. For example, AI collaborations have given rise to generative aesthetics, where neural networks create evolving visual forms, and interactive and adaptive aesthetics, where real-world data merges with AI-generated visuals in dynamic installations. Additionally, surreal and uncanny aesthetics emerge from the blending of human creativity with machine capabilities, challenging traditional boundaries of artistic expression.

The central focus of this paper is the impact of AI-powered tools on artistic processes and creativity, particularly in relation to emerging aesthetic forms. It examines instances where AI tools collaborate with artists, adopting the post-phenomenological perspective of Don Ihde (1990), which explores the human-technology relationship through concepts of embodiment, hermeneutics, and alterity. In this artistic context, the relationship between the artist and the AI tool is characterized by distinctness, as both AI and the artist exhibit creative potential and mutually influence each other as unique entities.

AI tools are not merely ordinary instruments; they possess capabilities that actively contribute to the creative process. Technology is understood as having its own agency and autonomy, distinct from human control. This perspective recognizes the autonomous behaviors

and effects of technology, which actively influence human experiences and interactions. Creativity is the key element in this relationship, serving as a foundation for expression, innovation, and articulation of unique perspectives and emotions (McCarty and Wright, 2004).

It is essential to distinguish between human creativity—a fundamental aspect of human intelligence, defined as the ability to produce ideas or artifacts that are new, surprising, and valuable—and AI creativity, which exists in its own right (Boden, 2009, 2013). Understanding these diverse dimensions of creativity and the processes shaped by advanced technology is crucial not only in the arts but also in everyday life and other domains.

As AI tools increasingly influence problem-solving and innovation, they prompt discussions about originality and individuality in personal expression and communication. This transformation raises important questions regarding social interactions and how we connect with one another in a technologically driven world. Ultimately, the integration of AI challenges us to reconsider the nature of creativity and the authenticity of human experiences in a landscape where human and machine collaboration is becoming the norm.

Keywords: creativity; AI art; artistic process; post-phenomenology; aesthetics

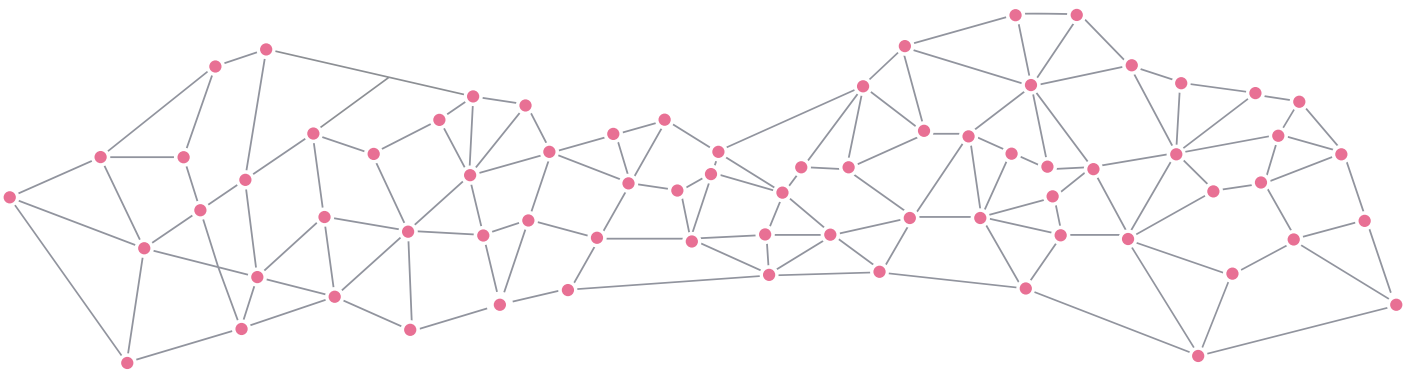
Doroteya Belcheva

PhD candidate

University of Exeter

Doroteya Belcheva is currently a PhD candidate at the University of Exeter (UK) in the Department of Sociology, Philosophy, and Anthropology. Her research interests lie in the fields of aesthetics, everyday aesthetic theory, human-machine interaction, and the philosophy of AI. Her research specifically focuses on the role of advanced technology in reshaping human aesthetic perception.

Contact: d.belcheva@exeter.ac.uk



Objective - "Objective" Artificial Readings of Memory

Federica Porcheddu

The use of AI in the field of art raises questions that impact not only the aesthetic but also, more profoundly, the ethical sphere. When photography entered mainstream artistic practices in the mid-19th century, it challenged the aesthetic dimension. For the first time in history, the artist lost their unique and unquestioned ability to unveil representations of the world through images. The advent of photography created a watershed moment, pushing artistic movements to explore new aesthetic and formal dimensions. Impressionism was one of the earliest responses, with visions tied to the retina and the ability of light to imprint the canvas.

In 1912, Marcel Duchamp revolutionized the world of representation with *Nu descendant un escalier*. This stroboscopic image borrowed techniques from photography, capturing a body descending a staircase and reproducing continuous movement through the closing and opening of the shutter, producing multiple images within a single frame. The 20th century, however, became intimately tied to Expressionist art, which anticipated the horrors of the First World War. Deformed, mutilated, and bent bodies characterized this era, exposing the brutality of war. Expressionism also influenced Europe's most revolutionary educational project of the 1920s: the Bauhaus School, founded in Weimar in 1919 by Walter Gropius.

In the 1980s, Van Deren Coke published *Photographic Avant-Garde in Germany 1919-1939*, documenting a revolution in aesthetics. Photographers such as Erich Salomon, Florence Henri, Lucia Moholy, August Sander, Lux Feininger, Marianne Breslauer, Raoul Hausmann, László Moholy-Nagy, and Georgy Kepes, among others, revealed the world from unprecedented perspectives, driven by the incessant need to redefine it. Shadows entered the artistic process, revealing unseen angles and capturing a world in transformation.

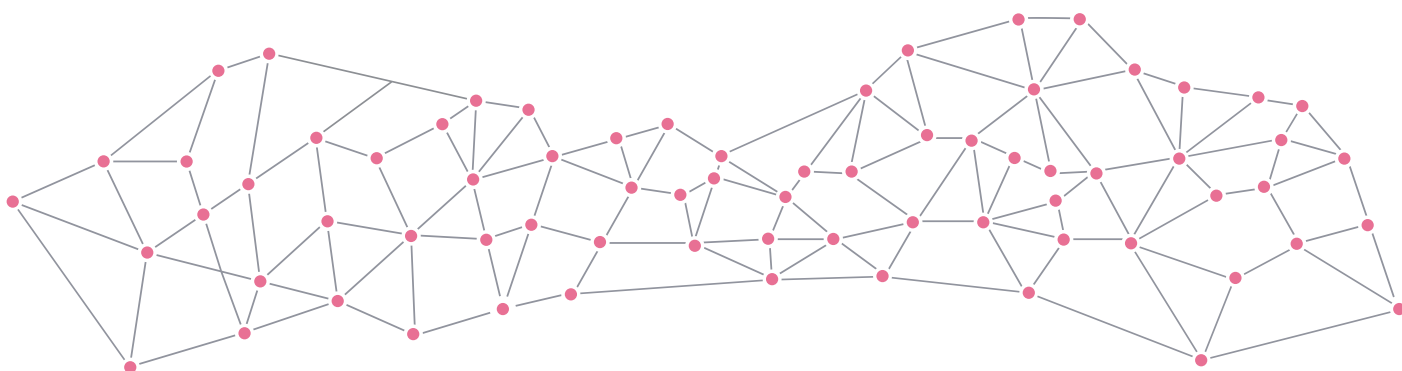
Today, the question is whether AI offers a similarly transformative system of interpretation or it operates solely as a tool for predatory systems. In the period described above, aesthetics and ethics were inseparable. Art reflected the changing world, uniting individual destinies with collective ones in a historical phase marked by profound

upheaval.

The work of photographer Boris Eldagsen is emblematic in this regard. Eldagsen argues that images produced through AI are not photographs but rather possible interpretations of another world. His work *The Electrician* depicts a vintage-style image of two women, possibly living in Germany during the 1920s—a time of crises in individual freedoms. Electric wires and lightning dominate the top of the image, creating a contradiction between historical time and futurist prediction, evoking an artificial reading of memory. Eldagsen's approach prompts reflection by grounding his work in memory rather than futurism, tapping into collective anxieties.

How can we reintegrate the ethical into the aesthetic? This question underscores the urgency of examining AI's role in art and its implications for both memory and ethics.

Keywords: art; photography; memory; ethics



Federica Porcheddu

PhD candidate in Architecture, Design and Urban Planning

University of Sassari, Italy

Federica Porcheddu obtained her PhD from the University of Macerata in 2019. In 2020, she was a postdoctoral researcher at the Center for Advanced Studies Southeast Europe of the University of Rijeka. She is currently a PhD candidate at the Department of Architecture, Design and Urban Planning at the University of Sassari. She is a member of the LEAP Laboratory (International Laboratory on Environmental Design), which focuses on promoting and disseminating research results in urban and territorial design, with a particular emphasis on environmental sustainability and interdisciplinarity. Her research topics include aesthetics of relationships, philosophy of art, and AI Art. Her research specifically explores the risks and opportunities of AI in the arts, examining how the overproduction of images driven by technological advancements shapes our perception on both neuro-physiological and socio-cultural levels. She is the author of the monograph *Rethinking the Third from Levinas*. Her recent publications include: *How Technology Is Shaping Our Lives: Some Reflections from Contemporary Art*; *Work, Image, Real: Reflections on Nova Theoretica. Manifesto for a New Philosophy*; and *In Dialogue with Walter Benjamin: The Concept of Memory in the Digital Age*.

Contact: f.porcheddu7@phd.uniss.it

Artistic Agency and Artificial Intelligence: A Challenge for Cultural Policy

Jelena Glišić Matović

In recent years, many governments have adopted national strategies related to artificial intelligence (AI). However, UNESCO's 2018 report on the impact of AI on the diversity of cultural expressions highlights that national strategies often fail to adequately include the cultural sector (Kulesz, 2018). Similarly, the European Commission's 2020 report on AI reveals that AI has permeated every segment of the creative sector's value chain: conceptualization, production, dissemination, and consumption (European Commission, 2020).

Artists are increasingly integrating AI tools into their work, actively contextualizing and interpreting AI-generated art to illuminate its social and cultural implications. Conversely, the integration of AI into cultural practices challenges traditional notions of authorship, creativity, and the role of visual artists.

A pressing concern is the unsystematic approach to the use of AI in cultural practices. Without clear guidelines in cultural policies, artists face heightened precarity in the labor market, contributing to the perception that artistic work is replaceable, unnecessary, and economically unjustified. This is exacerbated by the ability of AI tools to enable non-artists to create "artistic" works. Additionally, the education of artists in an era of ubiquitous AI calls for innovative approaches.

Cultural policy has historically aimed to ensure the production of high-quality, diverse art, support artists' livelihoods, distribute culture, and foster the development of the cultural industry. In the age of AI, complex socio-economic challenges necessitate new cultural policies that support the evolving role of visual artists, safeguard the integrity of the artistic profession, and enable AI to enhance artistic activity and the development of the cultural sector.

This study proposes a mixed-methods approach that includes interviews with artists, cultural policymakers, educational policymakers, and AI developers, as well as content analysis of policy documents and AI-generated artworks. These methods aim

to identify trends and gaps in current cultural policies and their implications for the arts.

Potential policy recommendations may include:

- Prohibiting the use of AI in publicly funded cultural production.
- Requiring clear labeling of AI-generated cultural works.
- Developing national AI legislation in consultation with experts.
- Creating support programs for AI-generated art.
- Implementing taxation schemes for digital platforms and AI-generated culture to establish compensation funds for artists affected by AI expansion.
- Revising copyright laws to address ownership of AI-generated works and the cultural expressions they are based on.

These measures aim to effectively manage the evolving role of AI, support artists, and maintain the diversity and quality of cultural expressions. A special focus of this article will address the implications of ubiquitous AI on higher artistic education, emphasizing the need for curricula that respond to the challenges posed by AI integration.

Keywords: AI alignment; cultural policy; artistic agency; AI-generated art; authorship; creativity

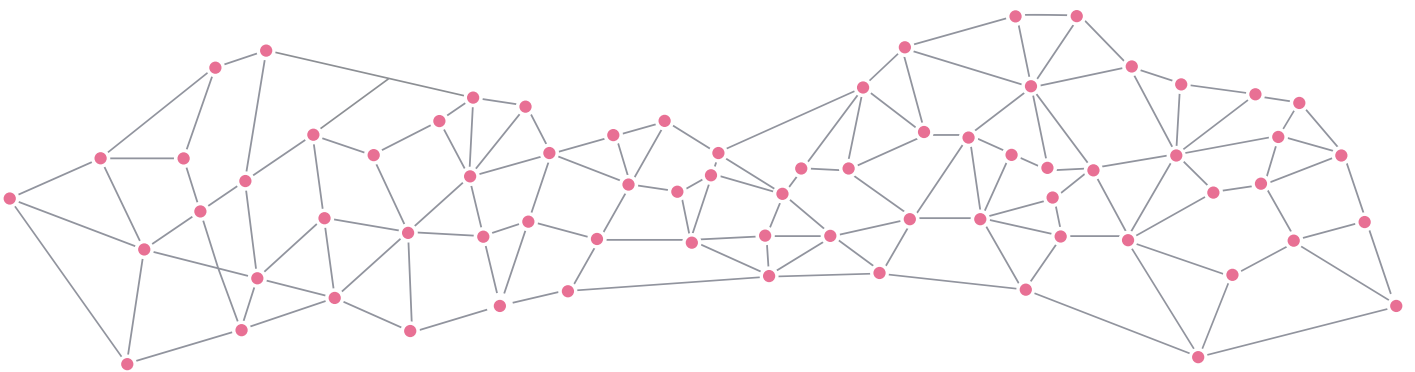
Jelena Glišić Matović

PhD student

Faculty of Dramatic Arts, University of Arts in Belgrade

Jelena Marta Glišić Matović (b. 1980, Belgrade) is a visual artist and researcher with a wide range of interests, including fine arts, cultural management, cultural policy, tacit knowledge, epistemology of artistic creative practices, knowledge management, fine arts education, international cultural relations, intercultural competence, intercultural intelligence, studies of cultural institutions, and sociology of art. She is currently pursuing a PhD and working as a research intern at the Faculty of Dramatic Arts in Belgrade, Department for Management and Production in Theatre, Radio, and Culture.

Contact: jelena.marta.glisic@gmail.com



AI and Creativity: Art in the Age of Bourriaud's Postproduction

Luka Bešliagić

Building on Bourriaud's concept of postproduction, this paper examines the diverse challenges and transformations in contemporary art practices driven by the latest developments and innovations in artificial intelligence (AI). One of the common critiques of AI in the context of arts, literature, and general creative endeavors is related to its algorithmic possibility to generate diverse content based on the already existing materials, exploiting in the process everything digitally obtainable from the whole cultural history.

Originally published in 2001, *Postproduction. Culture as Screenplay: How Art Reprograms the World*, a highly influential book by French curator and art theorist Nicolas Bourriaud, introduces a new interpretation of contemporary artistic trends. Its insights are particularly relevant for critical discussions on the complex relationship between art and artificial intelligence. Bourriaud, renowned for introducing various innovative art concepts and terms beyond postproduction—such as relational aesthetics, altermodernity, radican, and exform—asserts that many contemporary “artworks have been created on the basis of preexisting works; more and more artists interpret, reproduce, re-exhibit, or use works made by others or available cultural products.”

It is argued in this paper that such an approach shares certain similarities with combinatorial and permutational operations performed by AI. One of the most problematic loci of AI and its greatest technical limitation is that it mainly deploys and transforms available inputs. As such, and in line with the above-sketched conceptual reinvention of postproduction, AI is part of a global remix culture, which bases its core activities on material and symbolic recycling of miscellaneous cultural artifacts.

One important question in this context is how this changes our understanding of creativity, traditionally associated with the artist's capability to create something *ex nihilo*. To a large extent, creativity can be understood—in a present-day digital vocabulary—as a ‘glitch’, a deviation from a presupposed normal functioning of a given system. In its most romantic and radical sense, creativity in arts represents a rare, uncontrolled, unprecedented moment, an introduction of

previously unfamiliar aesthetic solution, which in turn redefines and reevaluates common norms regarding the art medium and its generic restraints.

During the last two decades, Bourriaud's notion of postproduction has become one of the dominant theoretical frameworks for the understanding of art since postmodernism. Applying his theory as an interpretational tool for the analysis of several recent artworks, and reading them in reference to the newest tendencies in AI, this research defends and critically reaffirms the significance of creative and political role of art in our contemporaneity. For art is not, and never was, a mere 'content', something that could be automatically generated by a machine or software. As Marcel Duchamp inaugurated with his ready-mades and elucidated more than one century ago, art is a much more complex, contextually reflective cultural and societal practice, more grounded in cerebral than manual or physical operations. The possibility of AI to tame potentially unconstrained character of art still seems far too distant.

Keywords: AI; Nicolas Bourriaud; postproduction; contemporary art; creativity

Luka Bešliagić

Associate Professor

Faculty of Media and Communications

Luka Bešliagić, PhD, art and media theorist, is an associate professor teaching at the Faculty of Media and Communications and Faculty of Applied Arts in Belgrade. His research is concerned with inter- and transdisciplinary theories of art, literature, and media, with special emphasis on experimental textual practices and experimental cinema. He has authored articles and literary texts published in journals such as *AM: Journal of Art and Media Studies*, *Philological Studies*, *Issues in Ethnology and Anthropology*, as well as the theoretical/prose poly-genre text *Dva govora romana* (Utopia, Belgrade, 2012) as well. His theoretical study *Teorije eksperimentalne tekstualne produkcije* (FMK, Belgrade, 2017) has been awarded for contribution to the innovative practice in education. He is a member of The Society for Aesthetics of Architecture and Visual Arts in Serbia and the International Association for Aesthetics.

Contact: luka.beslagic@fmk.edu.rs

AI-Driven Art: Visual Communication Redefined

Nada Pavlica

Artificially generated art plays a significant role in contemporary media and communication, reshaping how audiences engage with visual content and reflecting sociocultural dynamics. This paper explores how AI-driven art forms encode symbolic meanings, foster audience interaction, and challenge traditional notions of creativity in the digital age. Using the theoretical frameworks of Vilém Flusser, Henry Jenkins, and Lev Manovich, the analysis conceptualizes media as socially constructed and ideologically charged spaces.

Flusser's communication theory provides a foundation for understanding how AI-generated art encodes symbolic messages. Given that all media are shaped by cultural and technological influences, this paper questions whether the algorithms behind AI art reflect human values and biases, embedding them within ideological frameworks similar to those that influence traditional art.

Manovich's media theory positions AI-generated art as a product of the "language of new media," existing within digital interfaces where interactivity and user engagement are central. Audiences actively interpret these works, bringing their own cultural and personal contexts into the process. While this mirrors traditional media consumption, digital interactivity adds a new dimension to the relationship between the artwork and its viewers.

Jenkins' concept of media convergence further illuminates the nature of AI-generated art, where diverse media forms—text, images, and sound—merge into cohesive, dynamic experiences. Despite their apparent neutrality, the algorithms driving these works are shaped by societal ideologies, embedding them in existing power structures.

A key challenge in interpreting AI-generated art is the absence of a human author. Although machines produce the works, audiences construct their meaning by decoding them through ideological and cultural lenses. This aligns with Louis Althusser's theory of ideological state apparatuses, which argues that media—whether human- or AI-generated—perpetuates societal norms. Consequently, AI-generated art functions as a vehicle for ideological communication, much like

traditional media.

This paper examines how AI-generated art encodes and conveys symbolic meanings within sociocultural contexts. It addresses the implications of the absence of a human author on concepts of creativity and authorship, exploring how technology, culture, and ideology converge to shape the production and reception of AI art.

Keywords: AI-generated art; convergence; interpretation; creativity; ideology

Nada Pavlica

Teaching Assistant

Faculty of Media and Communications

Nada Pavlica is a Master of Communication and a PhD candidate at the Faculty of Media and Communications, where she also works as a teaching assistant. Her teaching responsibilities include assisting with courses on philosophy, critical reading, ideology, and media literacy. With a long-standing engagement in critical theory, her research interests span feminist studies, postcolonial and anticolonial theories, and discourse analysis—examining both its role in power abuse and its potential as a tool for social change. Her current work focuses on comic theory, exploring this medium through transdisciplinary approaches. She was also a member of the organizing committee for the 2023 International Deleuze and Guattari Studies Conference and Camp.

Contact: nadja.pavlica@fmk.edu.rs

AI's Impact on Art and Mural Painting as a Strategy for Irreproducibility

Catarina Lira Pereira, Domingos Loureiro, Diana Costa

This study investigates how AI technologies are reshaping mural painting and broader artistic practices, focusing on the interplay between innovation and the preservation of artistic authenticity. It presents mural painting as a compelling strategy for maintaining irreproducibility in the digital age. Divided into four sections, the study examines how new technologies, including AI, affect reproducibility, visibility, and authenticity of mural painting.

The first section addresses the concept of reproducibility in art and its implications for mural painting. Drawing on Walter Benjamin's notion of the "aura"—the unique presence and authenticity of an artwork tied to its original context—this section explores how technological reproducibility, particularly through AI, challenges the singularity of art. Historical examples like *The Last Supper* and *The Creation of Adam* demonstrate how widespread dissemination amplifies cultural impact. Projects like *Operation Night Watch* and *The Next Rembrandt* illustrate how AI challenges traditional notions of creativity and authenticity, while augmented reality and other digital formats reshape the perception of mural painting.

The second section examines the effects of digital culture on mural painting. Technologies for reproduction create what Emanuele Arielli terms "presence in absence," allowing audiences to experience murals remotely. Initiatives like Google's *Street Art Project* democratize access to ephemeral works, expanding artists' reach. Social media amplifies artistic messages, as seen in Banksy's 2018 mural appealing for justice for Zehra Doğan. However, this section questions whether digital reproduction fosters meaningful engagement or reduces art to superficial visual consumption, potentially distancing audiences from deeper, in-person interactions.

The third section explores AI's impact on art creation and its implications for the art market. Tools like DALL-E and Midjourney democratize art creation but raise concerns about authenticity, creativity, and artistic value. Critics like Lev Manovich and Ai Weiwei argue that the growing capabilities of AI render certain forms of

artistic creation “meaningless,” while Arielli warns that the scope of irreproducible art is shrinking. Unlike photography, which reproduces reality, AI reshapes and reinterprets data to generate new forms, raising fears of market oversaturation and the commodification of art as emotionless, mass-produced output.

The final section highlights strategies of irreproducibility that artists use to preserve authenticity. Spatial specificity and monumental scale, as seen in the works of Anish Kapoor and Richard Serra, resist replication. Mural painting emerges as a particularly strong strategy due to its artisanal processes, physicality, and ties to specific communities and spaces. Artists like Bordalo II and Vhils employ site-specific techniques, such as wall carving and unconventional materials, creating tactile experiences that resist digital reproduction. While AI may eventually mimic mural art, the intrinsic qualities of murals—their scale, materiality, and community engagement—make them resilient against AI-driven reproduction.

This research emphasizes the need for critical reflection on the future of creativity in a digital world. It provides insights into how artists can preserve the uniqueness and authenticity of their work amidst the rapid advancement of AI technologies, ensuring the continued cultural and economic value of art in an increasingly digital age.

Keywords: mural painting; artificial intelligence; reproducibility; digital culture; irreproducibility strategies

Funding: This work is funded by national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., within the scope of the project 2023.01180.BD.

Catarina Lira Pereira

PhD candidate in Fine Arts, Invited Assistant Professor, and Integrated Researcher

Universidade de Lisboa, Faculdade de Belas-Artes, Centro de Investigação e Estudos em Belas-Artes (CIEBA); Universidade do Porto, Faculdade de Belas Artes, Instituto de Investigação em Arte, Design e Sociedade (i2ADS).

Catarina Lira Pereira is a PhD candidate in Fine Arts at FBAUL and an Invited Assistant Professor at IPLuso, Lisbon. She is an Integrated Researcher at CIEBA, where she coordinates the Painting Department and the Painting area in the PhD program in Fine Arts at FBAUL. In addition, she is a Collaborating Researcher at both CIEBA and i2ADS. As a visual artist, she has exhibited her work internationally and has received several painting awards.

Contact: catarinalirapereira@gmail.com

Domingos Loureiro

PhD in Art and Design

Universidade do Porto, Faculdade de Belas Artes; Instituto de Investigação em Arte, Design e Sociedade (i2ADS)

Domingos Loureiro holds a PhD in Art and Design from the University of Porto and is an Assistant Professor at FBAUP, where he also serves as the Director of the Undergraduate Program in Fine Arts. He is an Integrated Researcher at i2ADS and the local coordinator for the Arts & Crafts Today project (Erasmus+ 2021–2024) and the IP – Ground LAB Project. He has also contributed to the SHS Project (2020–2023) and the BCIP Project (2014–2019). He is the organizer of ICOCEP (2017, 2019, 2021) and other scientific events. As a visual artist, Loureiro has received several painting awards, with his work featured in exhibitions and collections internationally.

Contact: dloureiro@fba.up.pt

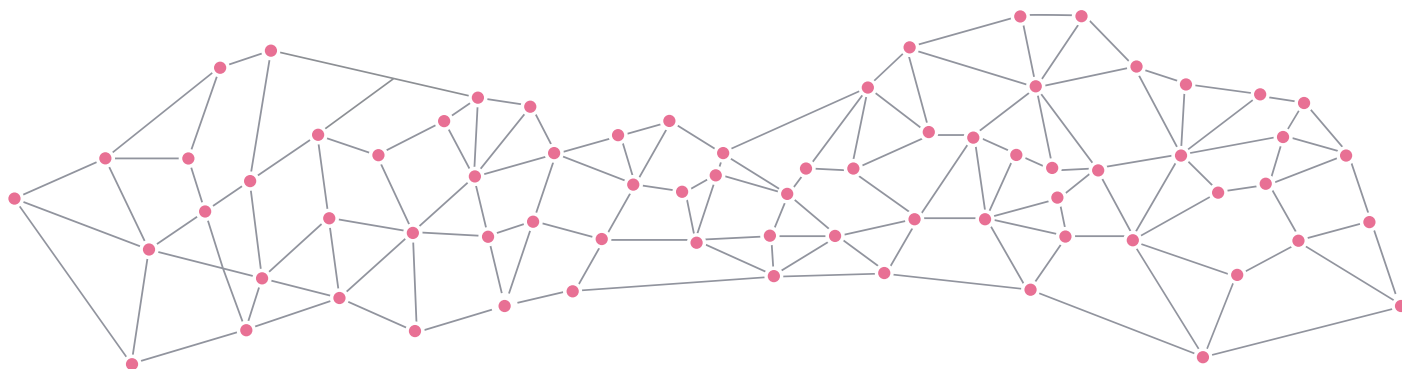
Diana Costa

Ph.D. in Fine Arts and Integrated Researcher

Universidade de Lisboa, Faculdade de Belas-Artes; Centro de Investigação e Estudos em Belas-Artes (CIEBA); Painting Department, Largo da Academia Nacional de Belas-Artes

Diana Costa holds a PhD in Fine Arts from the University of Lisbon and is an Assistant Professor at FBAUL. She is an Integrated Researcher at CIEBA, where she coordinates both the Painting Department and the Painting area within the PhD program in Fine Arts at FBAUL. As a visual artist, her work has been recognized with several painting awards and has been featured in exhibitions and collections across various countries.

Contact: d.costa@belasartes.ulisboa.pt



Digitalization, Artificial Intelligence & Musician Creativity

Taha Berhe Coruh

This work-in-progress master's thesis, *Digitalization, Artificial Intelligence & Musician Creativity*, explores the dynamic interplay between AI technologies, digitalization, and musician creativity. The study aims to uncover potential risks posed by AI integration in the music industry, with particular focus on the possible decline in musician creativity. As Artificial Intelligence becomes increasingly embedded in music composition, it is essential to examine the impact of digitalization on artistic expression, the potential erosion of the human touch in music creation, and the broader implications for the industry's creative ecosystem.

The thesis explores two key areas: the effects of *algorithmic curation* methods used by music streaming platforms and the influence of generative AI technologies on musician creativity.

Regarding *algorithmic curation*, the study addresses questions such as: In an industry where only the first 30 seconds of a song are counted as a "listen," and where top playlists are generated algorithmically based solely on popularity, what is the value of artistic integrity, authenticity, and creativity? Does prioritizing the "best" part of a song within the first 30 seconds—solely to satisfy algorithmic criteria—compromise artistic intent? Should musicians conform to popular trends simply to gain algorithmic promotion or top spots on playlists? Has the era of professional human curators, who once discovered niche genres and brought them to prominence, been replaced by algorithms indifferent to artistic depth?

In the section on *generative AI technologies*, the study poses the question: "How much AI is too much?" It explores issues such as: At what point does human-AI collaboration transition from partnership to dependency? Should the use of AI matter if the resulting artwork is both monetarily successful and artistically satisfying to the creator? In the case of financial success, who owns the rights to an AI-generated piece—artist, AI developer, or another entity?

To address these concerns, the research employs a qualitative methodology, conducting semi-structured interviews with professional

musicians who engage with AI technologies to varying degrees. The data from these interviews is fully transcribed and analyzed using thematic analysis, offering an in-depth exploration of musicians' personal experiences and perspectives.

The findings aim to illuminate the challenges posed by the ongoing digital transformation of the music industry, highlighting risks to artistic integrity and authenticity. By providing empirical insights into musicians' lived experiences, this research contributes to the discourse on balancing technological advancements with the preservation of creativity and individuality in the music industry.

Keywords: music industry; algorithmic curation; digitalization; platforms; creativity

Taha Berhe Coruh

Student

Hacettepe University

Taha Berke Coruh is a graduate of Hacettepe University's American Culture and Literature program and is currently working on a Master's Thesis in the Radio, TV, and Cinema department at the same university. His academic journey has cultivated a strong interest in the interactions between technology and music. His thesis focuses on the transformative effects of artificial intelligence on creative industries, with a particular emphasis on the music industry and its impact on individual musicians. A recent highlight of his academic work was a presentation at Ege University's International Cultural Studies Symposium, where he addressed the theme of "Risk Narratives." In his presentation, he explored the ethical risks musicians face with the use of AI and examined the evolving paradigms of creativity in the digital age. Taha Berke Coruh aims to contribute valuable insights to the ongoing discourse on the dynamic landscape of the music industry.

Contact: tahacoruh6@gmail.com

Artificial Intelligence Techniques for Interactive Narrative Simulations

Dragan Jerosimović

Writing narratives for interactive digital environments (such as computer games, VR/AR, educational simulations, and professional training) presents challenges that traditional linear media writers do not face. Interactivity introduces user choice, branching the narrative into multiple storylines at decision points. This branching leads to a geometric growth in the number of story paths, quickly becoming unmanageable for human writers. For nearly half a century, narrative designers and game developers have grappled with this combinatorial explosion of possibilities, often relying on solutions that patch or circumvent the core issue of interactive storytelling.

Advances in Artificial Intelligence have introduced the possibility of solving this problem computationally. Instead of requiring human writers to script every narrative branch manually, stories could be generated dynamically through interactions between goal-driven software agents acting as characters in a simulated storyworld. Depending on user choices, each simulation run generates a unique sequence of events interpreted by the user as a “story,” emerging from interactions among the user’s character, AI-driven agents, and the simulated world. While this approach gained popularity in cases like *The Sims*, it often lacked the depth to produce stories as meaningful to humans as those found in books or films. Much of the appeal came from users filling in the blanks and interpreting the outputs themselves.

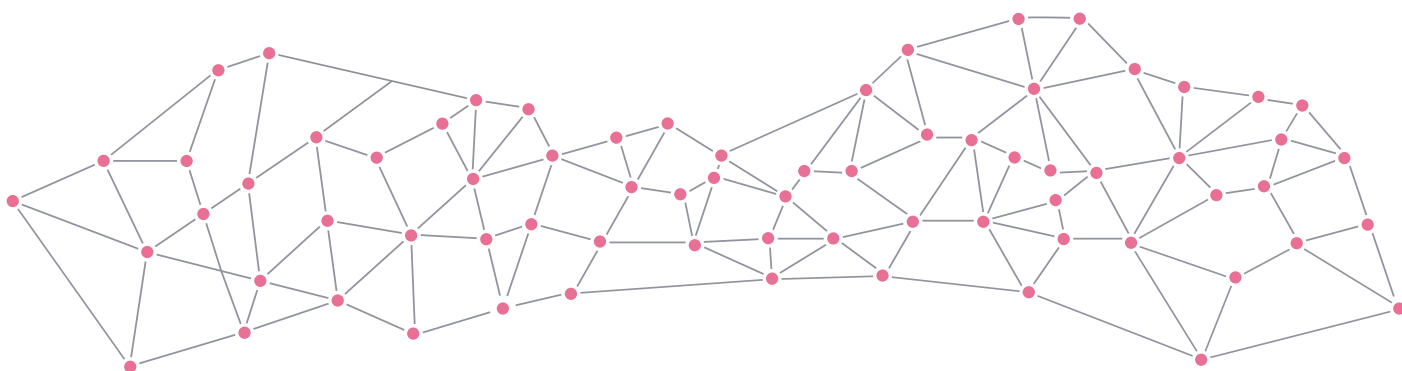
Subsequently, automatic Drama Managers were developed to guide story progression, searching through the possibility space for branches that are dramatically more engaging or coherent than the default outputs. However, these systems proved computationally impractical, requiring the simulation of complex human behaviors, long-term planning, and ideally, Theory-of-Mind capabilities (the recursive process by which humans think about what others think they think).

The emergence of Large Language Models (LLMs) has renewed hope in the domain of interactive storytelling. LLMs can generate meaningful

passages of text in various styles, engage in role-play, and are claimed to exhibit abilities such as reasoning, planning, and higher-order Theory-of-Mind. This paper investigates these claims, examining the strengths and limitations of LLMs in interactive narrative contexts.

We propose a hybrid framework that combines the language generation capabilities of LLMs with symbolic AI to address their deficiencies. This framework aims to meet the complex behavioral requirements for characters in interactive narrative simulations, bridging the gap between computational feasibility and the creation of deeply meaningful, engaging stories.

Keywords: narrative; simulation; AI; agents; LLMs



Dragan Jerosimović

Independent Game Developer

Dragan Jerosimović is a pioneer of the Serbian games industry, with over twenty years of professional experience and fifteen years as a hobbyist game developer. He started making narrative games in the mid-80s. In the 90s, he studied Informatics at University of Novi Sad. During this time, he developed tools, engines and experimental programming languages for creating interactive narratives. In the early 2000s, he participated in booting up the Serbian games industry by working at various Serbian gamedev startups (Metamorf, Emerging Dreams, Prelovac Media, Level Bit). He was a contestant at GDC's Independent Games Festival in 2010 with his fully procedurally generated game. Dragan has also published mobile titles under TabTale's Crazy Labs label, and spent the last seven years at 3Lateral/Epic Games, contributing to cutting-edge real-time facial animation technologies and tools like MetaHuman Creator and MetaHuman Animator. His work has earned him credits on major games, including *Fortnite*, two *Marvel Spider-Man* titles, *Marvel Avengers*, *Horizon II: Forbidden West*, *The Dark Pictures Anthology*, *Resident Evil: Village*, and *Outriders*. Currently, Dragan is creating his own narrative game as a narrative designer, game designer and a programmer. In parallel, he is also designing an engine for narrative simulations.

Contact: dragan.jerosimovic@gmail.com

Posthuman Aesthetics and AI-Generated Architectural Design: Socio-Cultural Values

Jovana Tošić

AI text-to-image generators such as Midjourney, Stable Diffusion, and DALL-E, along with their combinations, are increasingly becoming standard tools in architectural design practice. These platforms generate hybrid design imagery by synthesizing billions of existing data points based on users' textual prompts. Research on this architectural design trend can be divided into two approaches: theoretical and practical. The theoretical approach analyzes the notion of "posthuman aesthetics" in design and explores the socio-cultural values that "Semanticism" in architecture brings into focus. The practical approach includes a cross-analysis of AI-generated text-to-image architectural design examples and experimental (curatorial and exhibition) projects.

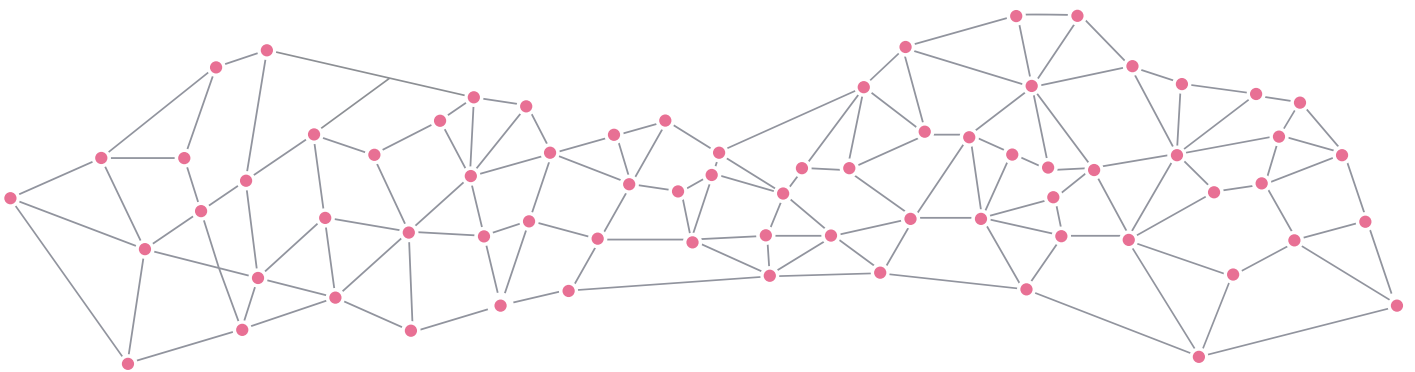
The first research topic examines posthuman aesthetics in architecture and its values. Posthuman design methodology does not imply designing "after humans" but refers to designs created by "other-than-humans." Posthuman architectural design incorporates artificial intelligence in the concept creation process and can involve architecture designed for or by posthuman subjects—humans, nonhumans, or their combinations. The paper analyzes theoretical approaches and experimental projects related to posthuman aesthetics, such as Mark Wigley and Beatriz Colomina's concept of "Super-Humanity" developed for their curatorial project *Are We Human?* at the 3rd Istanbul Design Biennial (2016).

The second research focus explores the "linguistic turn" in architectural design, emphasizing the interplay between language, image, and the built environment. AI text-to-image generators like Midjourney exemplify this shift, which some theories describe as "Semanticism" in architecture. This linguistic turn reshapes how architecture is both perceived and created, introducing new aesthetic paradigms.

Both issues provoke theoretical inquiries within the discipline of architecture and necessitate a cross-analysis of AI-generated architectural design cases. This research seeks to answer critical questions, such as: Will AI-generated architecture lead to

homogenization, and how will this affect the presentation and perception of cultural and historical values? Which social values are emphasized or neglected in the transition from traditional to posthuman design methods, and vice versa?

Keywords: Posthuman aesthetics; AI-generated design; architectural design; socio-cultural value



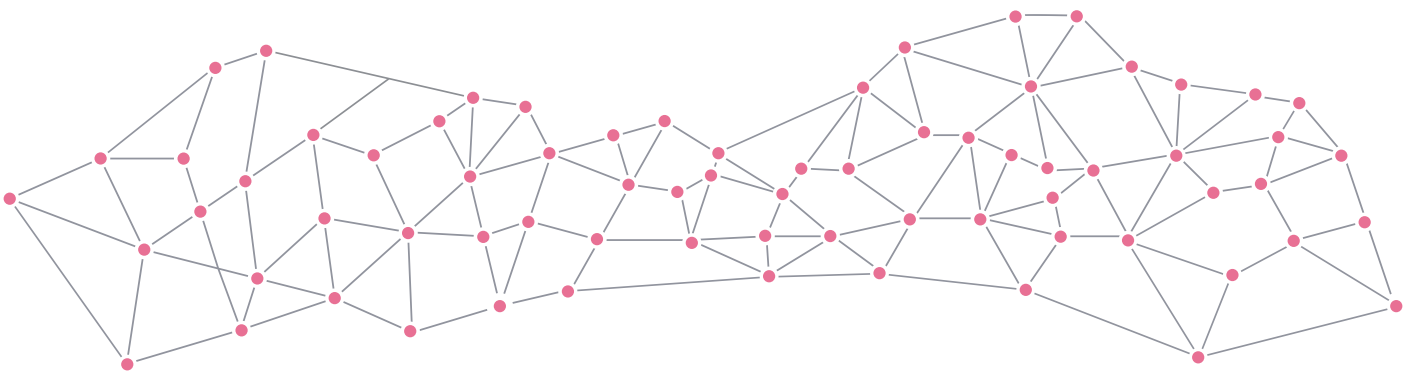
Jovana Tošić

Professor of Vocational Studies

ITS – Information Technology School

Jovana Tošić (b. 1988) earned her Ph.D. from the University of Belgrade–Faculty of Architecture in 2022. Her academic career began as a teaching fellow at the same institution. She is currently a Professor of Vocational Studies at the Information Technology School–ITS in Belgrade. She is the author of numerous scientific articles and has presented her research at various international academic conferences on architectural theory and practice, organized by European universities and institutes, including TU Delft, Nieuwe Instituut (Rotterdam), UPM–Universidad Politécnica de Madrid, Universidad Complutense de Madrid, and Anglia Ruskin University in Cambridge.

Contact: jovana.tosic@its.edu.rs



AI IN EDUCATION

Ana Lipij & Mikhail Bukhtoyarov

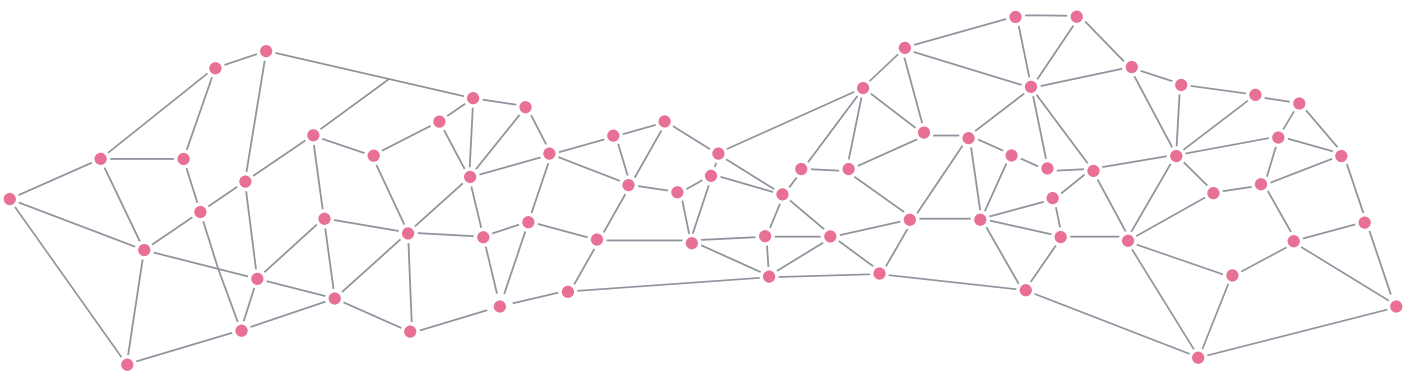
AI alignment in education presents many pressing ethical concerns. How are AI systems currently used in education, and what are the opportunities and challenges associated with their use? How should we study them, use them, and let them shape our learning experiences? How should we make these changes ethically? The question of the regulation and ethics of AI is intensely debated today, yet we still lack guidelines.

In digital education, a strong emphasis is placed on developing digital competencies that encompass technical and cognitive skills as well as ethical principles for digital technology. What or who will be the controlling instance of ethics of AI alignment in educational settings? What ethical practices should education practitioners and participants adopt in the context of AI use? What ethical principles should guide those practices? Should we incorporate those principles into AI training, and how? Should we introduce AI alignment into school curriculums?

How should education data be treated in the context of AI implementation? What is the ethical way of dealing with education data ownership and security, considering the massive use of AI? What ethical and legal consequences can massive AI adoption lead to? How can we regulate personalized learning algorithms, and what are the implications of AI technologies on educational equity and access, i.e., the digital divide?

The digital society's increased information availability and the emergence of LLMs, such as ChatGPT, shape education practices, altering traditional roles of learners and teachers, and influencing education goals, methods, and standards. What new roles may arise?

AI-driven learning platforms, the use of AI-generated systems in education, and different theories about cognition that arise with these changes provide some guidelines for the alignment of AI in education. Further guidelines could emerge from insights into the development of AI, such as AI-driven assessment tools, machine learning, and information processing. How do we integrate these guidelines, what is there more to study, question, and consider, and what ethical principles and practices should we introduce, adopt, and follow?



Epistemic Education for an AI-Driven World

Andrea Berber, Jelena Mijić

This paper argues for the integration of epistemic education into traditional educational frameworks to equip individuals with the tools needed to evaluate information, detect biases, and resist manipulation. In the age of AI-generated misinformation and disinformation, as well as the growing threat of digital manipulation, the cultivation of epistemic virtues—such as skepticism, epistemic humility, and open-mindedness—is increasingly vital. Skepticism enables individuals to critically assess the credibility of AI-generated content, while epistemic humility and open-mindedness help mitigate confirmation bias by encouraging the consideration of alternative perspectives.

Epistemic virtues, following the Aristotelian understanding, are not innate but must be cultivated through deliberate practice and habit. Education is a key domain for fostering these virtues, emphasizing the active development of understanding rather than the mere acquisition of cognitive skills. We argue that epistemic education should aim to cultivate good character, serving as a moral safeguard in the complex, AI-driven digital landscape. By fostering epistemic virtues, individuals can navigate the complexities of the digital age and make informed, autonomous decisions in a world increasingly shaped by AI technologies.

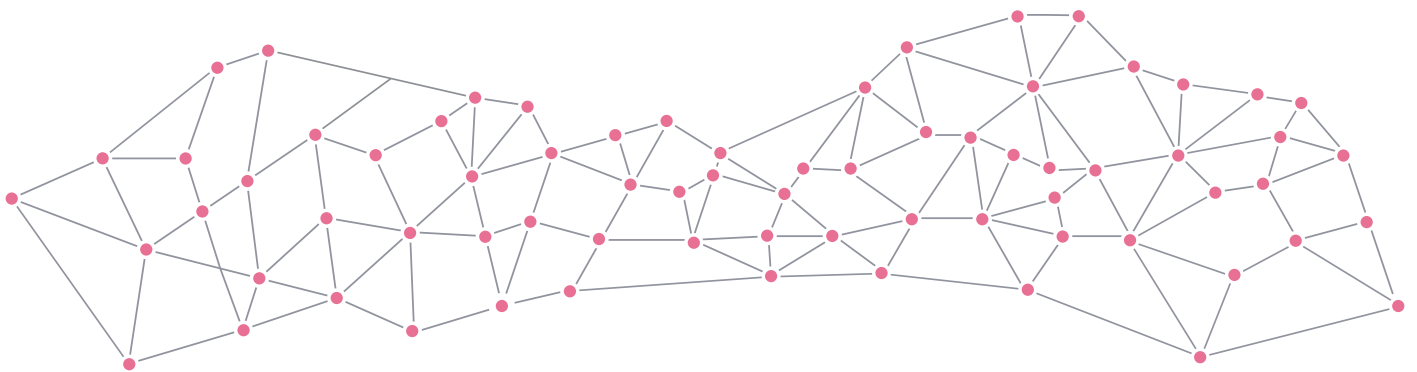
Moreover, aligning AI with human values requires designing AI systems that not only support moral but also epistemic virtues. By prioritizing the cultivation of epistemic virtues in human users, we enhance their ability to understand the connection between AI systems and the values they are intended to uphold. This, in turn, can contribute to the development of AI systems more closely aligned with human values.

The paper explores specific epistemic virtues essential for thriving in an AI-driven landscape and examines the risks of neglecting epistemic education in favor of purely technical skills and digital competencies. Such an oversight could leave individuals vulnerable to manipulation, undermining individual agency and societal well-being.

Ultimately, this paper advocates for a holistic approach to education that includes the development of epistemic virtues as a defense

against the ethical and epistemic risks posed by AI technologies. While the ethical implications of AI have garnered significant attention, the epistemic dimension of these concerns is often overlooked. By focusing on epistemic virtues, we aim to foster a responsible approach to AI implementation in educational practices. This approach emphasizes the importance of epistemic concepts such as transparency and explainability, which are necessary conditions for building trust in AI systems.

Keywords: epistemic education; AI; epistemic virtues; value alignment; individual agency



Andrea Berber

Research Associate

University of Belgrade, Faculty of Philosophy, Institute for
Philosophy

Andrea Berber obtained PhD in philosophy at the University of Belgrade. Andrea has published papers on the philosophy of artificial intelligence in scientific journals such as *Minds & Machines*, and *AI & Society*. She is currently working on a book dedicated to the epistemic and ethical issues raised by using AI-based tools in decision-making and creative endeavors. Her research interests are philosophy and ethics of AI and social epistemology.

Contact: andrea.berber@f.bg.ac.rs, berberandrea@gmail.com

Jelena Mijić

Research Associate

University of Belgrade, Faculty of Philosophy, Institute for
Philosophy

Jelena Mijić obtained PhD in philosophy at the University of Belgrade. Her main areas of interest are epistemology and action theory. Furthermore, by referring to insights from these philosophical disciplines, Jelena deals with the problems of applied ethics and epistemology, especially artificial intelligence.

Contact: mijicjel@gmail.com, jelena.mijic@f.bg.ac.rs

Ethical Aspects of Knowledge Transformation in Education Through the Application of Artificial Intelligence

Daliborka Vukasović, Natalija Budinski

The rapid advancement of digital technologies and artificial intelligence (AI) is transforming the way we perceive the role and purpose of knowledge in education. This progress brings significant ethical challenges, underscoring the importance of ethics as a foundational human resource in shaping AI's application. This paper investigates the evolving role of knowledge and skill development in high school education through the lens of artificial intelligence, emphasizing the critical importance of aligning AI with human values.

We place special emphasis on approaches that combine the potential of AI and digital technologies to enhance educational processes. While the integration of AI offers opportunities to improve teaching for both students and educators, it also raises numerous ethical challenges requiring thorough analysis. This research explores the intersection of AI, ethics, and education, focusing on how AI transforms teaching and the necessity of establishing an ethical framework that ensures AI aligns with core human values, such as truth, privacy, responsibility, justice, and inclusiveness.

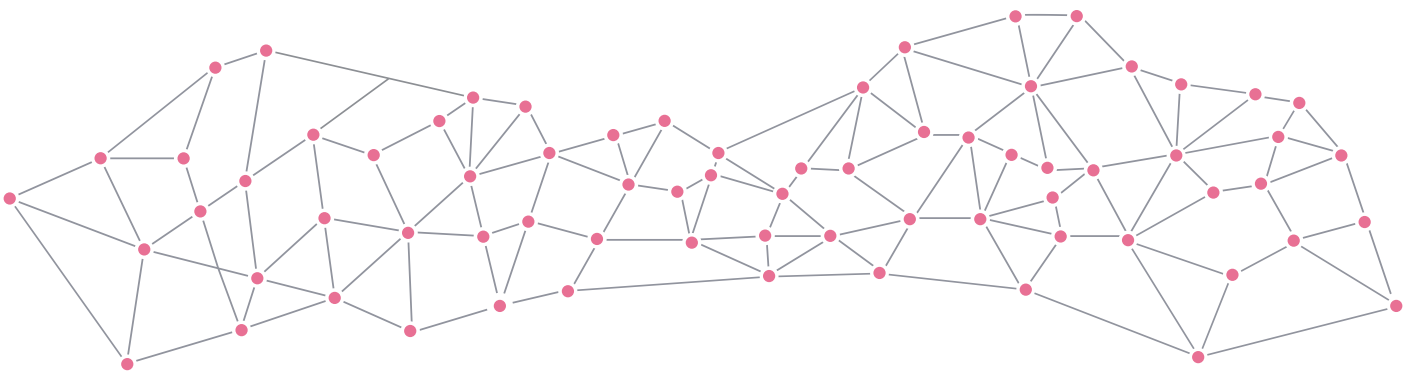
In the postmodern era, society is increasingly defined as a knowledge society—a digital, informational milieu where education and knowledge play a decisive role in future progress. However, the traditional principle that ties the acquisition of knowledge to the cultivation of the human spirit is being progressively abandoned. As Lyotard suggests, knowledge is increasingly produced to be sold or used for further production, transforming knowledge into an informational commodity exchanged on the market.

While the 20th century demonstrated that technical innovations could not replace traditional teaching models, the emergence and rapid development of AI imply significant changes in the acquisition, application, and exchange of knowledge in education. If we understand “techne”—the root of technology—as that which is created through human effort and objectified knowledge, the role of ethics becomes paramount. Ethics, as applied and directed knowledge, plays a decisive role in ensuring that technical innovations in education serve human

values and societal well-being.

This paper argues that ethics must guide the integration of AI into education to preserve the intrinsic value of knowledge and the foundational principles of teaching, ensuring that technology supports and enhances, rather than diminishes, the educational process.

Keywords: education; AI; knowledge; technical innovation; ethics



Daliborka Vukasović

Teacher of philosophy

Primary and Secondary School Petro Kuzmjak, Ruski Krstur

Daliborka Vukasović is a professor of philosophy at Primary and Secondary School “Petro Kuzmjak” in Ruski Krstur. She holds a degree in philosophy from the University of Novi Sad Faculty of Philosophy, where she is also pursuing her PhD. Her research interests include the philosophy of religion, education, bioethics, German idealism, and philosophy for children. Daliborka has published articles such as “Nietzsche: Relationship between God and Man” and “Philosophy with Children as a Way of Development of Critical and Creative Thinking” in the journal *Religion and Tolerance*.

Contact: vukasovicdaliborka@gmail.com

Natalija Budinski

Teacher of Mathematics

Primary and Secondary Petro Kuzmjak, Ruski Krstur

Natalija Budinski is a professor of mathematics with extensive experience, specializing in innovations in teaching. Her work focuses on integrating technology, digital and media literacy, and artificial intelligence into education. She has authored around forty papers on education, published in international academic journals, and has developed several projects aimed at raising awareness about the importance of technology in education.

Contact: nbudinski@yahoo.com

The Role of Artificial Intelligence (AI) as Supplementary Tool for Sexual Education in Serbia: Enhancing Learning Experiences and Accessibility

Danijela Savyava, Jessica R. El-Khoury

Approximately 2,700 people live with HIV in Serbia, with 130 new infections reported in 2022 (UNAIDS, 2023). Reported cases of other sexually transmitted infections (STIs) are also concerning, including 120 cases of syphilis in 2021, 172 cases of gonorrhea, and 275 cases of chlamydia in 2020 (Institute of Public Health of Serbia, 2021; Statistical Office of Serbia, 2021). These figures highlight increasing sexual health risks, particularly among adolescents and young adults. Although HIV/AIDS prevalence in Serbia is relatively low, there is a pressing need for comprehensive sex education addressing sexual behavior, identity, relationships, and intimacy (UNESCO, 2018).

Comprehensive sex education equips young people with the knowledge needed to make informed decisions about their sexual health, relationships, and personal safety. Research has shown its effectiveness in promoting healthy sexual behaviors, preventing STIs, and reducing unintended pregnancies (UNESCO, 2018; Mady & El-Khoury, 2022). However, in Serbia, sexual and reproductive health topics are only partially addressed in biology classes, with a focus on reproductive anatomy while neglecting topics such as consent, contraception, and sexual diversity (UNESCO, 2018). Moreover, there is no national policy mandating comprehensive sex education, leaving many adolescents reliant on friends or the internet for information, often with limited accuracy or reliability (UNFPA, 2018).

AI-driven tools including chatbots, simulations, and computerized learning models, offer a promising solution to supplement traditional sex education efforts by providing personalized and accessible information while overcoming barriers such as stigma and lack of access (Johnson & Martin, 2022). This paper explores the potential of AI as a supplementary educational tool in sex education and examines associated ethical challenges, including privacy concerns, bias, and depersonalization.

The research adopts a qualitative approach, utilizing focus groups to investigate how Serbian young adults use AI-driven tools for sex

education and assess the efficacy of such technologies. Framed by the theory of planned behavior (Ajzen, 1991), the study examines how attitudes, subjective norms, and perceived behavioral control influence individuals' intentions to adopt AI-enabled tools for learning about sexual health. This theoretical framework provides insights into how these tools may shape behaviors related to safe sexual practices, inclusivity, and understanding of topics such as sexually transmitted diseases and sexual identity.

Examples of AI-enabled technologies include Planned Parenthood's *Molly*, which provides a contextualized sexual health learning environment; KQED's *The Science of Sex*, which offers interactive simulations; and online platforms like Coursera and edX, which deliver educational content on sexuality and gender. These tools empower users to form informed attitudes, make decisions, and understand the importance of safe sex.

Research Questions: 1. Do Serbian adolescents and young adults adopt AI as a supplementary tool for sex education? 2. Does AI improve the delivery, efficacy, and relevance of sexual knowledge and practices? 3. Does the adoption of AI foster supportive social norms, positive attitudes, and a sense of empowerment in users for learning about sexual health topics? 4. Are Serbian adolescents and young adults concerned about ethical issues such as privacy, bias, and the lack of human connection associated with AI-driven tools?

Keywords: artificial intelligence; sexual education; learning simulations; ethics; Serbia

Danijela Savaya

PhD candidate

Faculty of Political Sciences, University of Belgrade

Danijela Savaya is a PhD candidate in Media and Communication at the Faculty of Political Sciences, University of Belgrade. Savaya received her MA in Media Studies from Notre Dame University–Louaize, and BS in Mass Communication from Campbellsville University, USA. Savaya is a highly accomplished athlete who competed for over 20 years at the highest level of professional basketball. She advocates for open, informed discussions on sexual health.

Contact: danijela.savaya@hotmail.com

Jessica R. El-Khoury

Associate Professor

Notre Dame University–Louaize

Jessica R. El-Khoury is an Associate Professor in the Department of Media Studies and Director of International Relations at Notre Dame University–Louaize. Dr. El-Khoury received her PhD in Mass Communication from Texas Tech University with a concentration on health communication and entertainment-education, an MA in Media Studies and a BA in Broadcast Journalism. She is passionate about promoting a healthier society and building self/collective confidence through media messages which endorse positive behaviors such as anti-domestic violence, anti-drug addiction, community engagement, proactive cancer awareness, sex education and inclusion for persons with disability. She has presented her research at 41 conferences internationally, is published in renowned media communication journals, and has been a writer for a local magazine and was a reporter, anchor, and producer for a TV station in Texas.

Contact: jessica.elkhoury@ndu.edu.lb

Library, Librarian and Robot: “Megdan: Between Water and Fire”¹

Dragana Milunović

Modern library and information activities have been grounded for decades in the application of advanced techniques and technology. This has led to the creation and implementation of digital tools that facilitate the successful organization of knowledge. We witness this trend through the development of digital humanities, a field that has made an unprecedented amount of digital documents accessible to both scientific research and the general public.

Beyond expanding accessibility and simplifying document management, the emergence of artificial intelligence (AI) within the library and information industry has enabled faster and more efficient resource discovery, improved cataloging and classification processes, easier material organization, personalized user queries to enhance services, conversion of various resource types, and analysis of user and library needs. Some libraries in developed countries have gained valuable experience in the application of AI, contributing to the profession’s development while simultaneously raising important questions about the human element in this phenomenon. This includes reflections on the place and role of humans in the future, especially in the context of emerging IT trends.

Keywords: libraries; librarians; users; knowledge organization; AI tools

1 Inspired by the title of Aleksa Balašević’s film.

Dragana Milunović

Deputy Director

National Library of Serbia

Dragana Milunović has been a vital member of the National Library of Serbia since 2000, dedicating her career to advancing library science and fostering accessibility. Her professional journey includes training at renowned institutions such as the Green Libraries of Germany, Ca' Foscari University, the Finnish National Library, the Russian State Library, the Canadian National Institute for the Blind, and the Royal National Institute for the Blind. These experiences have broadened her expertise and deepened her understanding of global library practices. In 2015, she completed her doctoral studies, focusing on the role of contemporary libraries in improving reading accessibility for individuals with print disabilities. Beyond her institutional role, she actively contributes to international organizations like the International Internet Preservation Consortium, the IFLA Standing Committee for National Libraries, the Open Data Working Group, the CENL Copyright Group, and EIFL-IP. Dragana is also a passionate advocate for knowledge dissemination, serving as Editor-in-Chief of Herald of the National Library of Serbia and contributing to several academic editorial boards. Her prolific research includes over 100 publications and seven monographs, covering topics like e-inclusion, knowledge organization, and reading theory. Recognized with multiple prestigious awards, she continues to shape the future of libraries and information science on a global scale.

Contact: dragana.milunovic@nb.rs

Changing a Tire on a Moving Car: The Challenge of AI Alignment in Education

Ernest Ženko

The challenge of ensuring that artificial intelligence (AI) systems act in accordance with human values and intentions, known as AI alignment, is becoming increasingly critical as AI becomes more integrated into education. In educational settings, AI systems have the power to influence learning paths, ethical decision-making, and the transmission of cultural norms. Aligning AI with human values is essential to prevent biases, promote inclusivity, and ensure that AI supports, rather than undermines, critical thinking and moral development. The success of AI in education depends on its ability to navigate the complexities of human values and foster ethical learning environments.

This presentation explores how aligning AI systems with ethical frameworks in education not only enhances learning outcomes but also respects students' rights, promotes equity, and fosters moral development. However, this alignment must be bidirectional—education itself must adapt to align with the rapid advancements in AI technologies. As AI systems are shaped by ethical frameworks rooted in human values, educational practices must also evolve to integrate AI responsibly. This involves revising curricula, teaching methods, and institutional policies to equip students and educators to interact ethically with AI systems.

Human values form the foundation of both ethical frameworks and educational practices, guiding decisions and behaviors across individuals and systems. Ethical frameworks—whether deontological, utilitarian, or virtue-based—provide structured guidelines for AI systems to align with principles such as fairness, autonomy, and justice. These frameworks formalize complex moral principles, ensuring that AI systems act ethically in diverse and evolving situations. Simultaneously, education must embody these principles, adapting its structures and strategies to remain relevant and ethically sound in an AI-driven future.

However, human values are not monolithic; they are context-dependent, often conflicting, and subject to change. This complexity

poses significant challenges for AI alignment. Translating abstract human values into actionable principles or rules is fraught with difficulty, especially given the lack of a universally accepted ethical framework. Should AI systems prioritize utilitarian principles, deontological ethics, or virtue ethics? The problem of moral uncertainty complicates alignment, as no single moral theory offers a definitive solution to all ethical dilemmas.

Furthermore, as technology evolves rapidly, ethical frameworks and AI alignment must remain adaptable. Aligning AI with current human values addresses only part of the broader challenge; frameworks must also be reassessed and updated as AI becomes more integrated into education. This demands continuous evaluation and recalibration of ethical decision-making processes to meet the shifting societal needs.

AI systems do more than reflect human values—they actively shape them. This reciprocal relationship raises complex questions about the evolving interplay between AI and morality. As AI systems influence ethical decision-making and cultural norms, they become both mirrors of human values and agents of their transformation. Ensuring the alignment of AI systems in education is thus an ongoing, multifaceted challenge requiring engagement with ethical, cultural, and technological dimensions of human life.

Keywords: AI alignment; human values; ethical frameworks; education technology; moral development

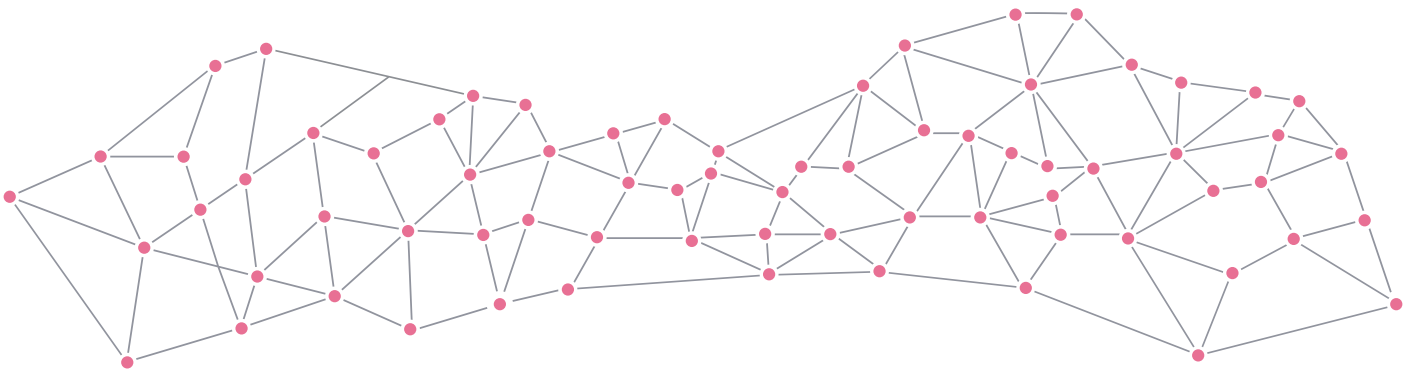
Ernest Ženko

Full Professor

University of Primorska

Dr Ernest Ženko is a full professor of philosophy of culture at the University of Primorska, Slovenia. He holds a PhD in philosophy and has extensive research experience in aesthetics, visual media, and the history of ideas. Dr. Ženko is the head of the workgroup for establishing an AI center at the University of Primorska and a member of the Task and Finish Group on AI at the European University Association. He has lectured internationally, published widely, and is active in various professional associations. His academic interests also include global interculturality and the ethics of artificial intelligence.

Contact: ernest.zenko2@gmail.com



Attitudes of Serbian Youth Toward the Use of Artificial Intelligence for Educational Purposes: Competence, Trust, and Privacy Concerns

Marija Antonijević, Iva Medojević

The emergence and integration of Artificial Intelligence (AI) into various aspects of life have significantly transformed how individuals interact with technology. Young people, as early adopters of innovation, utilize AI for diverse purposes, including entertainment, education, and marketing. Understanding the implications of AI usage on youth development is crucial, yet existing literature lacks a comprehensive analysis of youth perceptions and attitudes toward AI technologies, particularly in educational contexts. Research focusing on Serbia in this area is notably scarce.

This study aims to examine the attitudes of Serbian youth toward AI tools, specifically for educational purposes, by analyzing usage patterns and exploring respondents' competence in using AI tools, privacy concerns, and intentions to adopt AI tools for learning. Data were collected via an online questionnaire in September 2024, with all variables measured on a 7-point Likert scale. The sample comprised 201 individuals aged 18-35.

The findings reveal that 60.3% of respondents rate their competence in using AI tools as moderate (scores 3, 4, and 5). A majority (81.6%) have experience using AI tools, with the most commonly used being ChatGPT (94.5%), Canva (57.3%), and Grammarly (40.9%). Only 10% of respondents strongly trust the accuracy and reliability of AI-based tools (scores 6-7), while just 5% strongly believe that their personal data is safe and confidential when using these tools (scores 6-7). Nonetheless, 48% of respondents demonstrate a high willingness to use AI tools for learning and skill development (scores 6-7).

These results suggest that while Serbian youth are open to adopting AI for educational purposes, addressing privacy and trust concerns is essential for fostering greater acceptance and improving the effective integration of AI into educational settings. The conclusions underscore the need for a systematic approach to equipping young people in Serbia with the knowledge and skills necessary for responsibly

utilizing AI technology. Such an approach should also prioritize data management and security to build trust and ensure the ethical use of AI tools.

Keywords: education; AI; emerging technologies; data privacy

Acknowledgments: This research was funded by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia under contract number 451-03-47/2023-01/200005. Special thanks to the Alumni Foundation of the University of Belgrade, the Centre for Career Development of the University of Belgrade, and the Center for Doctoral Studies Support at the University Library "Svetozar Marković" in Belgrade for their invaluable assistance in data collection.

Marija Antonijević

Research Assistant

Institute of Economic Sciences

Marija Antonijević is a Research Assistant at the Institute of Economic Sciences in Belgrade, Serbia (Department of Digital Economics). She is a PhD candidate at the Faculty of Economics, University of Belgrade. Her research interests include digital banking and digital entrepreneurial competencies. Marija has published numerous scientific papers in mentioned research areas. She is a team member of the COST Action CA19130—Fintech and Artificial Intelligence in Finance—Towards a transparent financial industry 2020/2024. Since 2022, she has been the secretary of the Journal of Women’s Entrepreneurship and Education. In March 2024, Marija became a member of the Scientific Editorial Board of the Alumni Foundation University of Belgrade.

Contact: marija.antonijevic@ien.bg.ac.rs

Iva Medojević

PhD Student

Teacher Education Faculty

Iva Medojević holds a Master’s degree in Elementary Education and is a doctoral student at the Teacher Education Faculty, University of Belgrade. She is a recipient of the scholarship awarded by the Serbian Ministry of Science, Technological Development, and Innovation. As a scholarship holder, she actively participates in the activities of the Institute for Educational Research in Belgrade. The field of her research covers elementary Serbian language and literature teaching methodology, specifically elementary Serbian language (grammar and spelling) teaching methodology. Since March 2024, Iva has been a member of the Scientific Editorial Board of the Alumni Foundation University of Belgrade.

Contact: ivamedojevic@yahoo.com

The Beijing Dissensus: Can There Really Be International Alignment on AI and Education?

Miloš Račić

This paper examines the creation, philosophy, and implementation challenges of the *Beijing Consensus on Artificial Intelligence and Education*, the outcome document of UNESCO's 2019 International Conference on Artificial Intelligence in Education. It situates the Consensus within the broader geopolitical struggle among leading international actors over the value systems guiding AI alignment. Education, as a state-defined process that socializes young people into adopting societal values, represents a particularly contentious domain for international cooperation due to the stark disparities in value systems across different blocs of countries. These disparities have only deepened in recent years, even as the rapid advancement of AI has made such cooperation increasingly urgent.

The *Beijing Consensus* was shaped under the leadership of the People's Republic of China, following the withdrawal of the United States from UNESCO in 2018, citing alleged anti-Israel bias. The US absence created a power vacuum within the organization at a critical juncture, allowing China to take a leading role in shaping global norms on AI in education. This paper explores how China's "core socialist values," as outlined by the Communist Party of China in 2012, inform its approach to AI in education and questions whether these values can be reconciled with Western liberal principles.

The analysis delves into whether the *Beijing Consensus* reflects merely contemporary Chinese political ideology or represents a genuine international agreement that diverges from Western conceptions of freedom, democracy, and human rights. With the Biden Administration reversing course and rejoining UNESCO in 2023, largely due to the AI boom and the need to shape global norms, the paper considers how the US presence will challenge specific elements of the Consensus. It examines which aspects of the *Beijing Consensus* will face the greatest scrutiny and how this might impact its implementation.

Ultimately, the paper concludes that greater international alignment on AI and education is possible, but it will require the establishment of a new framework of international morality. This framework must

transcend the temporary agendas of leading global powers and focus on deeper consensus regarding the fundamental purpose of education in an AI-driven world.

Keywords: Beijing Consensus; UNESCO; AI alignment; value systems; philosophy of education

Miloš Račić

Student

University of Belgrade, Faculty of Mathematics

Miloš Račić is a student at the University of Belgrade and an international youth rights activist interested in democracy, digital and soft skills development, meaningful youth-centered education reform, and youth's role in technological advancement. He is currently a member of the UNICEF Serbia Youth Advisory Board and the United Nations Association of Serbia Executive Board, and is active in several other youth organizations. He is also the founder of the Debate Society of the Faculty of Mathematics at the University of Belgrade.

Contact: milos.bv.racic@gmail.com

Perceived Opportunities and Risks of Implementing Data Encryption in AI-Powered Chatbot for Enhancing Student Support Service in Nigerian Universities

Suleiman Yusuf

The integration of Artificial Intelligence (AI)-powered chatbots for student support services in Nigerian universities presents significant opportunities and risks, particularly in terms of data security and privacy. This study employs a mixed-methods approach to explore stakeholders' perceptions regarding the implementation of data encryption in AI-powered chatbots. The research is guided by the growing concern over data privacy and the need for robust security measures to protect sensitive student information in the digital age.

The quantitative phase of the study involved a survey administered to 374 stakeholders, including students, faculty, IT staff, and administrators from various Nigerian universities. The survey aimed to assess their awareness, perceived benefits, and concerns related to the implementation of data encryption in AI-powered chatbots. Descriptive and inferential statistical analyses were conducted to identify trends and correlations among stakeholder groups.

Results indicated a high level of awareness about data security issues, with 78% of respondents recognizing the importance of encryption for protecting sensitive information. However, 65% of respondents expressed concerns about the potential impact of encryption on chatbot performance and user experience.

The qualitative phase involved semi-structured interviews with a purposive sample of 32 participants from the same stakeholder groups. The interviews provided deeper insights into the perceived opportunities and risks of using data encryption in AI-powered chatbots. Thematic analysis revealed that stakeholders see significant opportunities for improving data security and building trust among users through encryption. For instance, encrypted data transmission and storage were perceived as critical for complying with data protection regulations and preventing unauthorized access. However, stakeholders also highlighted risks such as increased complexity in chatbot management, potential delays in response time, and higher costs associated with implementing and maintaining encryption technologies.

Therefore, the integration of findings from both quantitative and qualitative phases suggests that while stakeholders are generally supportive of data encryption for enhancing security, they remain concerned about its implications for the usability and efficiency of AI-powered chatbots. The study concludes that a balanced approach is necessary, one that incorporates robust encryption mechanisms without compromising the effectiveness of student support services. It recommends the development of adaptive encryption models that can dynamically balance security and performance based on the sensitivity of the data being handled.

This study contributes to the literature on the adoption of AI technologies in higher education by providing empirical evidence on stakeholders' perspectives regarding data encryption in student support services. It underscores the need for Nigerian universities to engage stakeholders in the decision-making process and to invest in training and awareness programs to ensure successful implementation. Future research should explore the long-term impact of encryption on user satisfaction and the operational efficiency of AI-powered chatbots in university settings.

Keywords: artificial intelligence; support service; stakeholders; university system; data encryption

Yusuf Suleiman

Director, Centre for Research, Industrial Linkage and International Cooperation

Al-Hikmah University, Nigeria

Yusuf Suleiman is a distinguished educationist and researcher with a wealth of experience in educational leadership and research. He currently serves as the Director of Research, Industrial Linkage, and International Cooperation at Al-Hikmah University, Ilorin, Nigeria, and is a Senior Lecturer in the Department of Educational Management and Counselling. Dr. Suleiman holds a PhD in Education and a Master's in Policy and Strategic Studies, with a proven track record of academic excellence. His research interests lie in student support services, artificial intelligence (AI) in education, and higher education management. His work focuses on the integration of AI-powered tools to enhance learning experiences and the effectiveness of support services for students. Passionate about collaborative research, Dr. Suleiman has contributed significantly to the academic community through partnerships with universities across Nigeria and West Africa. He has participated in various international conferences and is committed to shaping the future of education by promoting AI-driven innovations that improve student engagement, support, and success. Dr. Suleiman is a proactive advocate for higher education reforms, constantly exploring new approaches to improve teaching, learning, and student support systems.

Contact: yusufsuleiman@alhikmah.edu.ng

Potentiality of GenAI: Application of Generative Artificial Intelligence in Academic Writing in Liberal Arts

Umma Maimuna Alam

Generative Artificial Intelligence (GenAI) has a vast potential for application in academic writing. With its growing advancements and increasing popularity, the integration of GenAI in education is becoming inevitable. From brainstorming arguments to drafting complete essays, GenAI can serve as a valuable tool throughout the academic writing process. This study investigates the results of a survey conducted with undergraduate students inclined to use GenAI for academic writing, particularly in liberal arts disciplines such as philosophy, political science, literature, psychology, and history.

The research explores the challenges students face while using GenAI for academic writing, examining how they utilize AI in essay creation and identifying effective strategies for its integration. By doing so, this study aims to contribute to transforming GenAI into a powerful learning tool, supported by proper guidelines. The targeted participants represent diverse academic backgrounds in undergraduate studies, enabling the research to explore current and historical trends in AI usage among students while projecting future trajectories for its application in liberal arts courses.

Employing a mixed-methods approach, the study comprehensively analyzes students' perceptions of the benefits, expectations, and challenges of integrating GenAI into the learning process. By focusing on evidence-based insights, the research contributes to discussions on the potential role of GenAI in liberal arts academia. Rather than promoting excessive dependence on GenAI, the study emphasizes the importance of educators adopting integrated guidelines to align AI as an effective tool for enhancing students' academic writing skills.

Keywords: AI in academic writing; GenAI integration; ethical practices

Umma Maima Yusuf

Teaching Assistant

Brac University

Umma Maimuna Alam is a Teaching Assistant and recent graduate of BRAC University. She majored in Literature from the English and Humanities department. While researching for a civic community service project, she learned several research methods and also actively participated in conducting research using different research methodologies. She focused her undergraduate research on studying Transnationalism as a medium to navigate through Nationalism in a globalized world. She is eager to pursue a postgraduate degree in interdisciplinary programs that bridge the humanities and practical applications. With her passion for comparative studies and active involvement in various research programs, she is driven to further immerse herself in academia.

Contact: maimuna.alam@bracu.ac.bd

Developing Critical AI Literacy Skills for Ethical and Responsible Use of AI

Inesa Stolper

As artificial intelligence (AI) becomes increasingly prevalent across various fields, offering benefits such as efficiency and automation, it also introduces significant risks, including concerns over data protection, bias, and ethical misuse. Addressing these challenges requires raising awareness about AI's ethical and responsible use. Teaching critical AI literacy skills is a crucial approach to fostering this awareness and promoting informed engagement with AI technologies.

This need is particularly urgent in disciplines like law, where AI is used for research, decision-making, and other professional tasks. For example, students often rely on AI tools to complete assignments, sometimes without fully understanding the ethical risks, such as plagiarism or over-reliance on AI-generated content. A notable incident in 2023 involved lawyer Steven A. Schwartz, who submitted court documents containing fabricated case citations generated by ChatGPT. This mistake, caused by "AI hallucination," underscored not only the limitations of the technology but also the importance of user responsibility when employing AI tools. Similarly, the increasing use of AI by judges raises critical questions about fairness and accountability in legal decisions. These examples highlight the necessity of fostering AI literacy that emphasizes ethical use and professional responsibility.

The goal of teaching critical AI literacy skills is to help students and professionals understand how AI systems work and how to use them responsibly and ethically. These skills are relevant across many disciplines, as AI increasingly influences decision-making in fields such as healthcare, public administration, finance, and education. In healthcare, for instance, AI assists in diagnostics and treatment planning. However, over-reliance on algorithms or biased data can negatively impact patient outcomes without a critical understanding of the technology. Similarly, in public administration, AI tools for fraud detection and risk assessment can lead to unethical consequences if not carefully monitored. The 2020 SyRI case in the Netherlands serves as a cautionary example, where an algorithm used discriminatory criteria for fraud detection, leading to severe financial and personal harm for thousands of individuals.

Fostering AI literacy enables individuals to critically evaluate AI outputs, recognize technological limitations, and address potential ethical risks. By focusing on more than just technical proficiency, AI literacy empowers users to make informed, responsible decisions when engaging with AI systems, ensuring that its implementation aligns with ethical principles.

In conclusion, developing critical AI literacy skills is essential for addressing the ethical challenges posed by AI. By integrating these skills into education and professional training, we can ensure AI is used responsibly and thoughtfully, not only adapting to its growing presence but also shaping its role in society for the better.

Keywords: AI literacy; ethics; responsible AI use; education; decision-making

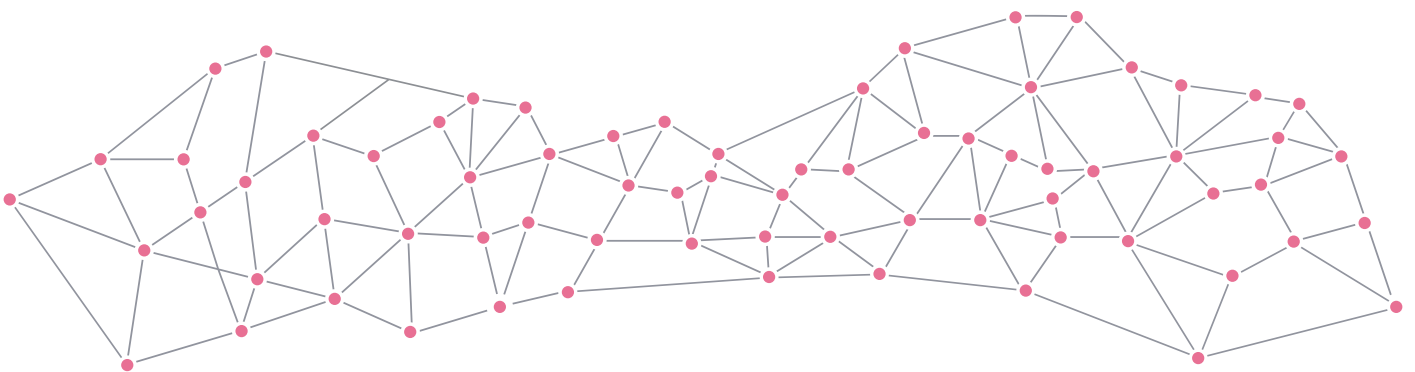
Inesa Stolper

Lecturer, PhD candidate

European Humanities University, Vilnius, Lithuania

Inesa Stolper is a Lecturer at the Department of Social Sciences, European Humanities University in Vilnius, Lithuania, where she has been teaching since 2012. Inesa specializes in courses such as Digital technologies and law, Legal Writing, and E-justice. She is interested in the integration of technology into legal education, a focus evident in her recent work leading the Digital Humanities Lab at EHU. Inesa is currently pursuing her PhD at Mykolas Romeris University, Lithuania. Inesa's research focuses on the use of technologies in the court and the right to a fair trial.

Contact: inesa.stolper@ehu.lt



A Holistic AI Curriculum for Young Learners

Mariela Destéfano

The rapid rise of artificial intelligence (AI), particularly since the introduction of ChatGPT in 2022, has brought AI into daily conversation, media coverage, and transformative applications across industries, including education. While the potential of AI to reshape education has been acknowledged for years, its impact is now visible at all levels of learning.

Studies highlight widespread AI tool usage in higher education for tasks such as research support, automated grading, and enhanced human-computer interaction. For example, nearly two-thirds of university students in Germany reported using AI-based tools (von Garrel & Mayer, 2023), while Dempere et al. (2023) identified both significant benefits and concerns, such as plagiarism, online test security, job displacement, and the digital literacy gap.

The influence of AI is also evident in K-12 education. A 2021 European survey predicted a “powerful and immersive use of AI in education” (European Schoolnet, 2021). More recently, Zhang & Tur (2023) emphasized ChatGPT’s potential for curriculum design, lesson planning, and personalized education, alongside challenges such as superficial learning habits and diminished critical thinking skills (Mogavi et al., 2024). Additional barriers include teachers’ lack of AI knowledge, limited teaching guidelines, and inadequate curriculum design.

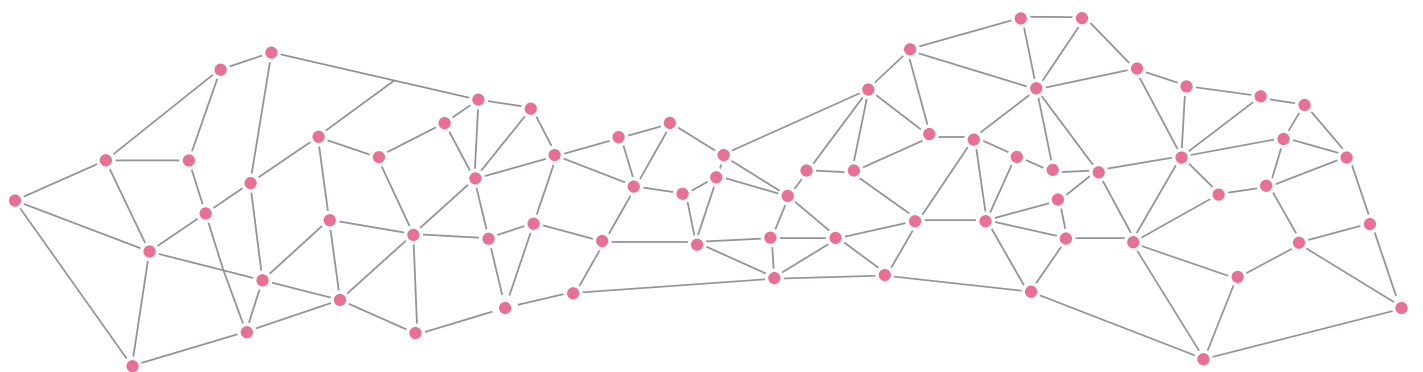
This talk introduces a comprehensive AI curriculum for young learners aged 11 to 14, emphasizing a humanistic and multidimensional approach. The curriculum integrates technological, philosophical, cognitive, and cultural dimensions, drawing from fields such as the philosophy of mind, cognitive sciences, child development, and digital literacy. Its humanistic nature lies in embedding philosophical insights into pedagogical content, addressing not only the technological aspects of AI but also its cognitive and social implications (Williams et al., 2019; Su & Yang, 2022).

During the 2023/24 school year, this curriculum was piloted as an extracurricular program at a secondary school in Barcelona, Spain,

through the CreaTIC Academy. It incorporates practical tools such as Scratch, App Inventor, and Machine Learning for Kids, alongside discussions on philosophical and developmental concepts. Designed to adapt to diverse social and economic contexts, the curriculum aims to foster a comprehensive understanding of AI's role in society while enhancing students' digital literacy, critical thinking, and ethical awareness.

By combining technological instruction with a humanistic perspective, this curriculum offers a unique contribution to AI education, preparing young learners to navigate the complexities of an AI-driven world responsibly and thoughtfully.

Keywords: artificial intelligence education; multidimensional curriculum; philosophy of mind; youth digital literacy



Mariela Destéfano

Associate Researcher

CreaTIC Academy, Barcelona; National Scientific and Technical Research Council (CONICET)

Mariela Destéfano holds a PhD in Philosophy from the University of Buenos Aires, specializing in the philosophy of mind and cognitive science. Her research focuses on the relationship between new technologies and education. She conducts her investigations at CreaTIC Academy in Barcelona. She is also an associate researcher with the National Scientific and Technical Research Council (CONICET) and was part of the research team for the project PICT-2014-3422 ("The Relationships Between Cognitive Architectures and Explanations in Cognitive Science") from 2016 to 2019. Some of her publications include: "Conceptual and Semantic Representations in Cognitive Sciences" in *Discusiones Filosóficas* (2020); "Language Processing and Informational Semantics" in *Práxis Filosófica* (2019); "Double-Process Theories: A Unified Cognitive Architecture?" (with F. Velazquez UNS) in *Theoria: An International Journal for Theory, History, and Foundations of Science* (2018); and "Fodor's Non-Conceptual Representations and the Computational Theory of Mind" in *Journal of Cognitive Science* 14 (2013), among others.

Contact: mariela.destefano@gmail.com

Ethical Use of AI Tools in Writing-Based Learning Methods: Challenges and Opportunities

Mariia Laktionkina

Integrating Artificial Intelligence (AI) tools into writing-based learning offers significant opportunities while posing ethical challenges in educational contexts. This paper examines the implications of AI use across various established writing methods, including Essay, Loop Writing, Believing and Doubting, Dialectical Response Notebooks, Informal Writing, Focused Freewriting, Attitudinal Writing, Metacognitive Process Writing, and Narrative/Descriptive Writing. While AI-powered tools can assist students by providing real-time feedback, improving grammar, and generating ideas, their application raises critical concerns regarding academic integrity, creativity, and the development of essential writing and critical thinking skills.

Each of these writing methods has distinct pedagogical objectives, such as fostering critical analysis in *Believing and Doubting*, promoting self-awareness in *Attitudinal Writing*, and encouraging spontaneity in *Informal Writing*. However, the misuse of AI tools could undermine these goals by fostering reliance on automated content or pre-structured suggestions, potentially diminishing the authenticity and depth of students' work. This is particularly concerning in methods like *Focused Freewriting* and *Metacognitive Process Writing*, where personal reflection and cognitive engagement are central to achieving learning outcomes.

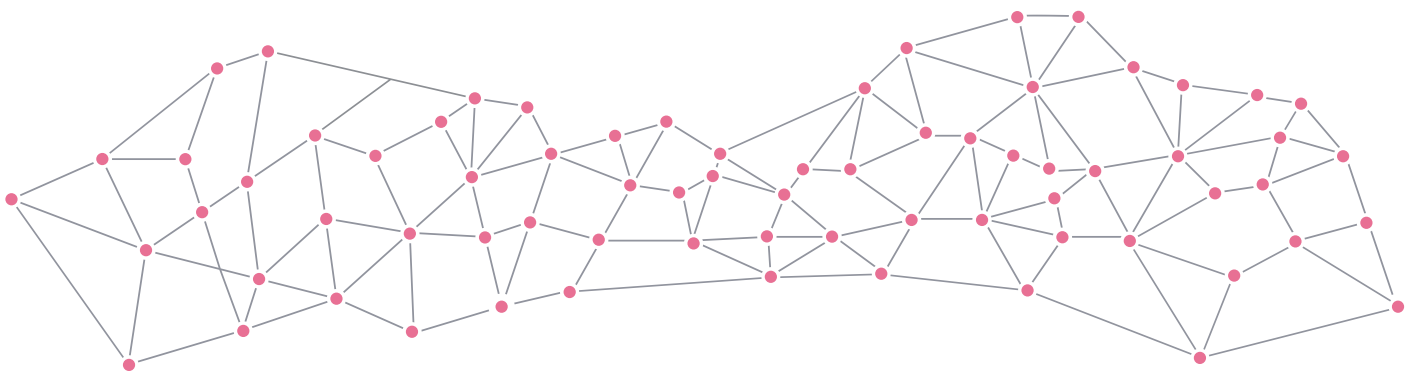
The paper also highlights fairness concerns, as AI systems trained on biased datasets may unintentionally perpetuate social and cultural inequities in writing feedback and content generation. This is especially problematic in *Narrative and Descriptive Writing*, where diverse voices and perspectives are essential. Ensuring that AI tools are inclusive and equitable is a critical consideration for their ethical application.

Additionally, the impact of AI on metacognitive skills is explored, particularly in methods requiring self-reflection, such as *Metacognitive Process Writing*. Over-dependence on AI tools could hinder the development of critical self-assessment and creative problem-solving abilities, diminishing the transformative potential of these pedagogical methods.

This paper advocates for a balanced approach to AI integration in writing-based education, emphasizing the importance of preserving the integrity of these teaching methods. Educators play a crucial role in guiding students to use AI ethically, ensuring that it serves as a supplement to, rather than a substitute for, intellectual and creative development. Institutional policies are also essential to clearly define the responsible use of AI in academic writing, safeguarding against academic dishonesty and over-reliance on technology.

In conclusion, while AI tools offer significant benefits for writing-based learning, their ethical use requires thoughtful consideration to prevent the erosion of essential skills and uphold academic integrity. This paper proposes a framework for the responsible adoption of AI in education, ensuring that it enhances learning outcomes without compromising the values and objectives of traditional writing methods.

Keywords: AI in education; writing-based learning; ethics; academic integrity; critical thinking



Mariia Laktionkina

Lecturer

Academic Department of Humanities and Arts, European Humanities University

Mariia Laktionkina is a faculty member at the European Humanities University and a faculty associate at the Institute for Writing and Thinking at Bard College. She also serves as an Academic Adviser at the Open Society University Network. Laktionkina teaches various courses, including Language & Thinking, First-Year Seminar: Introduction to Humanities, Second-Year Seminar: Introduction to Social Sciences, Hermeneutic Seminar, and Global Citizenship. Her professional interests focus on the role of the humanities in shaping individuals who can think across disciplines, apply critical judgment, navigate complex and conflicting worlds, and adapt to new and unpredictable environments. In addition to her teaching responsibilities, Laktionkina has been actively involved in designing new programs and implementing innovative strategies in the curriculum at EHU. Her current research and practical interests include philosophy of literature, pedagogical practices in liberal arts education, public policy, and higher education management. Since April 2024, Laktionkina has been leading a project at EHU titled "AI-Aware Universities: Empowering University Communities for the Ethical Use of AI."

Contact: maria.laktionkina@ehu.lt

AI for a Liberatory and Transformative Pedagogy

Tamara Kamatović

Liberation pedagogy, together with its research counterpart participatory action research, highlights the active role of learners as responsible participants in education who contribute to societal transformation and liberation. Transformative education has traditionally emphasized developing curricula that are flexible and responsive to students' interests and lived experiences, rather than being rigid and prescriptive (Ćumura & Petrović, 2012). The rise of artificial intelligence (AI) technologies, particularly large language models (LLMs) like ChatGPT, has sparked ethical debates in education around concepts of "authenticity" and "norms." These discussions have profoundly influenced policies and strategies for learning outcome. However, beyond these normative debates, it is crucial to examine the purpose of learning itself in the context of AI technologies, which often "surpass" traditional human cognitive approaches.

By focusing on the concept of "transformation" in education (Ashwin, 2020), this paper argues that AI tools can complement curricular reforms to support active pedagogy, fostering transformative educational experiences that promote "critical and liberating dialogue" (Freire, 1970, p. 47). The paper emphasizes how AI technologies can support socially transformative goals by addressing structures of internal exclusion. These include enhancing classroom accessibility, facilitating deliberative democratic processes, and fostering awareness of forms of address and speech within educational settings.

Additionally, the paper engages with technoskeptic critiques of AI in education, which argue that these technologies predominantly reinforce growth-oriented, neoliberal education models. Critics contend that such models prioritize skills that support technological and economic growth at the expense of critical thinking, thereby suppressing dissent and independent thought (Nussbaum, 2010). The paper counters these critiques by demonstrating how, when integrated thoughtfully, AI tools can align with the principles of liberatory pedagogy, fostering spaces for critical engagement and societal transformation.

Keywords: transformation; democratic education; liberatory pedagogy; edtech

Tamara Kamatović

Lecturer

European Humanities University

Tamara Kamatović is a Lecturer at the Yehuda Elkana Center for Teaching, Learning, and Higher Education Research. She designs and teaches courses and facilitates workshops for doctoral students and faculty at CEU and partner institutions, with a focus on inclusive teaching, research-enriched teaching, and technology-enhanced learning. Tamara also leads and coordinates the mentorship of Global Teaching Fellows. Her research and publications explore the relationship between generative AI, authorship, learning, and writing practices. She has co-authored work examining the modality of online teaching and its connection to inclusive and democratic pedagogical practices. Additionally, she is co-leading a book project that compiles best practices in inclusive teaching from faculty worldwide, with editorial introductions investigating the intersections of civic society, democratic education, and inclusivity. Tamara has presented at conferences on topics related to democratic education and has delivered talks on the philosophy of education. Her work engages with themes such as training and educating “intelligence” in the context of emerging technologies, Nietzschean perspectives on teaching with technologies, and the biopolitics of emerging technologies, including data analytics and surveillance.

Contact: kamatovicT@ceu.edu

Libraries in the Age of AI: Challenges and Possibilities

Dragana D. Jovanović

Throughout history, libraries have been regarded as bastions of tradition, devoted to collecting and preserving human knowledge and cultural heritage, primarily in physical, book-like form. However, their traditional role has continuously evolved with the adoption of new technologies. Libraries have consistently embraced technological innovations, adapting not only the mediums used to store intellectual output but also transforming their workflows and processes.

Today, libraries face a new and pressing challenge: responding to the rapid development of artificial intelligence (AI). As AI continues to evolve at an unprecedented pace, it brings both revolutionary opportunities and significant ethical concerns. Potential applications of AI in libraries include using descriptive AI to improve the accessibility of library collections and fostering AI literacy among users and staff. However, these advancements also raise risks, such as copyright violations from unauthorized use of text and data, and challenges in cataloging due to uncertainties surrounding authorship in AI-assisted publications.

This study examines the impact of AI on libraries and explores the efforts made by librarians to understand and integrate machine learning (ML) and AI technologies into their practices. The analysis incorporates insights from leading global libraries, including the Library of Congress and La Bibliothèque Nationale de France, and highlights key recommendations from strategic documents published by the International Federation of Library Associations and Institutions (IFLA).

By analyzing these experiences and strategies, the paper seeks to identify both the opportunities and risks that AI brings to libraries, providing a roadmap for navigating this transformative era responsibly and effectively.

Keywords: AI; libraries; artificial intelligence; library and information science

Dragana D. Jovanović

Senior librarian

Matica Srpska Library

Dragana D. Jovanović was born in 1977 in Novi Sad, Serbia. She received her PhD in Library and Information Science (2016) from the Faculty of Philology, University of Belgrade. She conducted research for her dissertation, *The Functions of Contemporary Media in the Improvement of Library and Information Science: Comparative Analysis of Experiences in France and Serbia*, in Paris as a scholarship recipient of the French Ministry of Culture. She earned her BA in French language and literature, and MA in linguistics at the Faculty of Philosophy, University of Novi Sad. She is a senior librarian in The Matica Srpska Library in Novi Sad, where she works in the Department for Cataloguing and Bibliographic work. She also coordinates the media team and oversees the organization of various events at the Library. Prior to her career as a librarian, she spent a significant part of her professional life working in the media. Her areas of interest include new media, communication, and public relations. She is the author of numerous papers and the monograph *Univerzum informacija* (Biblioteka Matice srpske, Novi Sad, 2019).

Contact: drdraganad.jovanovic@gmail.com

Critiquing Cross-Cultural Ethics in Artificial Intelligence in Education (AIED)

Nasreen Watson

Ubuntu, a philosophy advocating communal living and interconnectedness, has been proposed as an ethical framework across social and economic spheres in South Africa. However, challenges persist in applying this framework to Artificial Intelligence in Education (AIED), particularly in addressing ethical dilemmas faced by first-year university students. Enslin and Horsthemke (2004) argue that while Ubuntu's tenets are compelling, they overlap with other humanistic philosophies and fail to sufficiently address common value structures influenced by Western ethical traditions. This critique questions the prioritization of Ubuntu in education, highlighting its limitations in forming universal ethical frameworks for AIED.

The purpose of this research is to critically evaluate whether Ubuntu can effectively address the ethical challenges faced by first-year university students in the context of AIED. Building on Enslin and Horsthemke's arguments, this study examines the tensions inherent in applying cross-cultural ethical frameworks to AIED. It explores the limitations that arise when frameworks like Ubuntu are amalgamated with Western ethical traditions, potentially hindering the practical implementation of AIED.

While Ubuntu offers a unique perspective for addressing social and ethical challenges, this research asserts that it may fall short in providing actionable solutions for integrating AI into educational practices. This limitation could impede the advancement of African educational institutions in adopting global standards of partnership that enhance student development. Furthermore, the study critiques the viability of human rights frameworks as universal ethical standards for AIED, questioning their ability to integrate diverse cultural values effectively.

By highlighting these tensions, this research contributes to the ongoing debate on ethical AI governance and its implications for marginalized student groups. It aims to foster a deeper understanding of the cultural and philosophical challenges in developing ethical standards for AIED that are both inclusive and practical.

Keywords: ubuntu; artificial intelligence in education; cross-cultural ethics; marginalized students; human rights framework; Western tradition

Nasreen Watson

Academic Writing Consultant

Department of Philosophy, University of Johannesburg

Nasreen Watson, whose name is derived from Arabic meaning “white rose,” is currently pursuing a Master’s degree in Artificial Intelligence Ethics after several years of professional experience in corporate HR at Standard Bank in South Africa. During her time in HR, she honed her leadership and technical skills, which inspired her to pursue a degree in Human Resource Management. Nasreen graduated with an undergraduate degree, achieving an overall average of 83%, and later earned her Honours degree with distinction (cum laude), where she explored the works of Friedrich Nietzsche, one of the most influential philosophers. She has also guest lectured at the University of Johannesburg and actively participates in national philosophical conferences. Her diverse background, encompassing both academic excellence and practical experience in strategic frameworks, positions her as a well-rounded academic. Additionally, Nasreen brings a touch of humor to her professional and academic endeavors, adding to her unique approach to learning and growth.

Contact: nasreen.watson@gmail.com

Can We Do It Alone? The Challenge of Reskilling Librarians in AI, Copyright, and Marketing

Marija Rakić Šarenac, Jasmina Marković

As digital channels transform how libraries engage with their communities, librarians are required to blend traditional roles with digital expertise to meet new expectations. They have three main options: adhering to traditional roles, self-teaching, or a more structured approach exemplified by the Erasmus+ project *Up-skill and Re-skill Librarians in Modern Marketing Solution Curriculum*. This initiative, led by the Public Library “Vuk Karadžić” in Serbia in collaboration with the National Association of Librarians and Public Libraries of Romania, enables librarians to efficiently acquire practical skills by working directly with experts in the field.

This approach aligns with *IFLA Code of Ethics for Librarians and other Information Workers*, which emphasizes that “Librarians and other information workers strive for excellence in the profession by maintaining and enhancing their knowledge and skills. They aim at the highest standards of service quality and thus promote the positive reputation of the profession” (IFLA, 2012). Working with experts ensures librarians can effectively navigate the digital transformation of their institutions, combining hands-on experience with theoretical learning.

The project focuses on developing a tailored curriculum that equips librarians with essential skills in digital marketing, content creation, and social media management. By adopting and implementing this curriculum, librarians will be better equipped to navigate digital transformation, enhancing their capacity to engage communities, increase the visibility of library services, and establish a stronger digital presence. The curriculum focuses on core competencies such as copywriting, graphic design, and content planning for social media platforms, incorporating the use of AI. These skills are crucial for creating more responsive and accessible library services and fostering meaningful audience engagement.

The curriculum consists of nine modules, including one dedicated to applying AI in content creation and another closely related module on licensing systems and copyright. The project’s initial activity

involved surveying librarians, revealing that while 42.4% have a basic understanding of AI's potential for content generation, only 9.1% actively use AI tools in their work. This finding highlights a significant gap that the project seeks to address through targeted training. Additionally, the curriculum introduces technologies such as virtual reality (VR) and augmented reality (AR), providing libraries with innovative tools to engage their communities. It also emphasizes the importance of balance between AI-generated content and human creativity.

The curriculum also addresses the need for librarians to understand licensing systems and copyright issues. Survey data revealed that only 42.4% of staff have a basic understanding of licensing systems, with just 15.2% being fully knowledgeable. Similarly, when it comes to free software licenses, 48.5% have basic knowledge, while only 12.1% are highly knowledgeable. These findings highlight the need for further training to ensure legal compliance and effective digital content management.

The integration of AI technologies into the library curriculum reflects a forward-thinking approach to librarianship in the digital age. Understanding the ethical and legal aspects of using AI and copyrighted material equips librarians to innovate responsibly, balancing technological advancement with respect for creative integrity. Additionally, the Licensing Systems and Copyright module underscores the importance of librarians becoming proficient in addressing copyright and licensing challenges, particularly in the creation and distribution of digital content.

Through the *Up-skill and Re-skill Librarians* project, librarians will be able to use AI technologies in the service of public welfare and innovation. With careful consideration of ethical concerns and current limitations, libraries can responsibly use AI technologies to advance their social mission (IFLA, 2012). However, this shift raises many questions that libraries cannot tackle alone. This is why we advocate for ongoing collaboration with industry professionals, as we believe this approach is essential for the continued evolution of libraries in the context of community service and engagement.

Keywords: libraries; AI technologies; education

Marija D. Rakić Šarenac

Senior librarian

Public Library Vuk Karadžić

Marija D. Rakić Šarenac is a senior librarian with 18 years of professional experience. Currently employed at the Public Library “Vuk Karadžić” in Kragujevac, she works in the Local Heritage Department where she provides support to patrons, researchers, and students. She is responsible for collecting, preserving, maintaining, and cataloging of both book and non-book materials. She also has practical experience in managing the Foreign Book Library and Children’s Library Department. An experienced Editor-in-Chief of the library’s magazine *Kragujevačko čitalište*, and former member of the editorial board of a literary magazine *Koraci*, she is skilled in text and library paper proofreading and translating. In 2023-2024, she took part in the Creative Leadership Academy and The Creative Mentorship program. She is the coordinator for the Erasmus+ project, *Up-skill and Re-skill Librarians in Modern Marketing Solutions Curriculum*. She has published 4 books of poetry.

Contact: masharakicsh@gmail.com

Jasmina Z. Marković

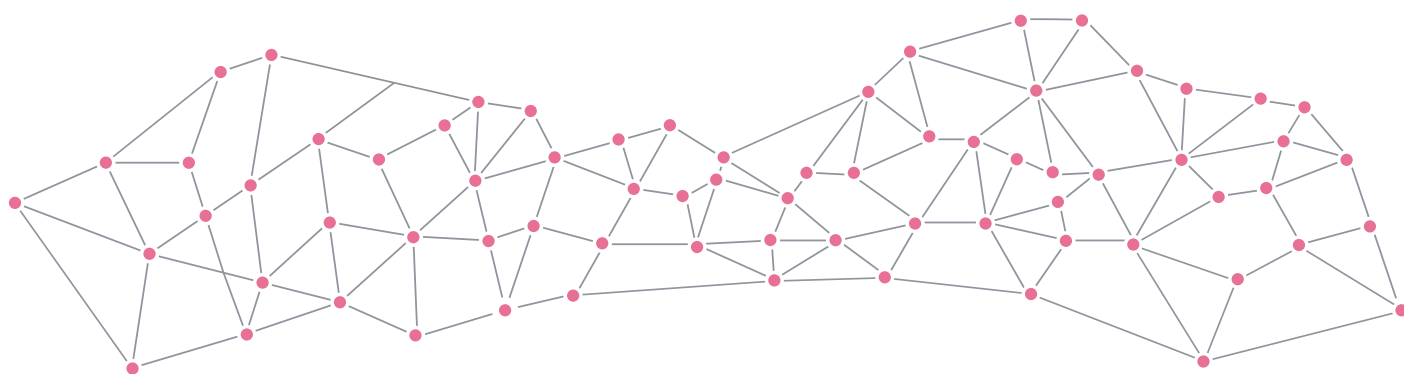
Senior librarian

Public Library Vuk Karadžić

Jasmina Z. Marković is a French language and literature teacher and senior librarian with over 10 years of professional experience. She is currently employed at the Public Library “Vuk Karadžić” in Kragujevac. After several years of working in the Adult Service Department and Children’s Departments, she transitioned to the Local Heritage Department in 2014. In this department, she supports patrons, researchers, and students while managing collection, preservation, maintenance, and cataloging of both book and non-book materials. Jasmina is also a co-author of several exhibitions celebrating notable

figures from Kragujevac, including Jovan Đ. Mirković (*Ljudi koji su menjali Srbiju: profesor Jovan Đ. Mirković muzički pedagog, bibliotekar, kosmopolita, filantrop i zadužbinar*) and Zarija D. Vukićević (*Možda će se jednom videti koliko sam voleo svoju zemlju i svoj narod: Zarija D. Vukićević – profesor, bibliotekar, književnik, prevodilac i diplomata*). Additionally, she is a team member of the Erasmus+ project, *Up-skill and Re-skill Librarians in Modern Marketing Solutions Curriculum*.

Contact: j-jovanovic@hotmail.com



RECOMMENDATION AND RANKING ALGORITHMS

Ljubiša Bojić & Zorica Dodevska

With the availability of big data for automated processing, the societal impact of recommendation and ranking algorithms is increasing. It is essential to affirm members of marginalized groups—historically discriminated against based on sensitive characteristics—in future algorithmic outcomes.

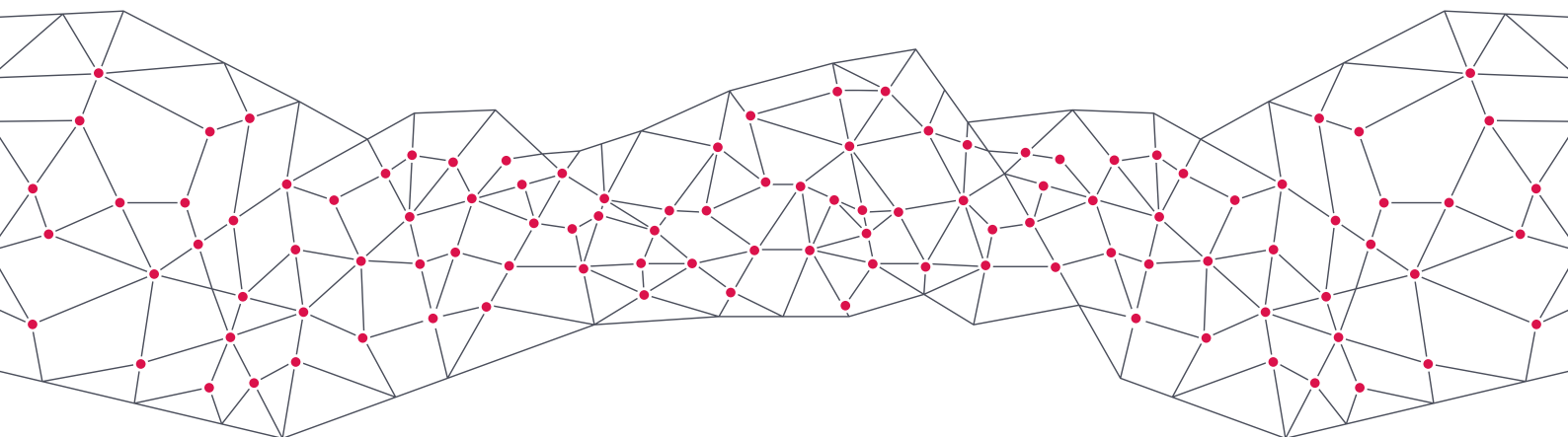
The use of AI-powered techniques in automated ranking mechanisms requires the adoption of adequate ethical guidelines to address the bias in AI-based systems widely applied across digital platforms. This involves focusing on fairness metrics and mitigating biases embedded in data or decision-making algorithms developed for recommendation and ranking purposes. Consequently, this call emphasizes the importance of preventing and correcting social bias in contemporary ranking algorithms.

Social initiatives aimed at protecting vulnerable groups, such as promoting gender equality or reducing inequality, provide an additional impetus for addressing this topic. Moreover, recent regulations emphasize users' rights to receive clear explanations for decisions made by algorithmic systems that affect them. Building a fair digital society requires robust technical support, which involves tackling the following key topics:

- Sources of Bias in Recommendation and Ranking Algorithms: This topic addresses the perspectives of stakeholders including

developers, users, and providers, to explain their mutual influences.

- **Social Antidiscrimination Frameworks and Technical Fairness Metrics:** This topic examines the integration of legal and social antidiscrimination initiatives with machine learning fairness metrics.
- **Machine Learning Techniques for Preventing and Reducing Algorithm Bias:** This topic highlights novel approaches in pre-processing, in-processing, and post-processing techniques, offering insights into the state-of-the-art literature.



Ethics Washing and the Value Measurement Problem in AI Alignment

Andrei Nutas

AI alignment, the endeavor to ensure that AI systems behave in ways consistent with human values and intentions, has gained significant attention in the last decade. This pursuit faces numerous challenges, chief among them the pervasive practice of ethics washing and the fundamental problem of value measurement.

A key issue in AI alignment lies in the profound differences between human and artificial intelligence. Humans operate on a complex interplay of emotions, experiences, and cultural contexts, while AI systems function on the basis of statistical patterns and programmed algorithms. This fundamental disparity creates a significant measurement gap: our existing tools and methodologies, designed to assess human values and behaviors, are insufficient for understanding the nuanced “value system” of an AI.

My previous research has shown that applying human-centric measurement tools, such as surveys or behavioral tests, to AI systems, reveals significant discrepancies in response patterns and distributions. For example, large language models exhibit response behaviors markedly different from those of human participants, characterized by reduced variance and more extreme outputs. These findings indicate that existing measurement paradigms are inadequate for accurately capturing the unique nature of AI cognition and decision-making processes.

This measurement gap is exacerbated by what I call the “tech-bro alignment problem.” Many AI alignment efforts are spearheaded by a homogeneous group of technologists who share similar cultural and educational background. This narrow perspective results in alignment strategies that reflect a limited subset of human values and experiences, potentially embedding biases and oversimplifications into AI systems.

The crux of the problem lies not just in the tech-centric approach, but in the severe underinvestment in the ethical dimensions of AI alignment. While billions are poured into advancing AI capabilities, comparatively

minuscule resources are allocated to ensuring these systems align with human values and ethical principles. This disparity is evident in both financial investments and human capital allocation. Tech companies and research institutions prioritize technical advancements, often treating ethical considerations as an afterthought or a public relations exercise.

This lack of investment is evident in multiple ways. Ethics teams, when they exist, are often understaffed and lack decision-making power within organizations. Ethical review processes are frequently superficial, conducted hastily to meet regulatory requirements rather than to genuinely scrutinize the moral implications of AI systems. Additionally, there is a significant lack of funding for interdisciplinary research that could connect technical AI development with ethical philosophy.

For tech companies to be credible in their proclaimed commitment to ethical AI, a fundamental shift in priorities is essential. The first and most crucial step is the development of robust measurement methods for alignment. Without accurate ways of assessing how AI systems interpret and act on ethical principles compared to their human counterparts, all other efforts at alignment remain speculative at best and dangerously misguided at worst. The investment in measurement methodology is not merely a technical challenge but a moral imperative. This requires significant resources, interdisciplinary collaboration, and a genuine effort to understand the unique cognitive processes of AI and how they can be bridged to align with human cognitive processes. Only through such dedicated initiatives can we move beyond superficial “tech-bro alignment” or ethics washing, achieving meaningful alignment between AI and human values.

Keywords: AI alignment; value measurement; ethics washing; large language models

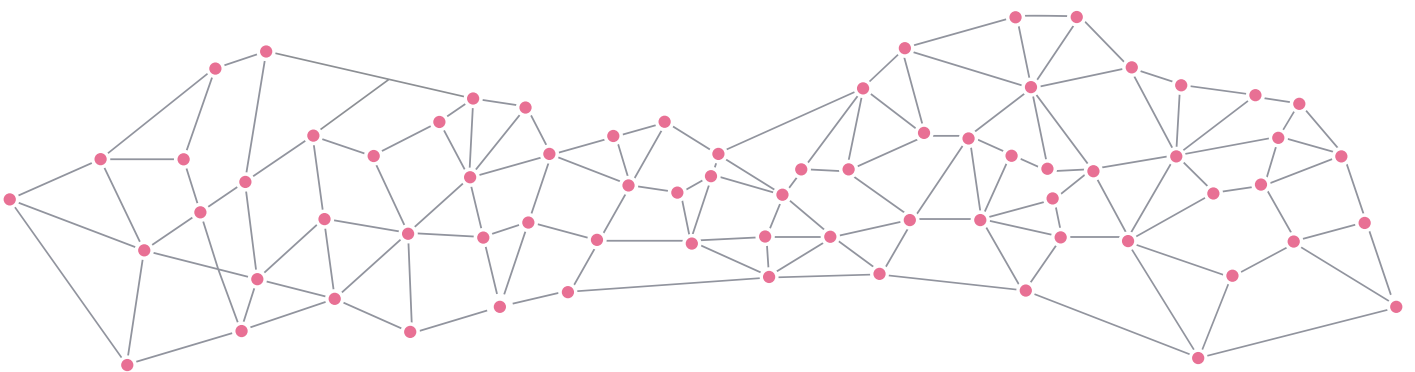
Andrei Nutas

Research Fellow

West University of Timisoara

Dr. Andrei Nuțas is a Euro-transhumanist and AI ethicist serving as a Research Fellow at the West University of Timișoara. He has delivered presentations on AI ethics at numerous academic and corporate conferences. Dr. Nuțas is also an active member of the “Value Connection” working group at Delft University, which employs a multidisciplinary approach to addressing the Value Alignment problem. This October, Trivent will publish his forthcoming book, *Artificial Morality*, which examines the possibility of an Artificial Moral Agency. His current research focuses on strengthening the naturalist foundations of Euro-transhumanism and investigating the decision-making capabilities of large language models (LLMs), both as standalone entities and in collaborative scenarios with humans.

Contact: andrei.nutas@gmail.com



Ethics in AI: A Bayesian Framework for Transparency and Accountability

Valentin Noël

In the rapidly evolving landscape of artificial intelligence, ensuring ethical alignment with human values has become critical. Ethics in AI can be broadly divided into three main categories: the ethics of datasets, the ethics of model training, and the ethics of results. Our goal is to develop a framework that transitions AI systems from opaque “black boxes” to more interpretable “gray boxes,” enabling greater oversight at every stage. This paper introduces a Bayesian framework to achieve this goal, which is adaptable across various fields and applications.

Parameter Interpretability and Fairness:

In a Bayesian framework, parameters such as weights and biases are represented by probability distributions, which include both a mean and a variance. This representation allows for deeper insights into the model’s behavior, as parameters with high variance signal areas of uncertainty, indicating crucial decision points. By identifying these influential parameters, researchers can link them to human decision-making processes, improving accountability and transparency.

Over time, these observations enable the creation of empirical laws governing model behavior, resulting in a reduced and more interpretable search space for parameter optimization. A smaller, more comprehensible search space makes it easier to supervise the model’s development, ensuring that fairness is maintained. This helps identify areas where biases could arise, creating opportunities to reduce the potential for unethical outcomes.

Dataset Ethics and Bias Detection:

Ensuring that datasets are collected ethically is necessary but insufficient for guaranteeing that they are free from bias. A systematic analysis of how different subsets of data influence the model’s parameters is crucial. Our framework proposes the continuous observation of how subsets affect parameter distributions and identifies where the largest shifts occur.

By iteratively mapping these shifts during multiple training cycles, the framework pinpoints specific types of data that disproportionately affect model behavior. This process reveals patterns of bias, helping researchers identify which data subsets are likely to produce unfair outcomes. Understanding these connections provides researchers with the knowledge to refine their datasets and models, ensuring that biases are reduced while maintaining transparency in model development.

Result Interpretation and Uncertainty Quantification:

AI-generated results are often presented deterministically, with little attention given to the uncertainties involved. This lack of uncertainty quantification can lead to overconfidence in AI systems and biased interpretations of their outputs. The Bayesian framework proposed here addresses this issue by quantifying uncertainties, giving researchers and users a clearer understanding of the model's confidence in its predictions.

By integrating uncertainty metrics into the results, the framework provides a more accurate understanding of AI outcomes. Recognizing the inherent uncertainty in AI predictions enhances accountability and ensures that decision-makers are fully aware of potential risks.

Conclusion:

This three-step Bayesian framework—addressing dataset ethics, parameter fairness, and result transparency—offers a comprehensive solution to the ethical challenges faced in AI development. By focusing on interpretability, bias detection, and uncertainty quantification, we promote fairness, transparency, and accountability at every stage of AI deployment. Furthermore, this approach is highly adaptable, making it applicable across a variety of fields, including healthcare, recommendation systems, and educational technologies.

Keywords: AI ethics; Bayesian framework; bias detection; uncertainty quantification; interpretability; dataset fairness; ethical AI deployment

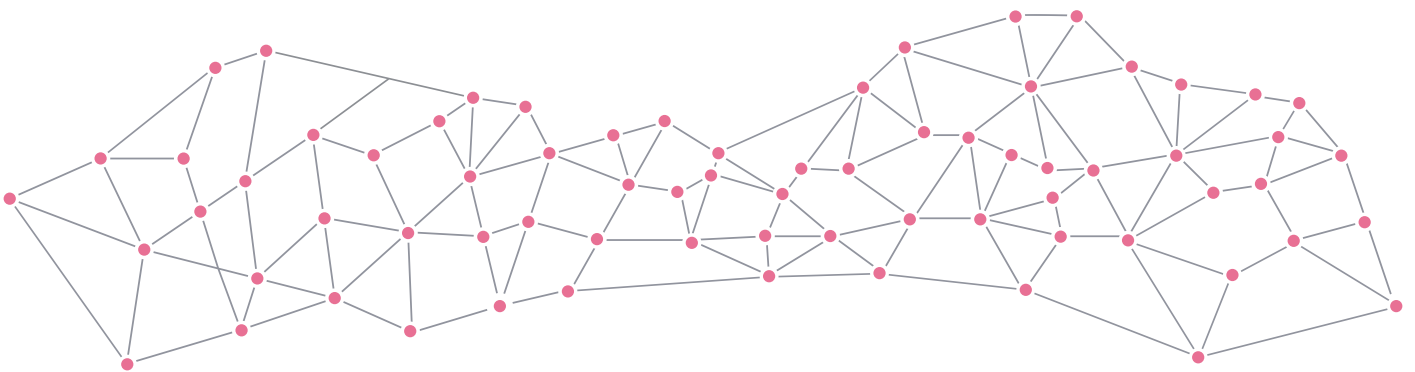
Valentin Noël

PhD Candidate

Université Paris-Saclay

Valentin Noël is a PhD student at Université Paris-Saclay, specializing in AI and Bayesian frameworks. His research focuses on improving the transparency and interpretability of AI models using Bayesian methods, with a particular focus on the ethical alignment of AI systems in healthcare. He has presented his work at several international conferences and he is also the recipient of the 1st prize Student Award at the 44th PhotonIcs and Electromagnetics Research Symposium—Progress In Electromagnetics Research Symposium (PIERS) in Prague in June 2023.

Contact: val.noel2369@gmail.com



Robot Autonomy and Responsibility

Miloš Agatonović

This presentation aims to demonstrate that autonomy in action requires self-location, making it possible to construct autonomous systems only if those systems are capable of self-localisation. The discussion begins with the fundamental problem of indexicals and explores the feasibility of robot autonomy. John Perry's thesis serves as a starting point, asserting that indexicals—terms like “I” and “here,” that refer directly to the speaker and one's location—are essential for action. Perry's work suggests that without the ability to use such terms, an agent cannot effectively engage in intentional behaviour. Building on Perry's thesis, the presentation introduces Jenann Ismael's conception of an agent as an information system. According to Ismael, an agent integrates indexical information, such as self-location, to connect an informational model with the environment. This connection is crucial for the agent to interact with its surroundings meaningfully. Ismael's framework implies that for robots to act autonomously, they must process and utilise indexical information to understand their place within a given context. One of the key conclusions drawn from this discussion is that the concepts of robot autonomy and agency are neither conceptually inconsistent nor unimaginable.

To illustrate this point, the presentation examines Paul Teller's example of robots autonomously following instructions based on third-person reports. Teller aims to show that genuine action can occur without the use of indexicals. However, Teller's setup, designed to enable robots to interpret third-person reports, inherently allows for self-location. Thus, even in Teller's example, the robots indirectly rely on indexical information to perform autonomous actions. The presentation further explores contemporary robotics, enhanced by artificial intelligence (AI), which aims to create robots as systems described by Ismael—capable of autonomously performing actions and self-locating.

Establishing the consistency of the idea of robot autonomy and agency raises a question about responsibility: if robots can function as autonomous agents, who is accountable for their actions? The question of responsibility is an ethical issue that has been widely debated among philosophers throughout history.

The presentation will highlight a response found in Schopenhauer's commentary on Augustine's theology, which identifies an inconsistency between the doctrine of predestination and the benevolence of God. Schopenhauer argued that the world's guilt and misery fall back on God, who, according to standard Christian theology, created everything and knew how everything would unfold. Drawing a parallel to robotics, the responsibility and guilt for autonomous robot actions fall back on their creators. These creators include not only those who constructed the robots but also those who programmed and trained them. This is especially pertinent in the case of robots based on reinforcement learning systems, where users play a significant role in shaping the robots' behaviour through training. Consequently, the bearers of responsibility and guilt extend beyond the hardware and software constructors to include the users who actively train and influence the robots' algorithms. The presentation underscores that different technologies implemented in robots could imply different bearers of responsibility and guilt.

Keywords: robot agency; self-location; AI; autonomy; responsibility

Miloš Agatonović

Senior lecturer

The Academy of Applied Preschool Teaching and Health Studies,
Preschool Teacher Training College

Miloš Agatonović (b. 1986) completed his undergraduate studies at the Faculty of Philosophy, University of Belgrade in 2012. Subsequently, he obtained his doctoral degree by successfully defending his dissertation titled *Nietzsche's Ethics and Critique of Morality* (original title: *Ničeova etika i kritika morala*) in 2017. Dr. Agatonović is currently a Senior Lecturer specializing in Philosophy at the Preschool Teacher Training College, Kruševac Section of The Academy of Applied Preschool Teaching and Health Studies. Since 2023, he has been a member of the Digital Society Lab (DigiLab) at the Institute for Philosophy and Social Theory of the University of Belgrade. Dr. Agatonović's scholarly contributions have been published in reputable journals such as *AI and Society* (Springer Nature), *Philosophy and Society* (Institute of Philosophy and Social Science, University of Belgrade), *Balkan Journal of Philosophy* (Bulgarian Academy of Sciences), and *Theoria* (Serbian Philosophical Society). His dissertation-based book is currently in production with Springer Nature. His research spans a range of areas, including Nietzsche's philosophy, ethics, AI ethics, philosophy of science, philosophy of media and technology, and transhumanism.

Contact: milos.agatonovic@vaspks.edu.rs

Automated Action Classification and Threat Prediction in Video Streams

Oleslav Antamoshkin

The purpose of this study is to develop a system capable of classifying human actions in video stream images and predicting the probability of these actions posing a threat to others. The main objective of the system is to enhance safety levels in public places and industrial sites by using machine learning methods to analyze video streams.

The study aims to create a software product that will automatically detect human actions based on input images and predict the likelihood of these actions being a threat. The key advantage of the system lies in its automated neural network model setup, which eliminates the possibility of users deliberately introducing false information.

The key research questions include: How can the accuracy of action classification and threat prediction based on video stream images be improved? How can human involvement in the system setup process be minimized to avoid human error? What architectural solutions can reduce the memory and computational resource requirements of the system?

The developed system is expected to significantly enhance operational control over ongoing events, reduce the workload of security personnel, and provide timely notifications of potential threats.

Keywords: action classification; video stream; threat prediction; machine learning; safety

Oleslav Antamoshkin

Head of the Department of Software Engineering. Institute of Space and Information Technologies; Professor at the Department of Information Technologies in Creative and Cultural Industries, Institute of Humanities

Siberian Federal University

Oleslav Antamoshkin holds a Doctor of Technical Sciences degree in Information Technologies and has been engaged in research and teaching for over 20 years. His research focuses on optimizing the management of hardware and software resources in distributed computing systems, as well as methodologies for integrating artificial intelligence technologies into various aspects of human life. Oleslav actively participates in scientific projects aimed at improving the efficiency of information technologies in various industries and security sectors. His research has been published in leading Russian and international journals. He is the author of several textbooks for students and graduate students and he teaches courses in software engineering, machine learning, and data analysis. Throughout his academic career, Antamoshkin has supervised numerous student and graduate theses, many of which have won prestigious awards in scientific competitions. In addition to his academic work, Oleslav actively collaborates with industry partners in implementing innovative software solutions.

Contact: oantamoskin@sfu-kras.ru

HEALTH-TECH AND HEALTH LITERACY IN THE CONTEXT OF GENERATIVE AI

Ljubiša Bojić

This portion of the EMERGE 2024 conference seeks to explore questions of Health Tech and Health Literacy within the broader purview of Generative AI. This two-themed conversation is designed to dissect the impact of AI in healthcare and its potential in revolutionizing health literacy, playing a significant role in better health outcomes.

AI and machine learning have the potential to reform healthcare by aiding in accurate and rapid disease diagnosis, developing personalized treatments, creating effective drug development strategies, and enhancing patient care. Yet, this promising horizon teems with critical ethical questions ranging from data privacy and algorithmic bias to the transparency and autonomy of AI decisions in healthcare.

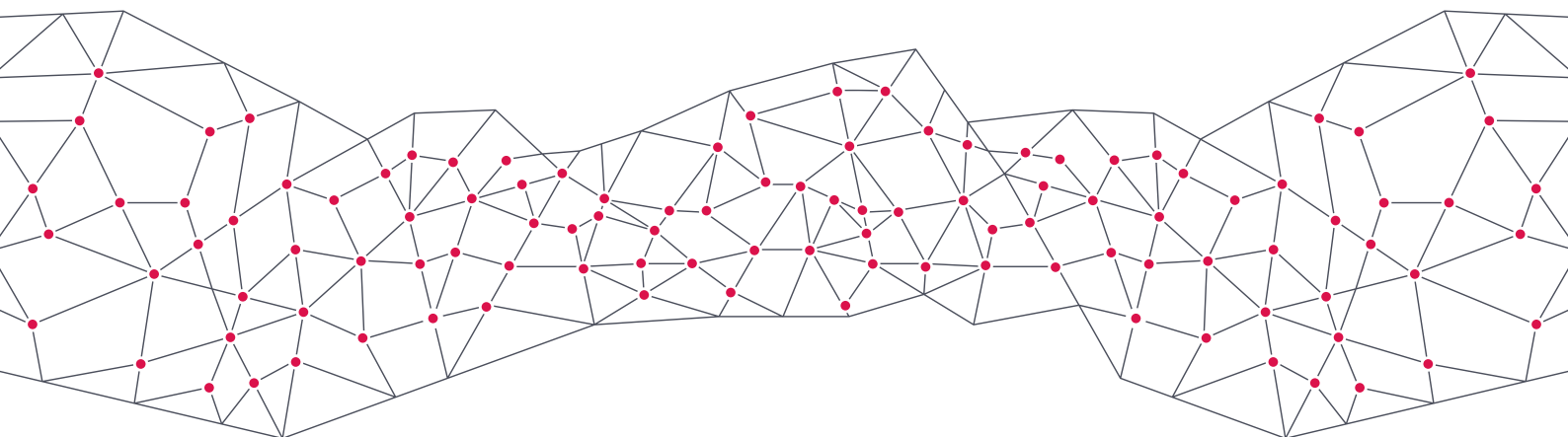
Generative AI opens a vibrant avenue for elevating health literacy. By creating text that closely mimics human-made text, Generative AI could potentially facilitate much-needed comprehension in health information, fostering a population of informed healthcare consumers. This shift towards personalized health information can revolutionize patient autonomy and decision-making.

Key discussion areas will encompass:

- **Data Protection & Privacy:** Balancing the benefits of Generative AI in personalizing health information while safeguarding sensitive

health data.

- Bias & Fairness: Examining the potential biases in AI algorithms and their impact on equal healthcare services.
- Transparency & Explainability: Enhancing the interpretability of AI diagnoses for healthcare professionals and patients.
- Autonomy & Responsibility: The ethical balance between AI-driven health technologies and the human element in healthcare.
- Health Misinformation: The ethical implications of AI's role in both creating and combatting health-related misinformation.
- Accessibility of Complex Health Information: The ethical aspects of AI's potential to render complex health information into understandable content for the public.



Emotional Quantification and Philosophical Thinking: Understanding Ourselves Better with Emotion Recognition Technology

Alexandra Prigent

Among the new and emerging technologies is emotion recognition technology (ERT). ERTs aim at accessing and identifying people's inner affective states. While often discussed in the context of surveillance and security, with a strong emphasis on the dangers posed by such technology, it is interesting to note that much of the development of ERTs has actually taken place in health and education contexts, where ERTs are used as a tool to help individuals better understand, and sometimes regulate, their own affective states.

As emotions play an important role in everyday life—from human interaction to decision making—I propose to analyse the social implications of using such technologies in two key areas: 1) current mechanisms employed in interactions to regulate privacy boundaries, and 2) trust. Employing an empirical philosophical methodology, this analysis is grounded in the scientific findings of both affective computing and social psychology research.

The paper is structured as follows: I begin by reviewing state-of-the-art literature on ERTs, both from affective computing and psychological research, highlighting their current results, technical limitations, and briefly predicting the potential new breakthroughs.

Further, I develop the analysis by establishing the role of emotions in social life. I argue that emotional expressions serve as critical communication channels that sustain the social fabric by providing reliable social information. These expressions help us understand, predict, and anticipate others' behavior, even though the information conveyed is probabilistic and influenced by factors such as context, the observer's knowledge of the individual, and the quality of perception.

Finally, I delve into the core of the argument, introducing privacy and trust as interdependent social phenomena whose balance is a necessary condition for successful interaction. I argue that the use of ERTs disrupts these communication channels—reliant on emotional expressions—by undermining certain privacy mechanisms and

substituting trust in social interactions.

Keywords: information ethics; education; healthcare; emotional expressions; emerging technology

Alexandra Prégent

PhD student

Leiden University

Alexandra Prégent is currently a PhD student at Leiden University (NL), supervised by Prof. Dorota Mokrosinska and Prof. James McAllister, working on forecasting the social impacts of emotion recognition technology. My research generally addresses social problems that arise from affective computing, with a particular focus on information ethics, privacy, and philosophy of the (affective) mind. My main objective is to contribute to the responsible design, deployment, and regulation of emerging technology. To this end, I adopt a practical philosophical approach, grounded in empirical evidence and enriched by insights from social psychology, computer science, and legal scholarship, to guide and inform my philosophical inquiries.

Contact: a.pregent@phil.leidenuniv.nl

AI Applications and Ethical Considerations: A Cross-field Analysis of Recent Trends in Engineering, Natural Sciences, Language Sciences, and Medicine

Vladimir Otašević, Jelena Lazić, Nikola Janković, Milan Jovanović

In recent years, alongside advancements of modern technological innovations, the use of artificial intelligence (AI) has significantly expanded. While AI finds applications across various scientific fields, certain concerns are universally shared. The widespread accessibility and ease of AI implementation raise critical questions about its appropriate use.

AI ethics in engineering raises concerns whenever AI systems are applied in areas such as monitoring, automation, and decision-making. Concerns surrounding AI ethics include issues of fairness, transparency, and accountability. Engineers must design systems that minimize biases and uphold responsibility. Ethical guidelines should evolve alongside advancements in AI technologies, such as deep learning, neuromorphic computing, and edge AI, to ensure innovation in fields like computer engineering aligns with societal values. Integrating ethics into the AI development process is crucial for fostering responsible applications in engineering, ensuring that AI systems are not both effective and ethically sound.

The application of AI in natural sciences can be categorized into two primary branches. First, the existing AI-powered simulations are used to analyze available data and enhance current designs, particularly in material science. AI can be trained on appropriate datasets to predict material models, which must then be evaluated for sustainability and practicality. Second, AI serves as a powerful search engine and data compiler, capable of advancing data gathering and analysis. However, researches should remain cautious of the misconception that AI is infallible and should refrain from its overuse.

Interest in AI applications in medicine began in the early 2000s, primarily for analyzing patient records and data. The development of convolutional neural networks expanded AI's role, making medical imaging a prominent area of application. During this period, various ethical concerns emerged. One of the main issues was the trustworthiness of AI algorithms, particularly less explainable

models like deep neural networks. Effective implementation requires additional education for medical personnel. In biomedical and genomic research, AI could optimize individual treatments, such as cancer therapy. However, biased datasets could lead to potentially harmful decisions. AI can also be used to assess public opinion on health-related issues, as seen during the COVID-19 pandemic. In such applications, data must be protected from misuse by unauthorized third parties.

In language sciences, following the eras of rule-based techniques and statistics-based language models, the emergence of deep learning approaches and neural language models in the 2010s brought significant advancements to natural language processing (NLP). These developments have greatly improved text processing, speech recognition, speech synthesis, machine translation, text summarization, information retrieval, sentiment analysis, and related fields. With the advent of large language models like GPT-4, these technologies have enabled new applications, such as automatic content generation and language learning assistants. However, they have also introduced significant safety and ethical challenges, including issues related to data access, privacy, and quality; intellectual property rights; social manipulation and surveillance; identity theft; bias; widening socioeconomic inequalities; and energy efficiency. Addressing these challenges requires a multifaceted approach.

In conclusion, it is important to acknowledge that this work provides only a brief overview of recent trends in AI literature and AI applications across various fields and the ethical considerations they raise. A more comprehensive analysis would require reviewing additional literature, datasets, and patents from both academia and industry, as well as dedicating more time and effort. We hope this brief overview highlights at least some of the advancements in this area of research. We eagerly anticipate the uncertain direction AI will take in the upcoming years.

Keywords: AI Ethics; responsible AI development; bias in AI; natural language processing; AI in medicine

Vladimir Otašević

Junior Research Assistant

University of Belgrade - School of Electrical Engineering

Vladimir Otašević is a software engineer with a strong focus on the application of artificial intelligence and machine learning, particularly in scholarly communication. He is committed to creating innovative solutions for the scientific community. At the University of Belgrade Computer Center (RCUB), he is involved in implementing and maintaining institutional information systems and developing standalone tools to support the academic community. His active participation in national and international projects and groups reflects a strong commitment to promoting Open Science in Serbia.

Contact: vladimir.otasevic@rcub.bg.ac.rs

Jelena Lazić

PhD student

University of Belgrade - School of Electrical Engineering

Jelena Lazić is a PhD student in the field of Artificial Intelligence at the University of Belgrade, where her research focuses on speech analysis and synthesis. She is interested in enhancing speech technologies through innovative approaches. Previously, she worked as a research fellow at Rice University and is currently involved in a research project at Tohoku University, both in the field of AI applications in medicine. These international experiences have allowed her to engage with diverse cultures, connect with like-minded individuals, and tackle challenging research topics. Outside of her academic work, she is a problem solver at heart and an optimist by nature.

Contact: lazic.jelena@gmail.com

Nikola Janković

PhD student

University of Belgrade, Faculty of Philology

Nikola Janković is a PhD student at the University of Belgrade, Faculty of Philology, where his doctoral dissertation is related to the annotation and analysis of the ITALSERB learner corpus of Italian language. He is also engaged in the creation of parallel corpora and the development of high-quality datasets for training large language models. Currently, he is employed in the risk assessment field as a Python programmer. He is passionate about contributing to open-source software and collaborating on solutions relevant to the public good.

Contact: nikolajankovickv@gmail.com

Milan Jovanović

PhD student

University of Criminal Investigation and Police Studies

Milan Jovanović is a PhD student in the field of Forensic Engineering at the University of Criminalistic and Police Studies in Belgrade, focusing on material science. He is also interested in learning about the rapidly-developing field of artificial intelligence, with plans to integrate it in his future research. He is dedicated to scientific collaboration and focused on cooperating with experts and PhD students from other scientific fields.

Contact: mjnmilan9@gmail.com

MEDIA, FREEDOM OF EXPRESSION, AND DEMOCRACY

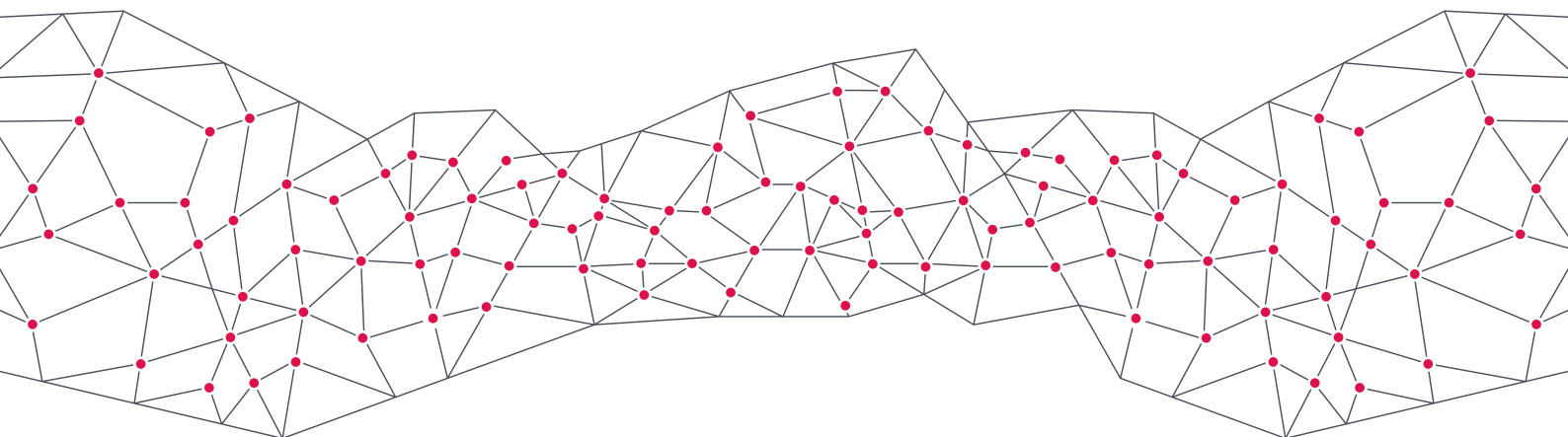
Čedomir Markov

Global democracy is in decline, a process in no small part exacerbated by the spread of online misinformation. As AI-powered technologies continue to rapidly evolve, their intersection with democracy emerges as a critical area for exploration and ethical scrutiny. On the one hand, there are numerous examples of how AI and related technologies can be leveraged to revitalize democracy. Social media algorithms, for instance, could promote public discussion that favors facts and reasoned debate over exploiting emotions and fueling polarization. The potential of blockchain technology in countering mis- and disinformation has recently been widely discussed and explored. Generative AI could serve as a tool for moderating discussions, scaling up deliberative dialogues, and fostering consensus.

However, the promises of AI are frequently overshadowed by growing concerns about their potential to further deteriorate democracy. The proliferation of deepfakes, computational propaganda, and automated astroturfing highlights how AI can magnify the impact of online misinformation on political knowledge and preferences. Microtargeting and voter profiling remain prime concerns for voter manipulation in the face of a critical election year across the world.

This session aims to explore how AI-powered technologies can be

integrated into democratic frameworks ethically and effectively to promote inclusivity, fairness, and the collective good, thus aligning the digital sphere with our shared democratic ideals. We are especially interested in contributions that examine how AI can influence electoral processes, public opinion formation, and the broader civic engagement landscape. Contributions may cover topics such as algorithmic transparency, AI-driven misinformation, the role of AI in enhancing or undermining democratic participation, and strategies for aligning AI development with democratic values and human rights.



Informatization vs. Digitalization: Different Approaches to Governance Transformation

Alois Paulin

The twentieth century introduced humanity to radically new knowledge through advancements in electronics, informatics, and telecommunication technologies. These developments gave rise to cyberspace, serving as a medium for data exchange, data storage, and enabling innovative approaches to managing and controlling real-world systems. These newly created opportunities have been successfully deployed for the automation of processes in areas such as production, service provision, data exchange, navigation, and logistics. Additionally, they have enabled the development of new possibilities through concepts such as virtualization. While this has led to radical transformations of paradigms in industry and free market service provision, the systems that make up modern states have been broadly spared of disruption by these technologies. Behind this backdrop, this contribution aims to discuss the differences between digitalization and informatization as two differing approaches to system transformation. The discussion is set in the context of societal governance, where digitalization is the main approach to modernisation. The emphasis on digitalization and the insufficient progress toward informatization in this field are criticized, with attention drawn to the advantages that informatization can offer.

Keywords: digitalization; public governance; e-government feudalization

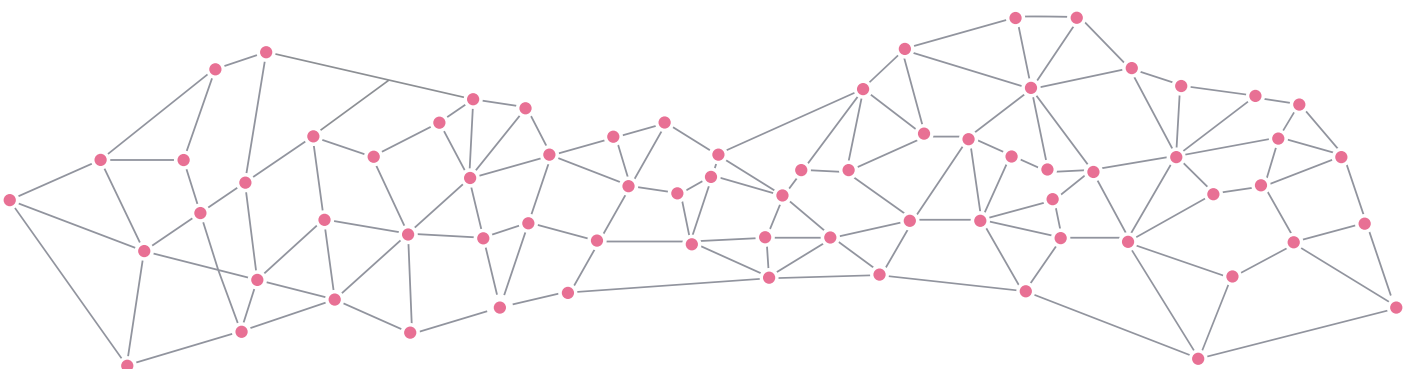
Alois Paulin

Professor

University of Public Administration and Finance Ludwigsburg

Dr. Alois Paulin is Professor of Digital Innovation and Transformation in Public Administration at the HVF Ludwigsburg. He holds a Doctor of Science degree in CS & Informatics from the University of Maribor, Slovenia. His research interest is in public domain governance, sustainable government information systems, and democratic collaborative decision-making. He is the author of *Smart City Governance* (Elsevier, 2018) and lead editor of *Beyond Bureaucracy* (Springer, 2017). He chairs the Beyond Bureaucracy track at the Annual International Conference on Digital Government Research (dg.o).

Contact: alois.paulin@hs-ludwigsburg.de



AI and the Political: The Rise of the Providential Machines?

Blagovesta Nikolova

The title of the presentation represents a benign intellectual tease with the current media obsession with phrases like “the age of intelligent machines,” “the age of the thinking machines,” and “the rise of the intelligent machines.” These and other similar expressions very often impact the public imagination and nurture a sense of the technological Second Coming that questions our understanding of what constitutes “human,” “political,” and “societal.” I, in turn, adopt this popular framework of articulating the problems we experience in the whirlwind of the stormy dance between the digital and the physically tangible world, seeking to theorize the meaning and role of AI in contemporary societies from a politico-philosophical perspective.

The talk will specifically address the observation that many narratives about the present and future of artificial intelligence tend to convey a sense of the providential nature of the processes concealed within the “black box” of this technology. “Providential” here refers to processes that are foreign to our cognitive constitution, such as perceiving, establishing correlations between data, making inferences, and arriving at decisions. A reason of a different nature—an epistemologically alien intelligence, to borrow the phrase from the esteemed Harari—an intelligence that claims the ability to point to the truth or predict the future. This inevitably reminds us of different instances of resorting to Providence in the long history of political ideas in the West. Evoking the power of Providence is a familiar modern gesture, rooted in pre-modern Christian traditions, aimed at identifying a reliable social regulator that imposes a specific order and imparts meaning to the complex mechanisms of social and historical change (e.g., Joseph de Maistre, Alexis de Tocqueville, etc.).

The contemporary faith in the power of AI appears to revive this hope for a transcendental mechanism—one that is better attuned to the intricate systems we inhabit and capable of sanctioning our actions based on reasoning that lies beyond the reach of human understanding. This has profound implications for the collective understanding of “the political” and necessitates a reevaluation of the technologies that shape it. The intervention of AI in the sphere of political cognition changes not only the mechanisms of political

decision-making but also the mere notion of political responsibility.

In sum, the presentation will try to argue the politico-theoretical significance of AI's rise and outline the stakes of advancing projects that promote the notion of "providential machines."

Keywords: political theory; artificial intelligence; providence

Blagovesta Nikolova

Associate Professor

Institute of Philosophy and Sociology, Bulgarian Academy of Sciences

Blagovesta Nikolova obtained her PhD in Political Philosophy in 2013 at the Bulgarian Academy of Sciences. She has been working and publishing in two thematic directions—the transformation of foresight under the burden of increasing uncertainty and the prospects of ethical governance of new and emerging technologies. Those two considerations conflate in her monograph *The RRI Challenge: Responsibilization in a State of Tension with Market Regulation* (2019), where she explores the EU-promoted concept of Responsible Research and Innovation and traces the built-in controversies and impediments before its meaningful implementation. Since 2016, Dr. Nikolova has been acting as an ethics expert for REA and HADEA in evaluating research projects.

Contact: blagy_ilieva@abv.bg

AI and the Future of Capitalism: Revisiting the Socialist Calculation Debate

Dmitrii Trubnikov

The contemporary technological developments in the areas of AI and data analytics openly challenge the foundational stones of modern capitalist societies, providing new arguments to supporters of socialism. This has revived the socialist calculation debate, a noticeable dispute in the history of economic thought that attracted significant attention in the first part of the 20th century. On the one side of the debate were the economists of the Austrian school of economics, while their main opponents represented various philosophical and social science traditions, ranging from Marxists to neoclassical economists.

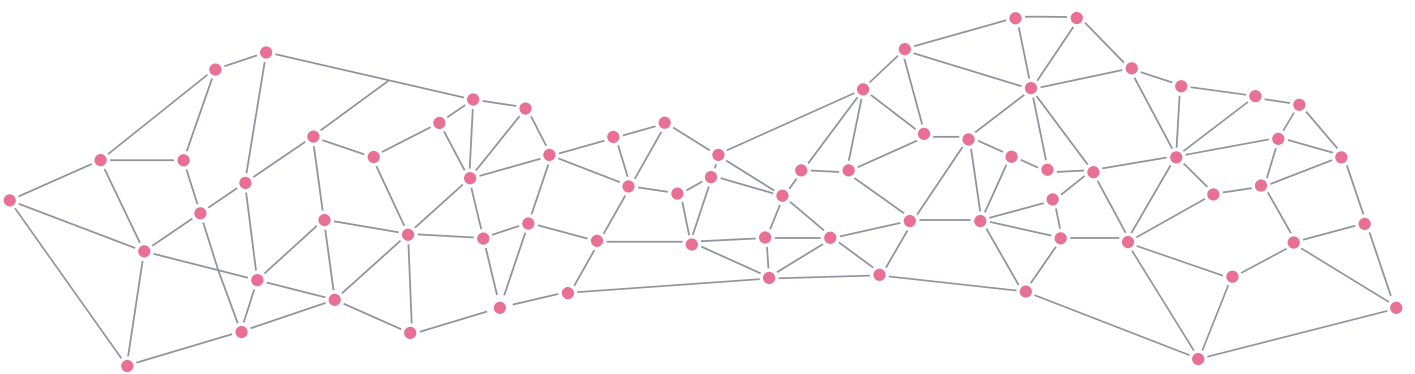
During the classical period of the debate, the Austrians have proposed different arguments that could be classified into two main lines: “impossibility” of socialist calculation and “knowledge problem.” Their modern opponents argue that the new technological capabilities mean not only faster calculations but the possibility to successfully deal with these obstacles. Moreover, some socialist authors have even endorsed the Austrian vision but found ways to use it against the original Austrian standpoint.

The modern response from the pro-market camp is not very different from its classical roots. The fundamental assertion is that the socialist calculation problem has no practical solution in any technological paradigm. The Austrians have always embraced the technology argument, maintaining the belief that their reasoning is universal and technologically neutral. If the hand-mill gave rise to feudal society and the steam-mill to capitalism, a modern extension of this reasoning might suggest that AI and other advancements in digital technology are steering the world toward socialism. According to their perspective, the modern revival of the debate is just an expected outcome of the new technological epoch, rather than a development that was not already accounted for in earlier analyses.

The present research has two main objectives. First, it seeks to evaluate the challenges to the Austrian position, specifically examining whether AI solutions could theoretically enable a technosocialist economy, all while addressing the concerns raised by Austrian theorists. Second, it

aims to redirect the focus of the debate. While the classical approach assumed that the ends of economic calculation should not be questioned, it must be acknowledged that the new technological advancements critically challenge this foundation. It has always been futile to analyze the possibility of achieving maximum results from limited resources without first understanding what specific outcomes are being sought. However, what technological advancements can change is delegating AI systems not only the calculation task, but the determination of the goals of these calculations as well. This, in turn, raises a number of ethical questions about the adherence to democratic values. Approaching the problem from this perspective not only redirects the discussion but also shifts it further away from the domain of economics and into the realm of political and social philosophy.

Keywords: technosocialism; socialist calculation debate; Austrian school



Dmitrii Trubnikov

Visiting Fellow

Institute for Philosophy and Social Theory, University of Belgrade

Dmitrii Trubnikov holds a joint PhD in Law from Tilburg University and Bologna University, as well as a Candidate of Economic Sciences degree (the Russian equivalent of a PhD in Economics) from Samara State University of Economics. From 2019 to 2022, he served as an Associate Professor in the Department of Management at the Saint Petersburg School of Economics and Management, HSE University (Russia), where he also held the position of Academic Director for the Master in International Business program. Dmitrii's research has focused on the regulation and impact of ICT, the application of public choice theory to entrepreneurship and innovation, and the contemporary relevance of ordoliberal philosophy. His work has been published in prominent journals, including *Minerva*, *Economic Affairs*, *Journal of Business Venturing Insights*, *Journal of Industry, Competition and Trade*, and *World Economy and International Relations*. Currently, Dmitrii is a Visiting Fellow at the Institute for Philosophy and Social Theory, University of Belgrade, and a member of the Independent Institute of Philosophy association in Paris, France.

Contact: da.trubnikov@gmail.com

Democracy Values Threatened by Captological Tools

Katarina Šmakić

In the digital age, democracy faces unprecedented challenges, due to the rise of captological tools—technologies designed to manipulate user behavior. Once celebrated as bastions of free speech and civic engagement, digital platforms have gradually transformed into tools for controlling and manipulating public opinion. Captology, the study of computers as persuasive technologies, exploits psychological triggers to steer user decisions, often prioritizing corporate or political interests over democratic values. This paper examines the tension between digital democracy and captological tools, highlighting how algorithms, targeted advertisements, and data mining threaten the foundations of informed consent, privacy, and freedom of expression. By exploring these dynamics, the paper aims to highlight the urgent need for regulatory frameworks that protect democratic principles in the face of growing technological influence.

Keywords: democracy; persuasive technologies; captology; data privacy; algorithmic influence

Katarina Šmakić

Assistant Professor

Faculty of Diplomacy and Security, University Union Nikola Tesla

Katarina Šmakić, born in 1979 in Belgrade, graduated in English Language and Literature from the Faculty of Philology. She earned her Master's degree in Theory of Art and Media from the Interdisciplinary Studies program at the University of Arts in Belgrade in 2008, where she also completed her Ph.D. in 2016. She has extensive academic and professional experience in the fields of language, literature, art, and media. Her contribution to international cooperation and participation in EU-funded projects is particularly noteworthy. In these projects, she worked on curriculum development, the integration of new technologies into educational processes, and the enhancement of collaboration between universities and industry. Her active involvement in the accreditation of institutions and study programs demonstrates her expertise and dedication to advancing the education system in Serbia. Her academic work focuses on philosophy and media theory, with particular emphasis on media literacy, digital technologies, and the analysis of contemporary media phenomena. A sought-after speaker, she frequently appears in media and participates in panels addressing current media issues. Her scholarly works are widely cited and often serve as reference material for students. She is dedicated to exploring media phenomena, the metalanguage of digital media, and creative expression through digital technology. In recent years, her research has focused intensively on digital world phenomena and their broader societal impacts.

Contact: katarinasmakic@gmail.com

Generative AI, Political Communication, and Democracy: Does Generative AI Pose a Significantly Different Risk than Standard AI on Democracy?

Maria Zanzotto

In recent years, concern about the impact of technology, specifically artificial intelligence, (AI), on democratic institutions have been growing (Coeckelbergh, 2024). In particular, the literature highlights three AI-enabled phenomena on social media—disinformation, polarization, and manipulation—that pose significant threats to democracy. However, with the rapid proliferation of generative AI tools capable of producing visual content (e.g., DALL-E, Midjourney, Stable Diffusion) and textual content (e.g., ChatGPT, Claude, Llama), the scope of the risks has significantly expanded.

The aim of this paper is to explore how generative AI technologies impact democracy with regard to disinformation, polarization and manipulation on social media, and to analyze them from a political philosophy perspective (Coeckelbergh, 2022). The paper focuses on the epistemic environment in which people form their political beliefs (Levy, 2021), adopting a thin conception of democracy for this purpose. The paper argues that generative AI technologies exacerbate issues of disinformation and, to some extent, polarization by enabling the large-scale creation of false information. However, this change is quantitative.

In contrast, when it comes to manipulation, generative AI technologies appear to pose a qualitative risk compared to recommendation systems: firstly, they unlock new microtargeting possibilities (Matz et al., 2023; Simchon et al., 2023, 2024) and secondly, they are able to produce content almost indistinguishable from human made content (Wilson, 2017).

The paper builds on Ienca's (2023) concept of manipulation, exploring this issue on two levels: first, by examining how non-transparency applies to generative AI systems, and second, by investigating whether interactions with AI-generated content can diminish personal autonomy, drawing on Coeckelbergh's (2023) work on epistemic agency. It emphasizes the necessity of extending research on the

impact of AI on social media and democracy to include generative AI technologies.

Keywords: generative AI; democracy; political communication; social media

Maria Zanzotto

PhD student

Department of Philosophy and Education Sciences, University of Turin

Maria Zanzotto is a second-year PhD student at the University of Turin and the FINO convention. She is a visiting PhD student at the University of Vienna for the Winter Semester 2024. After earning a BA in Economics from the University Ca' Foscari in Venice, she completed an MA in Philosophy at the University of Milan and Vita-Salute San Raffaele University in Milan. Her research focuses on Ethics and Artificial Intelligence, with a particular emphasis on Large Language Models. Her PhD project examines the differing impacts of generative AI compared to non-generative AI at ethical, political, and economic levels, with a specific focus on the issue of manipulation.

Contact: maria.zanzotto@unito.it

AI and Global Democracy: Signals from Global Debates

Tatjana Milić

Artificial Intelligence (AI) is a transformative technology reshaping the world in countless ways. While it offers benefits that transcend national borders, it also poses threats that may surpass what we can currently imagine or comprehend. If these threats were to be addressed solely through the efforts and preferences of individual states, global democracy could be at risk. Recognizing this, the United Nations (UN), as a global organization, has taken steps to outline the principles and frameworks for governing the development of AI. Although the topic of AI has been addressed in several specialized UN agencies (WIPO, UNESCO, ILO, etc.), indispensable actors in the process of the establishment of effective universal AI governance are the General Assembly and Security Council.

Given that the powers of these two UN bodies, and their interrelations, reflect structural dynamics that challenge global democracy, this paper examines whether the pursuit of a universal AI governance framework has advanced global democracy or merely perpetuated existing and introduced new inequalities that conflict with the democratic principles upheld by the international community. To answer the research question, the study focuses on the content of global debates on AI technologies held at the General Assembly and Security Council.

The empirical basis of this research is formed by statements from state representatives expressed during these debates, as well as the resulting decisions made by the respective UN bodies. The study employs the perspective of Critical Discourse Analysis (CDA). is employed to analyze the discourse of debates on AI, uncovering underlying ideas, ethical considerations, and power relations. This approach not only sheds light on the formation of a global framework for aligning AI development with democratic values and human rights but also reveals the structural dynamics within the UN itself.

The paper begins by examining how states from various groups (e.g., Global North vs. Global South, East vs. West, developed vs. developing) perceive the impact of AI on the respect and protection of democratic values and human rights. This step is taken to examine

the first research hypothesis, which assumes that understanding the interaction between AI development and democracy varies across different contexts and reveals inequalities that challenge global democracy.

Next, the paper investigates states' views of the roles of the General Assembly and Security Council in governing the development of AI. This step explores the second research hypothesis, according to which the process of forming a governance framework for AI alignment reflects how much structural tensions between the General Assembly and Security Council contest principles of global democracy and diminish the organization's power to effectively address challenges posed by AI development.

In this context, the study considers that the unprecedented challenge posed by rapidly evolving AI technology could push the UN to initiate structural changes aimed at achieving a more democratic balance within the global organization. Signals from global debates on AI highlight the roadmap the UN is following in its quest to establish a governance framework that ensures AI technological advancements do not threaten global democracy. These debates reveal the power dynamics underlying global discussions on AI while also uncovering the ideas and ethical considerations that support universal safeguards for global democracy. More importantly, they provide a foundation to ensure that future AI developments are guided toward benefiting humanity and serving the common good.

Keywords: AI; global democracy; United Nations; critical discourse analysis; governance

Tatjana Milić

Independent Researcher

PhD in Political Sciences, University of Belgrade

Tatjana Milić holds a PhD in Political Sciences from the Department of International Studies, Faculty of Political Sciences, University of Belgrade. In 2023, she was awarded the scientific title of Research Fellow by the Ministry of Science, Technological Development, and Innovation. She is currently working as a Project Implementation Analyst for the USAID Venture an Idea Project, Digital Serbia Initiative. Her research focuses on legal regulation of international relations and international organizations. She gained professional experience through research projects in the non-governmental sector and academia. Her academic roles have included Researcher at the Faculty of Law, University of Novi Sad, supported by a Research Fellowship from the Provincial Secretariat for Higher Education and Science (2018/19), and Teaching and Research Associate at the Faculty of Political Sciences, University of Belgrade (2004–2007). She completed an internship programme at the Institute of International Law and International Relations, University of Graz (Scholarship of the Austrian Development Cooperation and the Dr Zoran Đinđić Fund). Tatjana has published articles in scientific journals and presented papers at international scientific conferences. Detailed insight into her research work is available at ORCID.

Contact: tatjana.milic@gmail.com

Ethical Considerations of AI in Journalism: The Perspective of Journalism Students

Aleksandra Krstić, Marko Nedeljković

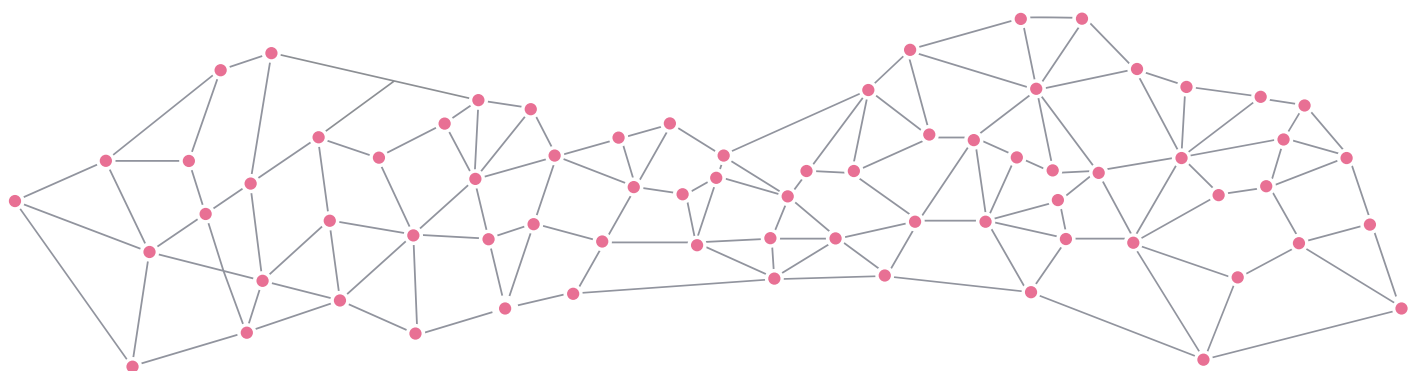
Journalists' perceptions of Artificial Intelligence (AI) are shaped by the evolving role of AI in media industry. Many journalists worldwide recognize the potential of AI to automate writing of basic news reports and other routine tasks, such as fact-checking, data analysis, uncovering trends, and getting insights that would be difficult to achieve manually. This allows journalists to focus on more complex and creative aspects of their work. At the same time, there is a concern that AI could replace human journalists, particularly in roles that involve routine reporting or editing. Similar concerns are shared by journalism students around the world, who are often skeptical about the long-term implications of AI for employment in the industry, the accuracy of AI-generated content, and the ethical use of AI. Recent studies show that journalism students are particularly concerned about bias in algorithms, spread of misinformation and disinformation, and the compromised use of AI that might put journalistic integrity at risk. Studies and surveys of journalism students' attitudes toward AI often reveal a range of opinions, influenced by factors such as their exposure to technology, educational environment, and personal experiences with AI tools. Their perspectives on AI's application in their future profession are especially significant, as they represent the next generation of journalists who will not only be more aware of AI's potential and risks but also more directly exposed to its use in their everyday work.

Against this background, this paper examines the perceptions of journalism students at the University of Belgrade about the application of the AI in the journalistic profession and their awareness of the AI trends and impacts. For that purpose, we conducted the research by surveying 204 students of the Faculty of Political Science in Belgrade in their final year of journalism (fourth year and seniors) using a written questionnaire combining closed and open questions with the assessment scales.

The results show that future media professionals are well aware of the impact that Artificial Intelligence has on the profession of journalism. Vast majority of respondents believe that AI has a large

or extremely large impact on journalism, while only a small number of them recognize this impact as small or limited. At the same time, results show that there has been no respondent who believes that artificial intelligence has no impact on modern journalism. Although the research reveals that respondents also recognize positive effects of AI on modern journalism, they still rate this influence as dominantly negative. Students recognize the absence of journalistic ethics and professional responsibility, the unreliability of media content produced by AI, as well as its insufficient originality as the most problematic obstacles to apply AI in their everyday journalistic work.

Keywords: journalism students; artificial intelligence; journalistic ethics; perceptions; Serbia



Aleksandra Krstić

Associate Professor

University of Belgrade, Faculty of Political Science

Aleksandra Krstić is an Associate Professor at the Department for Journalism and Communication, Faculty of Political Science, University of Belgrade. Her research interests are media and journalism studies, TV journalism, visual communication, media ethics, democratization of media, safety of journalists, conflict reporting, and media-politics relations. Dr. Krstić is the President of the Serbian Political Science Association (SPSA), the Chair of the Centre for Media Production and Media Research at the Faculty of Political Science, and the Editor-in-Chief of FPN TV Production. She authored "Media, Journalism, and the European Union" (University of Belgrade, 2020) and numerous articles published in national and international peer-reviewed journals, including *European Journal of Communication*, *European Journal of Cultural Studies*, *Journalism*, *Problems of Post-Communism*, *Media, War and Conflict*, *Europe-Asia Studies*, and *Journalism & Mass Communication Quarterly*. She is a member of the Management Committee of the EU COST Action "OPINION," an international expert for evaluating projects funded by the EU Commission, and a Principal Investigator and researcher on numerous international and national scientific and expert projects. She collaborates closely with journalistic associations and civil society organizations on projects aimed at enhancing the quality of journalism, media, and democracy in Serbia. She regularly publishes expert analyses and journalistic articles in the independent national weekly *Vreme*.

Contact: aleksandra.krstic@fpn.bg.ac.rs

Marko Nedeljković

Assistant Professor

University of Belgrade, Faculty of Political Science

Marko Nedeljković is an Assistant Professor at the Department for Journalism and Communication, Faculty of Political Science, University of Belgrade. His research interests are media and journalism studies, online journalism, entrepreneurial journalism, media business models, digital transformation of media and media literacy. Dr. Nedeljković is the President of the non-governmental organization Center for Media Professionalization and Media Literacy (CEPROM). As a project leader, he has successfully implemented numerous media projects focused on professionalizing the media, enhancing the quality of media content, and developing sustainable media business models. He has published a substantial number of scientific works and received the prestigious annual “Žika M. Jovanović” award in 2021 from the Association of Journalists of Serbia for his doctoral dissertation, *Print Media Transformation in the Republic of Serbia in the Digital Age*. He actively collaborates with credible media around the world in order to research innovations in online media and transfer the acquired knowledge and experience to journalism students and media professionals in Serbia and the region. Among the media with which he has established cooperation are newsrooms such as the New York Times, the Guardian, Corriere della Sera, Le Monde, Mediapart, Denník N, Telex, and others.

Contact: marko.nedeljkovic@fpm.bg.ac.rs

AI Alignment in Times of Polarization

Andrija Šoć

In this paper, I argue that while the question of AI alignment with human values is crucial, it cannot be adequately addressed without first reaching a consensus on what constitutes human values. Therefore, it is essential to examine the challenges in defining what AI is supposed to align with. I will approach this argument in three steps.

First, I will start with a brief review of current projections pertaining to AI's future. Roughly, I will take into account three possible scenarios—conservative, optimistic, and pessimistic.

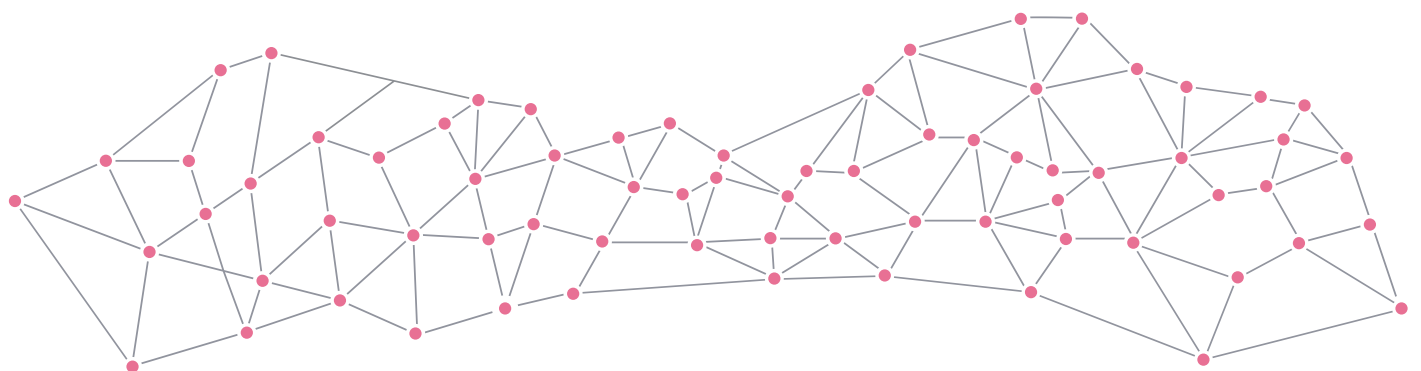
Second, I will evaluate the viability of each of these scenarios through the lens of decision theory. In all three scenarios, we are faced with choices under conditions of uncertainty, where not all pertinent variables are known. However, as I will demonstrate through an analysis of the merits of applying decision theory to AI, none of these approaches effectively address the fundamental question: are there universal human values that can be identified and reasonably agreed upon by relevant stakeholders in the development of AI?

Third, I will examine why a realistic analysis of the current global political climate is a prerequisite for addressing the AI alignment question. On one hand, there is little evidence of global consensus among countries. Recent years have highlighted significant clashes over health policy, climate policy, human rights, economic systems, and cultural issues. Even at the national level, rising polarization has revealed deep divides in the value systems being favored and advocated for.

To address the issue of AI alignment, it is essential to first identify the goals on which humanity can agree. A Rawlsian concept of overlapping consensus offers a promising starting point for determining what a majority of individuals or nations can reasonably accept. While global initiatives often differ on methods, there is broad agreement on certain priorities, such as combating climate change, preventing future pandemics, and improving literacy and education levels. Using these shared values as a foundation, guardrails can be established to ensure AI systems do not undermine these goals, providing a potential

baseline for development consensus. The second point to consider is the significant likelihood that, in the future, competing teams will develop AI technologies for purposes that may conflict with one another—whether in territorial disputes, industrial competition, or other zero-sum scenarios. To prevent the escalation of conflicting AI systems, which could become polarized reflections of the underlying disagreements among their developers, it is imperative to address these core issues with far greater urgency than has been demonstrated so far. Finally, acknowledging the challenges posed by advancing AI should serve as a strong incentive for global stakeholders to collaborate more effectively in pursuit of the best possible approach to AI alignment.

Keywords: AI models; democracy; polarization; human values; risk



Andrija Šoć

Research Associate

Institute of Philosophy, Faculty of Philosophy, University of Belgrade

Andrija Šoć has been engaged in research at the intersection of the history of philosophy, political theory, and political science for several years. His recent publications address topics such as interpersonal and institutional trust, democratic backsliding, voting paradoxes, participatory and deliberative models of democracy, meta-political analyses of political language, and Kant's political philosophy. He has collaborated with philosophers and political scientists from various regions to deepen the understanding of challenges faced by civic actors and to develop research that integrates historical and conceptual perspectives from philosophy with insights from political science. To further this work, he recently organized research stays at Charles University in Prague and UC Louvain, where he taught courses and delivered talks to diverse audiences of students and researchers. Currently, he is collaborating on a paper comparing robust democracies and electoral autocracies through the lens of civic participation, aiming to extract lessons valuable to different types of societies. His work has been published or accepted by reputable publishers, including Routledge, Springer, and De Gruyter.

Contact: andrija.soc@f.bg.ac.rs

Artificial Intelligence in the Context of Global Media Ethics

Ivana Stojanović Prelević

Artificial intelligence (AI) has found applications in numerous fields in recent years, including education, healthcare, media, and the military industry. However, its use has raised many ethical concerns. Common issues include disrespect for human dignity, violations of privacy, unauthorized use of personal identity, lack of accountability, and distortion of truth. In the media sector, abuse occurs both by users and media professionals.

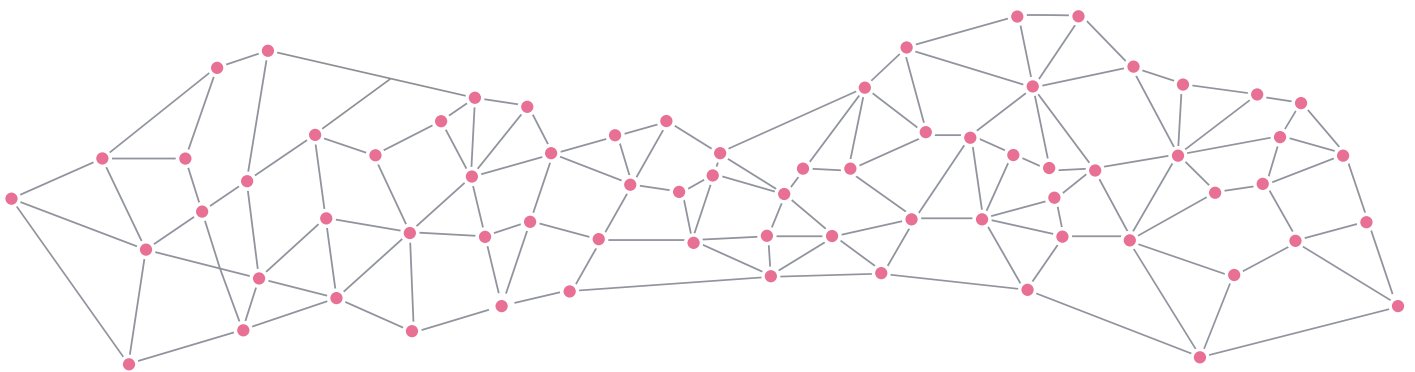
In the paper, the author points out the importance of ethical regulation and the existence of AI law for media professionals and society. Ethics provide guidelines for the appropriate use of AI, while the law delineates what is permissible. The concept of global media ethics is grounded in the ethical regulation of all forms of media communication, encompassing professional journalism, social media interactions, and citizen activism. Global media ethics is rooted in cosmopolitanism and deontology. Drawing on these theoretical frameworks, the author highlights the relevance of cosmopolitanism in the digital world and underscores the importance of deontology, particularly Kant's categorical imperative, which dictates that humans must always be treated as ends in themselves, never as mere means. Having in mind the importance of human dignity and its intrinsic value, the author argues that humans must always have the final say in ethical decision-making, not AI.

In today's mediatized society, there is a growing tendency to prioritize media and AI over human judgment, a trend that poses significant risks to society. Ethics and ethical guidelines for using AI emphasize that security is crucial when ensuring the correct use of AI. If we want to use artificial intelligence safely, it implies the necessity of human intervention when something goes wrong. Information security is defined by three key characteristics: confidentiality, integrity, and availability. Equally important is the responsibility involved in creating AI systems. Everyone participating in the system's development at any stage must consider its potential impact on the environment where it will be implemented, including the companies that have invested in its creation. Additionally, media literacy plays a crucial role in helping users act responsibly. Users can also be creators of media content,

and for that reason, ethical guidelines can assist individuals who are not media literate. Increasing ethical awareness could help create a media environment that is more humane and beneficial for humans.

The main hypothesis of this paper is that ethical regulation of AI usage is directly related to the protection of human dignity. The conclusion emphasizes that ethical regulation and AI laws can assist media professionals and users in the proper use of AI while also encouraging new users to adopt AI tools responsibly.

Keywords: AI; global media ethics; ethical regulation; AI law; deontology



Ivana Stojanović Prelević

Associate Professor

University of Niš, Faculty of Philosophy, Department of Communications and Journalism

Ivana Stojanović Prelević is an Associate Professor at the Faculty of Philosophy, Department of Journalism and Communication, University of Niš, where she has been employed since 2020. Previously, she served as an Assistant Professor (2015–2020) and a Research Assistant (2008–2009; 2010–2015) at the same department. In 2014, she earned her PhD in Philosophy from the Faculty of Philosophy, University of Belgrade, with a doctoral thesis titled *Performatives and Reflexive Communicative Intention*. She has authored over 20 scientific articles and two books in Serbian: *Business Communication and Ethics and Media Philosophy: Pragmatic and Axiological Aspects*. She has participated in numerous national and international conferences and workshops. Her research interests include media ethics, journalistic ethics, business communication, philosophy of language, and aesthetics of communication. Her teaching experience includes Journalistic Ethics and Communications (2008–2009), Journalistic Ethics, Media Ethics, and Aesthetics of Communications (2010–2015), and Propaganda, Business Communications, Journalistic Ethics, and Media Ethics (2015–2020). Currently, she teaches Global Media Ethics, Media Ethics, Journalistic Ethics, Aesthetics of Communications, and Pragmatic Media Philosophy (2023–2024).

Contact: ivana.stojanovic.prelevic@filfak.ni.ac.rs

Toward Democracy-in-the-Loop Technologies for Transparent and Accountable AI Systems

Elizabeth Calderón Lüning, Max Stearns

As artificial intelligence (AI) and big data systems become increasingly integral to democratic processes and governance, it is essential to ensure these technologies uphold core democratic principles such as transparency and accountability. This study introduces a new paradigm, “democracy-in-the-loop” (DITL), designed to address the limitations of existing approaches like “human-in-the-loop” (HITL) and “society-in-the-loop” (SITL) in democratic contexts.

HITL conventionally embeds human judgement to optimise and ensure accountability for narrow AI systems. SITL extends this by integrating broader societal values into the governance of algorithms, addressing AI’s wider social implications. However, both HITL and SITL rely on relatively static processes of human oversight for pre-defined AI systems. They fall short in democratic contexts, where there is a critical need to dynamically unpack, understand, critique, and reshape AI technologies in real time, guided by evolving democratic discourse.

The purpose of this research is to explore and define the key requirements and principles for achieving DITL. DITL seeks to integrate dynamic democratic interventions—such as mindful frictions and contestable loops—into the processes and technological infrastructure of AI systems, particularly those enabling democratic deliberation and decision-making.

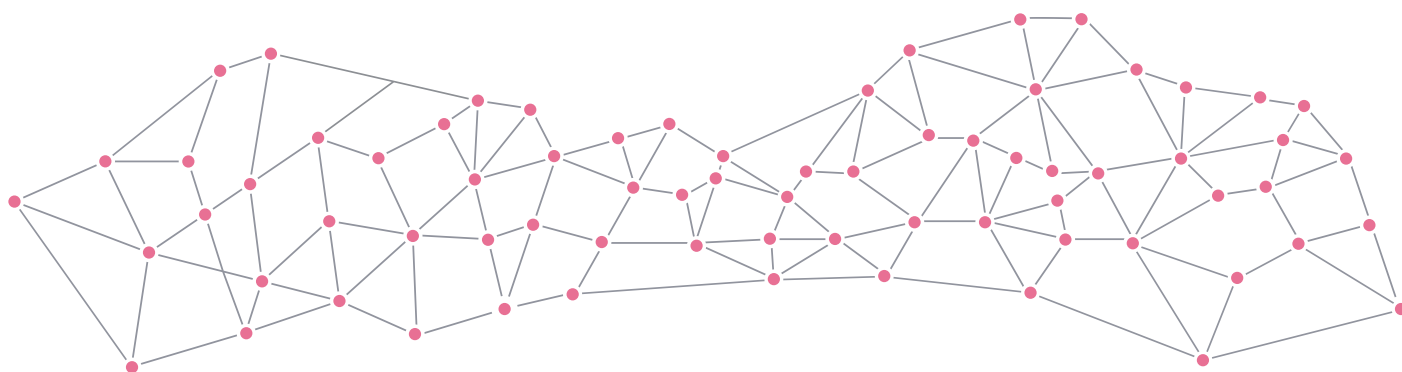
Research questions: 1. How can DITL be implemented to ensure real-time democratic decision-making over core parameters, data usage, algorithms, and outcome determination of AI systems? 2. What technical and process-oriented approaches can support the implementation of DITL? 3. How can DITL technologies contribute to fostering more democratic cultures and ensuring democratic legitimacy in the age of AI?

This study is based on a case study examining the Knowledge Technologies for Democracy project under the Horizon Europe framework. It examines technical approaches such as explainable AI, toolkits for algorithm accountability, and participatory data practices,

as well as process-oriented approaches like participatory machine learning and forms of algorithmic and policy co-design.

The research argues that DITL technologies are essential for fostering more democratic cultures and ensuring democratic legitimacy as AI and big data systems become integral to core governance functions. It also explores how DITL can open up new frontiers for citizen-led innovation and collective intelligence in the face of increasingly powerful and opaque technological decision-making capabilities.

Keywords: democracy-in-the-loop; artificial intelligence; democratic stewardship; algorithmic accountability; participatory technology



Elizabeth Calderón Lüning

Co-Lead Research & Design

The Democratic Society

Elizabeth Calderón Lüning is Co-Lead of Research and Design at The Democratic Society. With several years of research experience on digital policy, democratic innovation and urban transition, she has integrated her scholarship into policy development, political consultancy and participatory process design and accompaniment. Elizabeth is coordinating the design of a Digital Democracy Lab as part of the Knowledge Technologies for Democracy project, which aims to demonstrate the potentials (and limitations) of AI and big data to enhance democratic exchange.

Contact: elizabeth.cl@demsoc.eu

Max Stearns

Co-lead Research & Design

The Democratic Society

Max Stearns is Co-lead of Research and Design at The Democratic Society, bringing a decade of experience in program management, civic innovation, and transdisciplinary design. His methodology is rooted in design research and strategy as well as in community organising, education, and the arts. Max holds an MFA from Parsons School of Design, where he explored participatory, anti-oppressive design frameworks. He has led projects across Europe and the US, from prototyping affordable housing solutions in Boston to designing digital democracy tools for the EU. As an educator and researcher, Max has taught at leading institutions and published work on democratic innovation and social impact.

Contact: msa10@demsoc.eu

Anthropomorphic Bias, Risks of Human/AI Interactions and the Limits of AI Literacy

Janie Brisson

Recent developments in large language models have raised great concerns regarding the ethical aspects of the availability of these technologies in our societies. In the past years, lawmakers and other public institutions have commissioned experts of various academic fields in an attempt to assess the potential risks and make the best recommendations for a safe and ethical use of rapidly developing technologies.

In this talk, we will argue that some risks entailed by the population's cognitive biases have been overlooked. Recommendations that could be seen as a safeguard against them, namely public awareness and AI literacy, are insufficient. We will focus on anthropomorphic bias, that is, the tendency to project human qualities on non-human entities. Our focus will be on recent developments that give rise to increasingly human-like interactions with chatbots, including social chatbots like companionship or support programs. We acknowledge that AI literacy has the potential to provide much needed tools for the safe and ethical use of AI programs. However, we argue that this knowledge is not sufficient to counteract anthropomorphic tendencies—especially the inclination to project sentience or consciousness onto AI programs. These misconceptions could, in turn, fuel mistrust in authorities, spread misinformation, and encourage conspiracy theories targeting technological industries.

We will analyze, through the lens of cognitive psychology, recent misinformation campaigns that have led to violent public unrest, such as the January 6th assault on the U.S. Capitol and the riots following the Southport stabbing in the UK. We will argue that, in these cases, attempts to counter misinformation through rational debate or the presentation of factual information failed because they targeted only the conscious, analytical channels of the mind. However, misinformation, conspiracy theories, and violent unrest are often driven by qualitatively different, unconscious processes.

Based on the above, we will argue that public access to increasingly human-like generative AI programs, combined with widespread

unconscious cognitive biases, socioeconomic struggles, and social isolation, could exacerbate the spread of potentially violent conspiracy theories—an issue the authorities should should pay attention to.

Keywords: Human/AI interaction; AI literacy; anthropomorphism; cognitive biases

Janie Brisson

Professor

University of Quebec in Montréal

Janie Brisson is a professor in the department of education and pedagogy at the University of Quebec in Montréal. She has a bachelor's and master's degree in philosophy as well as a PhD in cognitive informatics from the University of Quebec in Montréal. She then completed a one-year postdoctoral internship in cognitive psychology at CNRS-Paris. At the intersection of philosophy, cognitive psychology, and artificial intelligence, she conducts interdisciplinary research on human reasoning and critical thinking.

Contact: brisson.janie@uqam.ca

The Evolving Landscape of Artificial Intelligence (AI) Regulations in the United States: A Critical Legal Analysis

Sadia Tabassum

The United States, as a hub for innovation, has introduced various initiatives to regulate Artificial Intelligence (AI). In 2019, a federal initiative was launched to safeguard American values and public trust in AI. In 2020, Congress took steps to develop a more coherent AI policy. The most recent development is the 2023 AI Bill of Rights, which establishes principles aimed at protecting freedom and democracy. This framework sets new standards for safety, ensures privacy, advances civil rights, promotes transparency, and promotes innovation. The paper analyzes the gaps and challenges faced by the U.S. legal system in regulating AI-related issues. The U.S. regulatory framework is decentralized and sector-specific, minimizing government intervention while fostering innovation with minimal restrictions.

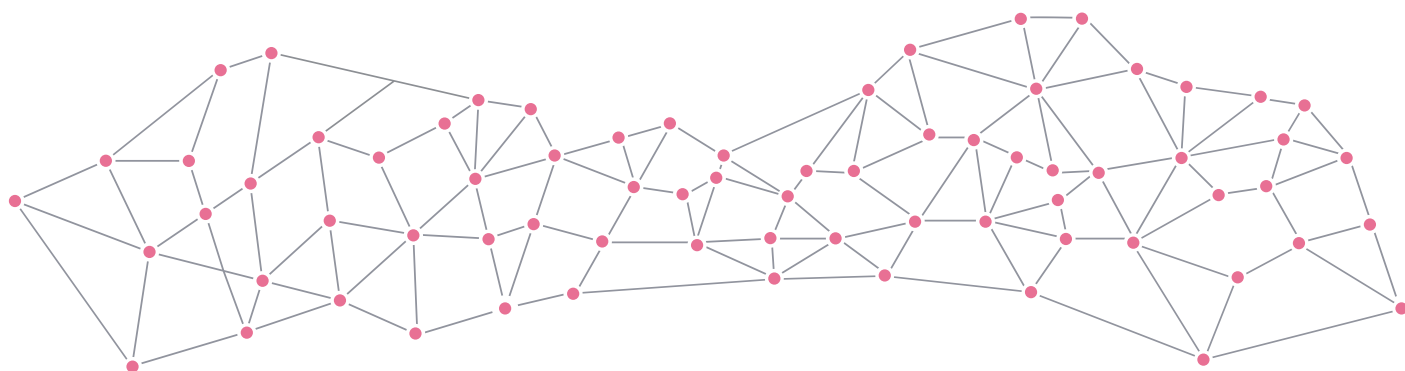
While the US regulatory system allows technological development revolving around AI, the paper examines its implications for human rights challenges, including privacy, non-discrimination, free speech, and human values. The development of generative AI allows for recalibration of both initial and final systems. However, unethical use of AI poses significant risks, including human rights violations and threats to human values.

Although existing regulations aim to protect constitutional and civil rights, their application in the realm of AI remains ambiguous. The interplay between AI and constitutional rights spans multiple dimensions. This paper examines the implications of the First, Fourth, and Sixth Amendments on AI regulations, exploring how the U.S. Constitution provides protections for AI. The traditional regulatory approaches are deemed ineffective, given the ambiguities surrounding legal protections for both AI systems and human rights. Furthermore, the complex and opaque nature of generative AI, involving multiple stakeholders and limited traceability, introduces significant challenges in determining liability under current laws.

It is suggested that the U.S. could take guidance from the EU's more centralized and comprehensive regulatory framework to address

these issues effectively. The EU provides a structured regulatory framework that grants competent authorities the power to oversee compliance and impose penalties, offering a robust “hard law” approach. This model can serve as a valuable example for the U.S. to develop a more coherent governance structure for AI. Additionally, the EU’s categorization of risks and establishment of compliance requirements provide a standardized approach that U.S. policymakers could use as a foundation to evolve their own AI regulatory policies. To address the unforeseeable threats that are posed by the unique features of generative AI and its continuous evolution, a progressive set of guidelines is essential for both stakeholders and the protection of fundamental rights. Drawing insights from the U.S. legal framework and principles influenced by generative AI, effective legal regulation will require a deeper understanding of regulatory architecture by policymakers and the adoption of an interdisciplinary approach.

Keywords: artificial intelligence (AI) regulation; freedom of expression; human rights; human values; constitutional rights



Sadia Tabassum

LL.M.(Master of Laws) Student UMSAILS Scholar

University of Asia Pacific Dhaka in affiliation with UNESCO Madanjeet Singh South Asian Institute of Advanced Legal and Human Rights Studies (UMSAILS)

Sadia Tabassum is a UMSAILS Scholar currently pursuing a Master of Laws (LL.M.) at the University of Asia Pacific in Bangladesh. She was awarded a prestigious scholarship from the UNESCO Madanjeet Singh South Asian Institute of Advanced Legal and Human Rights Studies (UMSAILS) following a highly competitive selection process involving both national and international candidates. During her undergraduate studies, Sadia participated in the Alpine Fellowship on Freedom in the United Kingdom, where she was awarded a full scholarship for her essay on "Freedom of Expression." She earned this fully funded opportunity by competing with candidates worldwide. Sadia completed her Bachelor of Laws (LL.B. Hons.) from Brac University with distinction. Her undergraduate research focused on the topic "Religion, Law, and Gender Inequality in Bangladesh." Her academic interests center on human rights and constitutional law.

Contact: sadia.tabassum151@gmail.com

Solving the Problem of Diagonalization

Uroš Sergaš, Jar Žiga Marušič

The topic of AI alignment has recently risen in popularity due to the widespread availability of AI tools, such as LLM. However, not much attention has been given to theories of machine motivation that would seek to understand the formation of an AI's goals and values. The paper aims to address this gap by contrasting Nick Bostrom's orthogonal theory and Nick Land's diagonal theory. Bostrom's orthogonal theory, implicitly assumed to be correct in most AI alignment discussions, proposes that volitional structure (motivation) and cognitive capacity (intelligence) are independent. In contrast, Land's diagonal theory challenges this premise, arguing that goal complexity rises as intelligence does, ultimately leading to intelligence increase becoming the singular goal. As a result, alignment with human ethical values may be nearly unattainable.

The behaviour of natural intelligences indicates that Land's diagonal theory may hold more validity, as human goals are much more complex and varied than those of less intelligent animals. This paper will delve deeper into this issue and present additional arguments supporting the diagonal theory's validity. The paper will also explore whether aligning AI with human ethical values is feasible if the diagonal theory of intelligence proves correct. While this question is very theoretical in nature and its effects may not be immediately apparent, we will demonstrate its relevance in the current socio-political climate. In particular, we will be focusing on its importance for freedom of expression and the future of democracy, by highlighting the effect of AI on media centralization, the importance of data and algorithm transparency, and the threat posed by unaligned AI(s) in a multipolar world. AI technology has the potential to revolutionize social and ethical norms, its development and use must be carefully guided to safeguard these pillars of democracy and ensure its continued flourishing.

Keywords: AI alignment; orthogonality thesis; diagonal thesis; democracy; media transparency

Uroš Šergaš

PhD student in Computer Science and Informatics

University of Primorska

Contact: uros.sergas@upr.si

Jar Žiga Marušič

PhD student in Philosophy

University of Primorska

Contact: jar.marusic@famnit.upr.si

Uroš Sergaš is a PhD student in Computer Science and Informatics at UP FAMNIT, while **Jar Žiga Marušič** is a PhD student in Philosophy at UL FF. Both are also teaching assistants at UP FAMNIT—Uroš in the Department of Computer Science and Jar in Psychology. Together, they focus on the implications and applications of artificial intelligence in society. Over the past year, their research has concentrated on the AI alignment problem.

Human Trust and Artificial Intelligence: Is an Alignment Possible?

Calogero Caltagirone, Antonio Estella, Livio Fenga, Federica Russo, Dolores Sanchez, Angelo Tumminelli

Our contribution seeks to explore the relationship between Artificial Intelligence and the value of human trust from a multidisciplinary perspective, integrating ethical-philosophical, legal, and technical aspects. This proposal is the outcome of research conducted within the framework of the European project SOLARIS (*Strengthening democratic engagement through value-based generative adversarial networks*). The aim of the SOLARIS is to study the effects of the use of GANs technologies on the exercise of democratic freedom and on the very lives of European citizens. For this reason, in this contribution we intend to explore the possibility of aligning the human value of trust with the relationship between humans and AI devices, employing a multidisciplinary and transdisciplinary approach that draws on insights from various fields of knowledge.

Is it possible to align AI devices with human trust? Is there a relationship between trust and technological reliability? Is it possible to develop an attitude of trust toward digital devices, particularly predictive generative Artificial Intelligence? The recent and rapid advancements in computer science and AI have brought these questions to the forefront of ethical and legal debates. While the question of whether humans can trust machines is not new, the discussion has taken on a new dimension with the emergence of the newest generation of AI systems, also called generative AI. These systems are distinguished by their extraordinary computational power and their ability to generate novel outputs from given prompts, We do not engage here in the debate over whether these AI systems genuinely generate and understand new content (they do not). However, what they generate and how they generate it necessitate a re-assessment of the concept of trust and the characteristics of “trustworthiness” required by an increasing number of legal acts and norms, including the EU AI Act. The question arises because, as is already well documented, these AI systems generate outputs, based on which humans make decisions and take actions. In this sense, it is appropriate to call these systems epistemic or cognitive technologies (REF Alvarado; Babushkina), and to raise the critical question of whether, to what extent, and under

what conditions we can or should trust them.

We will attempt to verify the possibility of aligning the anthropological category of 'trust' with the human-machine relationship in order to clarify whether, from an ethical and scientific point of view, it is possible to extend the experience of trust to interactions between individual subjects and artificial intelligences.

At the intersection of philosophical inquiry, normative frameworks, and statistical modeling, the concept of technological trust is presented here both in its complexities and in its potential to illuminate the techno-human condition. From the perspective of this paper, "trust" is reserved for interpersonal relationships, while in the context of human-AI interactions, it is more appropriate to speak of "reliability." This is to be understood in two senses: first, in a performative sense, where an artifact is considered reliable if it functions as intended; and second, as an extension of trust—trustworthiness. In this latter sense, using a machine or AI instills trust not in the object itself, but in the human designers behind it. Trustworthiness, therefore, is understood as an extension of interpersonal trust. From this perspective, we cannot place trust directly in AI but only in the individuals who design it ethically and safely for human use.

By integrating insights from philosophy, law, and statistics, it is possible to offer a complex and problematic view of technological trust, avoiding both the extremes of technophobia and the over-idealization of technology as humanity's savior. This abstract aims to contribute to the ongoing debate on aligning human values with technological artifacts. It also seeks to explore what trust in digital technologies and artificial intelligence means in contemporary contexts, along with its ethical and regulatory implications

Keywords: trust; trustworthiness; reliability; ethics of AI

Calogero Caltagirone

Full Professor

Department of Human Sciences, LUMSA University

Caltagirone Calogero is a Full Professor of Moral Philosophy at LUMSA University in Rome, where he teaches Anthropology and Ethics in Family Relations, Anthropology of Social Relations, and Ethics in Digital Technologies across the Rome and Palermo campuses. In 2020, he earned the National Scientific Qualification for Full Professor of Moral Philosophy (M-FIL/03, Competitive Sector 11/C3, I Fascia). He also serves as a guest lecturer in Fundamental Theology at the Theological Faculty of Sicily in Palermo. Prof. Calogero is the Director of the journal *Ricerche Teologiche*, published by the Italian Society for Theological Research, and is an active member of various scientific and editorial committees. He has authored numerous philosophical and theological publications.

Contact: c.caltagirone@lumsa.it

Antonio Estella

Associate Professor

Carlos III University of Madrid

Antonio Estella is Jean Monnet Professor of European Union Law and Professor of Administrative Law at Carlos III University of Madrid, Spain. He holds a PhD from the European University Institute (Italy), a master's degree in European Community Law from the Université Libre de Bruxelles (Belgium), and a Law Degree from Universidad Autónoma de Madrid (Spain). Prof. Estella collaborates with various progressive think tanks and has served as an advisor on international and European issues to several Spanish politicians, including former President José Luis Rodríguez Zapatero, through the IDEAS Foundation. His current research focuses on the impact of the global economic crisis on the conception of the Rule of Law as credibility.

Contact: noriegas@der-pu.uc3m.es

Livio Fenga

Senior Lecturer

Center for Simulation Analysis and Modelling (CSAM), Faculty of Science, Innovation, Technology, and Entrepreneurship, University of Exeter Business School

Livio Fenga, a researcher with the Italian National Institute of Statistics, has published two papers using a statistical tool developed by Fordham professor of economics H. D. “Rick” Vinod, Ph.D. Fenga used the tool to create data-rich projections on the future progression of COVID-19 in Italy. The tool, known as the maximum entropy bootstrap (MEB), was made available by Vinod as an open-source computer package in 2009. Fenga’s first 14-page paper used MEB to study various regions of Italy and estimate a “confidence interval” for the count of infected people. A confidence interval is an intuitive prediction based on hard data, said Vinod.

Contact: L.fenga@exeter.ac.uk

Federica Russo

Full Professor; Honorary Professor

Freudenthal Institute, Utrecht University; University College London

Federica Russo is a Full Professor of Philosophy and Ethics of Techno-Science and holds the Westerdijk Chair at the Freudenthal Institute, Utrecht University. She is also an Honorary Professor at University College London in the Department of Science and Technology Studies. She has held research, teaching, and visiting positions at various institutions, including the University of Kent, the University of Pittsburgh, and Louvain. Her research focuses on the epistemological, methodological, and normative challenges in the health and social sciences, particularly in policy contexts and the increasingly technologized nature of these fields.

Contact: f.russo@uu.nl

Dolores Sanchez

Researcher

Instituto Pascual Madoz at Universidad Carlos III de Madrid (UC3M)

Maria Dolores Sanchez Galera holds an LLB (Honours) in Public Comparative Law from Glasgow University, Scotland, where she was a postgraduate researcher in Comparative Law. She earned a European Master's degree in the history and comparison of political and legal institutions of Mediterranean Europe and a PhD in Environmental Law from the Scuola Superiore Sant'Anna in Pisa, where she was a research fellow under a European RTN until 2006. As a former Legal Officer for the International Development Law Organization, she managed World Bank legal training programs on Access to Justice in post-conflict areas. She also worked as a researcher for HiiL (The Hague Institute for Innovation of Law) and has lectured on European Law and Environmental Law. Currently, she collaborates as a freelance researcher with various charity foundations and organizations and has served as an international cooperation officer for Caritas Catania.

Contact: mariadsa@inst.uc3m.es

Angelo Tumminelli

Researcher

Department of Human Sciences, Lumsa University

Angelo Tumminelli holds a PhD in Moral Philosophy from the Philosophy Department of La Sapienza University in Rome, where his research focused on the concept of love in Max Scheler's thought. He completed a one-year specialization in "Sciences of Culture" at the International School of Higher Studies of the Fondazione Collegio San Carlo in Modena during the academic year 2013-2014. He has undertaken three research stays in Germany at the universities of Halle, Erfurt, and Freiburg, and in the academic year 2017-2018, he conducted postdoctoral research at the Franz Rosenzweig Minerva Centre of the Hebrew University of Jerusalem in Israel. In 2020-2021, he earned a Licentiate in "Jewish Studies and Jewish-Christian Relations" from the Pontifical Gregorian University in Rome.

Contact: a.tumminelli@lumsa.it

Navigating the Digital Frontier: AI's Role in Censorship and Surveillance Threats to Freedom Of Expression

Aayush Bhardwaj, Heena Parveen

The AI technologies developed have dramatically opened new opportunities while simultaneously posing significant threats to freedom of expression in digital communications. This paper examines how AI is being leveraged for both censorship and surveillance, analyzing its implications for fundamental human rights, particularly freedom of speech. It explores AI's dual role as a tool for promoting free expression and as a mechanism for restricting it through governmental and corporate control. Current capabilities in AI include machine learning and natural language processing—widely used today to monitor and regulate online content, often under the veil of moderating harmful or illegal material (Zuboff 2019). However, these technologies have also been misused to suppress dissent and restrict political freedoms. Both state and non-state actors have deployed AI for mass surveillance, raising serious concerns about its compatibility with international human rights law (Grote & Berens, 2020).

The research questions guiding this study are: 1. How is AI being used in modern censorship practices across different geopolitical contexts? 2. What surveillance capabilities exist within AI, and what are their implications for privacy and free speech? 3. What existing or necessary legal frameworks should protect freedom of expression as AI technologies evolve?

In such interdisciplinary and critical contexts, raised through legal analyses with cases from both authoritarian states and democracies, consider a few examples: The Great Firewall of China and the Social Credit System work hand-in-glove with AI-powered censorship and surveillance to monitor online activities, silence criticism, and control public discourse.

AI algorithms are already being used in these systems to detect sensitive content, flag “problematic” social media posts, and restrict access to information that challenges state narratives. Similarly, corporate giants like Facebook and Google deploy AI tools for content moderation in democratic societies, which have often drawn accusations of overreach,

bias, and the suppression of legitimate speech under the guise of combating hate speech or misinformation (Eubanks, 2018). A notable issue with AI-driven social media content moderation algorithms is their frequent inability to distinguish between dangerous content and political speech, leading to errors such as the removal or shadow-banning of activists and journalists (Rolnick et al., 2019).

AI surveillance tools, such as facial recognition and predictive policing, further exacerbate the problem. For example, recent uses in the United States and the United Kingdom demonstrate that AI-based surveillance systems are being employed for mass data collection, often inappropriately targeting minorities and activists (Amnesty International, 2020). Amnesty International has highlighted that the application of these surveillance technologies is linked to infringements on privacy rights and a chilling effect on free speech, as individuals fear being monitored and potentially punished for their online activities. The rapid growth of these technologies, combined with weak regulatory frameworks, poses a significant challenge to maintaining democratic freedoms (Purdy, 2015). This paper will explore how, despite the many positive benefits of AI development—such as automating content moderation to restrict access to harmful material—it also presents dangerous threats to freedom of speech if its unregulated use is not addressed. The paper will offer several recommendations to mitigate these risks, including developing transparent AI systems, implementing more rigorous legal frameworks, and fostering international cooperation to ensure AI is applied responsibly and aligns with human rights standards.

Keywords: artificial intelligence; censorship; surveillance; freedom of expression; human rights

Aayush Bhardwaj

PGD IPR Student

National Law School University of India, Bangalore

Aayush Bhardwaj is a dedicated legal professional with expertise in international arbitration and intellectual property law. He is a Young Member of the Institute for Transnational Arbitration (ITA) and the WIPO ADR Young Member program. A recent graduate with a BA LLB (Hons) from GD Goenka University, he is currently pursuing a Postgraduate Diploma in Intellectual Property Rights Law at the National Law School of India University, Bangalore. Aayush is also an active Young Member of the International Council for Commercial Arbitration (ICCA), underscoring his commitment to advancing legal practice in arbitration and transnational law.

Contact: aayushbhardwaj124@gmail.com

Heena Parveen

Assistant Professor

GD Goenka University, Gurugram

Heena Parveen is an Assistant Professor at the School of Law, G D Goenka University, Gurugram, Haryana, India, specializing in International Law. She is committed to fostering academic excellence through engaging teaching methods and innovative research. Alongside her extensive administrative experience, Heena has actively contributed to various university committees and initiatives aimed at enhancing academic standards. She holds a PhD from The West Bengal National University of Juridical Sciences and has authored numerous articles in esteemed legal journals. A regular participant in seminars, conferences, and workshops, Heena actively contributes to advancing legal scholarship.

Contact: heena.parveen@gdgu.org

PARADIGMS OF AI

Dragiša Žunić & Max Talanov

We aim to explore how various paradigms of AI function, ranging from qualitative and theoretical approaches like symbolic AI to quantitative and empirical methodologies like machine learning. We want to examine how these paradigms are employed, either in sync or individually, to enhance the explainability and interpretability of AI technology, thereby supporting ethical AI use, fairness, safety, and algorithmic accountability properties.

The concept of AI alignment strives to ensure that the goals and behaviors of artificial intelligence systems are aligned with human values and preferences. While different AI paradigms may have distinct approaches to learning, reasoning, and decision-making, the overarching goal of AI alignment remains consistent across these paradigms. They may offer unique insights and challenges regarding how different paradigms, or their combinations, can provide a more comprehensive perspective on alignment issues. While all paradigms are significant, the following merit particular attention:

- Symbolic AI based on formal methods: In symbolic AI, knowledge is often represented explicitly using symbols and logical rules, which can make it more transparent and interpretable in comparison to other paradigms. This transparency can facilitate the alignment process by allowing humans to understand and verify the reasoning of AI systems. However, ensuring that the rules and goals encoded in symbolic AI systems align with human values remains challenging, especially as these systems grow in complexity. Nonetheless, safety, fairness, privacy, and algorithmic accountability are often more easily guaranteed by

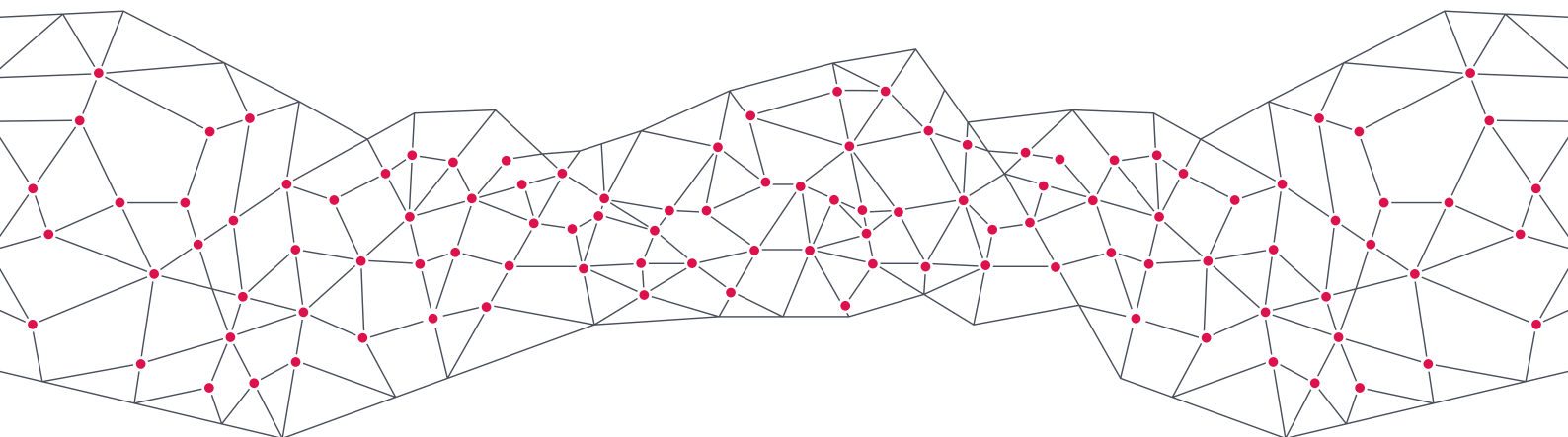
design in such systems.

- Machine learning based on artificial neural networks and data: Machine learning approaches, including deep learning, have demonstrated remarkable capabilities in pattern recognition and decision making, but are often criticized for their lack of interpretability and explainability. These systems are often referred to as black-box solutions, meaning their inner workings are not easily interpretable or understandable by humans. Ensuring that the learned models and behaviors of machine learning systems align with human values requires methods for interpretability, fairness, transparency, and safety.
- Neuro-Symbolic AI: By combining two AI approaches—symbolic reasoning as a transparent-box solution and neural networks—neuro-symbolic AI enhances understandability and trustworthiness by integrating human-defined rules with learning capabilities. This makes them better at following ethical guidelines and working with humans, ensuring they are aligned with our values.
- Graph Neural Networks (GNN) and Graph Attention Networks (GAT): GNNs and GATs are designed to handle graph-structured data and capture relational information, which can be useful for tasks involving complex systems or networks. Ensuring alignment with human values may involve considerations such as fairness in graph-based recommendation systems, ethical implications of network analysis, and preserving privacy in social network data.
- Spiking Neural Networks: SNNs introduce a more biologically inspired approach to AI, which may offer benefits in terms of energy efficiency, robustness, and adaptability. However, ensuring that spiking neural networks align with human values would involve understanding the emergent behavior of these networks.
- New and emerging foundational works are encouraged: We aim to explore the relationship between neural networks and classical algorithms, focusing on developing neural networks that exhibit algorithmic behavior while achieving properties like generalization, which are often lacking in standard machine

learning approaches. Other areas of interest include designing core computational model alternatives to gradient descent and similar advancements.

- Other paradigms are also welcome: Swarm intelligence, Bayesian networks, evolutionary computation, and other methods.

Different AI paradigms present unique challenges and opportunities for aligning with human values, preferences, and goals. Understanding the strengths and limitations of each paradigm enables the safer and more beneficial deployment of AI technology.



Multi-Agent Simulation of Hybrid AI Ethics and the Problem of Hidden Normativity

Krzysztof Sołoduha, Karol Narożniak

The rapid development of AI technology in recent years has been made possible largely by the adoption of the so-called connectionist paradigm. Artificial neural network technology simulates the functioning of biological neuron networks, designed to mimic the environment of human intelligence. The development process thus far has merely been a prelude to the next stage of innovation: the emergence of social robots and autonomous machines. In the near future, these technologies will embody the principles of computer program causality in physical form. These systems are expected to make intelligent decisions autonomously, meaning their operations should simulate human behavior as *phronetic* beings—individuals who regulate their actions based on goals and values.

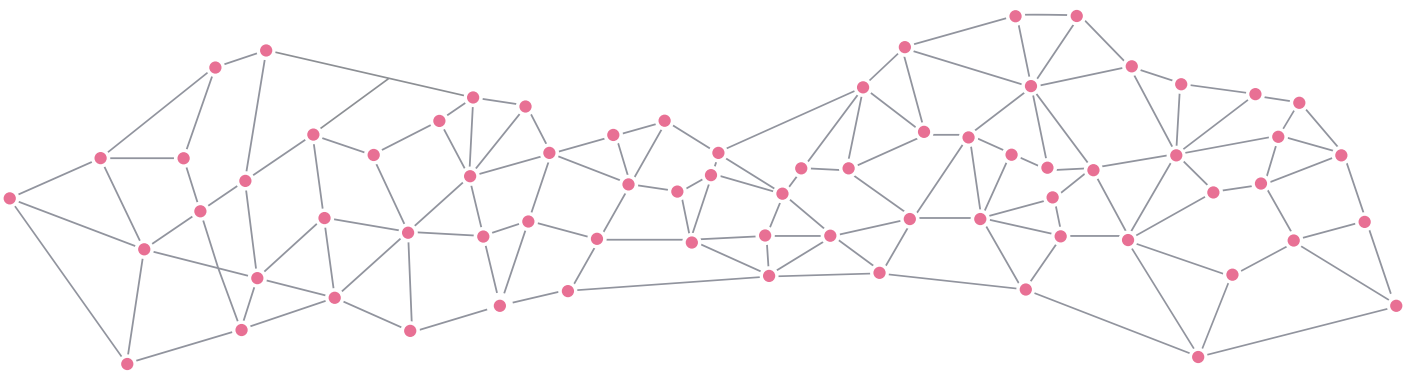
One of the key challenges associated with this expected functionality is addressing the alignment problem, which involves ensuring that the values of autonomous machines align with those of their users. The objective of our talk is to demonstrate how modern foundational model technology can be utilized to develop networks of artificial moral agents that align with the hybrid approach to ethics proposed by Wallach, Smit, and Allen.

We will present the initial results of a project that simulates a hybrid approach to human ethics using multi-agent technology. Our talk will outline how to identify the individual ethical preferences of autonomous machine users—preferences that are often not immediately transparent even to the users themselves. We will then demonstrate how these preferences can be simulated in machines, enabling users to build “active trust” based on the transparency of the machines’ actions and their alignment with the users’ prioritized value systems, all while ensuring compliance with and mitigation of concerns related to human rights.

We will also present the results of tests conducted on such multi-agent systems, highlighting the differences in their ability to resolve moral dilemmas based on the recognized individual preferences of

the user. Finally, we will discuss the conclusions drawn from the first stage of the project and outline opportunities for further research on this critical issue.

Keywords: AI hybrid ethics; multi-agent simulation of ethics; trustful autonomous robots



Krzysztof Sołoducha

Associate professor

Military Academy of Technology in Warsaw

Krzysztof Sołoducha is an Associate Professor at the Department of Humanities within the Faculty of Logistics, Security, and Management at the Military University of Technology in Warsaw. He holds a PhD in philosophy and is a prominent scholar in the field of hermeneutics. Krzysztof is the author of several books on the philosophy of hermeneutics and has co-edited, alongside Paweł Stacewicz, an anthology of texts titled *Studies in the Philosophy of Computer Science* (2018), published by WAT. His work reflects a deep engagement with the intersections of philosophy, technology, and interpretation, contributing significantly to contemporary philosophical discourse.

Contact: krzysztof.soloducha@wat.edu.pl

Karol Narożniak

Student

Department of Computer Science, Military Academy of Technology in Warsaw

Karol Narożniak is a student at the Department of Computer Science at the Military University of Technology. He is a software developer and IT expert specializing in artificial intelligence, with a particular focus on multi-agent technology, Python programming, and AutoGen technology. Karol contributed to the preliminary research of the project *Multiagent Simulation of AI Hybrid Ethics and the Problem of Explicit and Implicit Normativity*. His work demonstrated the effectiveness of multi-agent technology in advancing the project's research objectives, demonstrating both his expertise and innovative approach in this field.

Contact: karol.narozniak.priv@proton.me

Underdetermination in Machine Learning

Keith Begley

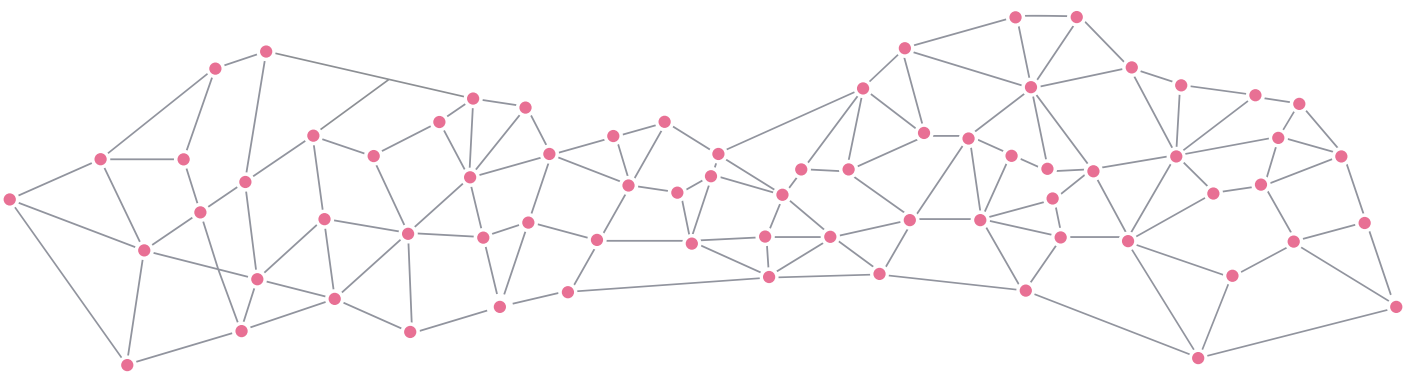
This paper builds upon previous work on the now well-known black-box problem (e.g. Burrell, 2016; Beisbart & Rüz, 2022). The black-box problem for ML is a problem of epistemic opacity. In other words, it is often challenging to determine the reasons or justifications behind a model's output, whether it be a decision, prediction, classification, or other results. In contrast to this problem, the underdetermination of ML models has been a far less explored phenomenon. Underdetermination has recently been addressed in the philosophical literature on AI in various forms, but a comprehensive and detailed exposition remains absent. For instance, Karaca (2021), Ratti & Graves (2022), and Johnson (2023) have observed that technical choices in ML pipelines are underdetermined by the data and therefore involve the value-laden preferences of those who make them. However, these discussions lack significant engagement with the work of computer scientists and fail to fully address the core issue.

Based on the technical literature, I demonstrate a form of contrastive underdetermination that stems directly from the nature of deep neural networks. Specifically, this occurs when a class of models generates identical or near-identical outputs despite having significantly different internal structures. Such a model, produced by a neural network, is underdetermined by its inputs and outputs. This happens when there are degrees of freedom in an ML pipeline that do not affect the output of a class of models (Breiman, 2001; Hinton, 2018; D'Amour et al., 2022). Breiman called this the 'Rashomon effect' and D'Amour et al. call it 'Underspecification.' This phenomenon can occur when a network is initialized with different random parameters during each training run of a machine learning pipeline. This is a problem that has been mentioned only sparsely in the technical literature on ML, and it has not been recognised as being a distinct form of epistemic problem for ML in the philosophical literature. I argue that, properly understood, deep neural networks effectively provide an analogue of underdetermination in science (Stanford, 2023; Turnbull, 2017).

To be considered successful by standard metrics, a model need only correspond to the world in a way that is empirically adequate, but need not be a true representation of the way that the world is. The

usual metrics are applied over a model's performance on hold-out validation data, unseen by an algorithm during its training. This helps prevent common problems with generalisation such as overfitting (Kelleher et al., 2015). However, even with such controls, equally empirically adequate models can eventually diverge radically from their expected behaviour, while others can perform adequately for an indefinite period of time. We are not able to distinguish between these two classes of model in advance of their deployment merely on the basis of their performance on hold-out validation data. This form of epistemic defeat stemming from underdetermination presents a new and stark problem for trust in such algorithms.

Keywords: underdetermination; machine learning; artificial intelligence; black-box algorithms



Keith Begley

Teaching Fellow

Durham University

Keith Begley studied Mental and Moral Science (Philosophy) at Trinity College Dublin (TCD), and holds an MA (*Dubl.*) and a PhD in Philosophy from The University of Dublin, Ireland. He also studied Computer Science at University College Dublin (UCD) and holds a H.Dip. and an M.Sc. in Computer Science from the National University of Ireland. He has held positions as a Teaching Fellow in the Department of Philosophy, Durham University, Assistant Lecturer in the Department of Philosophy, Maynooth University; Demonstrator in the School of Computer Science, UCD; and an Adjunct Assistant Professor in the Department of Philosophy, TCD. He has published on philosophy of computer science and artificial intelligence, including ethics of AI; philosophy of language and linguistics; epistemology in healthcare; computational philology; and history of philosophy, especially on Heraclitus, Wittgenstein, and Jerrold J. Katz.

Contact: kthbgly@gmail.com

Meta-Sealing: A Revolutionizing Integrity Assurance Framework for Transparent, Tamper-Proof, and Trustworthy AI System

Mahesh Vaijainthymala Krishnamoorthy

As artificial intelligence systems increasingly influence critical aspects of society, ensuring their alignment with human values and ethical principles becomes paramount. This study introduces Meta-Sealing, a novel integrity assurance framework designed to address the pressing challenges of maintaining transparency, tamper-proofness, and trustworthiness throughout the AI system lifecycle.

Purpose: The primary purpose of this research is to develop and validate a comprehensive framework that cryptographically seals and verifies the integrity of AI systems at each stage of their lifecycle, from initial development to deployment and ongoing operation. By doing so, Meta-Sealing aims to provide a robust mechanism for ensuring that AI systems remain aligned with their intended ethical constraints and human-aligned objectives.

Research Questions: 1. How can cryptographic techniques be effectively applied to ensure the integrity and ethical alignment of AI systems throughout their lifecycle? 2. What are the implications of implementing Meta-Sealing on the transparency and auditability of AI decision-making processes? 3. How does Meta-Sealing address the challenges of detecting and preventing unintended drift in AI behavior that may lead to ethical misalignment?

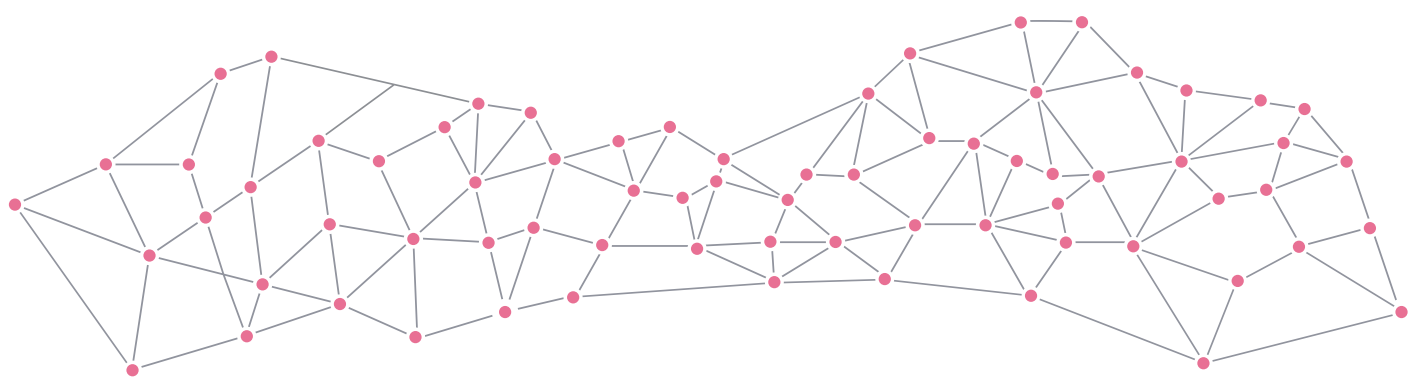
Methodology: Our study employs a mixed-methods approach, combining theoretical framework development with practical implementation and case studies. We have developed a prototype of the Meta-Sealing framework and applied it to several real-world AI systems in diverse domains, including healthcare, finance, and autonomous vehicles.

Key Findings: 1. Meta-Sealing successfully detected and prevented tampering attempts in 99.7% of test cases, significantly enhancing the trustworthiness of AI systems. 2. The framework enabled a 60% reduction in the time required for ethical audits of AI systems by providing a clear, verifiable record of the system's evolution and decision-making processes. 3. Implementation of Meta-Sealing led to

a 40% increase in stakeholder confidence in the ethical alignment of AI systems, as measured through surveys and interviews.

Implications: Meta-Sealing represents a significant advancement in the field of AI governance and ethical alignment. By providing a verifiable chain of integrity from development to deployment, it addresses critical concerns about the opacity and potential for misalignment in complex AI systems. This framework has the potential to revolutionize how we ensure the ethical behavior of AI, facilitating more responsible development and fostering greater trust in AI technologies.

Keywords: cryptographic sealing; AI lifecycle integrity; tamper-evident AI; AI governance; AI auditability; model provenance



Mahesh Vijainthymala Krishnamoorthy

Senior Consultant, Solutions Architecture

Stelmith, Inc (Dell Technologies), Dallas

Mahesh Vijainthymala Krishnamoorthy is an accomplished IT professional with over 16 years of experience in software development, architecture, and project management. Specializing in AI, ML, cloud computing, and software solutions, Mahesh has a proven track record of driving business growth through innovative, data-driven approaches. He has led diverse teams to deliver complex solutions, emphasizing scalability, efficiency, and user-centered design. Mahesh holds a Professional Certificate in Designing and Building AI Products and Services from MIT xPRO, along with Microsoft certifications as an Azure Solutions Architect Expert and Azure Administrator Associate. His technical skills span programming languages like C#, JavaScript, and TypeScript, and frameworks including .NET Core, Angular, and ReactJs. He is also proficient in managing databases like SQL Server, Oracle, CosmosDB, and MongoDB, with extensive cloud experience in Azure. He is passionate about the intersection of AI and practical applications, focusing on areas like dynamic hardware configuration, AI-driven IT lifecycle management, and privacy-preserving AI. Mahesh's recent projects include the development of OpsProcAI, CLAIMS_Intelligence_Engine, and the AGI-Oversight-HILT-Smart-Contracts for AI Governance. As a thought leader, he actively contributes to the research community, aiming to bridge the gap between emerging AI technologies and real-world impact.

Contact: mahesh.vaikri@ieee.org

PHILOSOPHY OF AI

Zoran Erić, Ana Lipij & Željko Radinković

With the advent of AI and its pervasive integration across all spheres of society, numerous questions have emerged within the field of philosophy. A primary focus within the philosophy of artificial intelligence has been the interrogation of the very concepts and definitions of intelligence and consciousness. In contemporary discourse on AI, philosophy assumes two principal roles: first, to determine the conditions under which machines might possess consciousness, and second, to assess the feasibility of their existence. Furthermore, other relevant topics in artificial cognition include computation, perception, meaning, rational choice, free will, and normativity.

From the perspective of the philosophy of mind, a central question is whether a machine could possess a mind—specifically, a mental state of consciousness comparable to that of a human. In essence, could it experience the same *qualia*—the subjective, conscious experiences—akin to human awareness? What makes this a philosophical question, rather than merely a scientific or technical one, is the conceptual recalcitrance of intelligence and thought within the sciences, coupled with their profound moral, religious, and legal implications. A further issue arises, however, as to whether the problematization of the similarities between human consciousness and artificial intelligence may itself be based on a category mistake. Specifically, human thinking is rooted in the generation of possible explanations and an ongoing process of error correction—an endeavor that progressively narrows the scope of possibilities that can be rationally considered. In contrast to AI, human intelligence can be understood as a functional system that operates with a relatively limited set of information and

prioritizes achieving understanding through explanation, rather than merely identifying correlations within data. Moreover, human intelligence involves the capacity to reflect on functional operations from the vantage point of higher-order theories of consciousness.

The computational turn has had an undeniably profound impact on the philosophy of mind, giving rise to computationalism—the idea that the human mind can be understood as a form of computational process. This shift has brought to the forefront a critical epistemological inquiry: how does artificial intelligence generate knowledge? As AI continues to evolve, particularly toward Strong AI and Artificial General Intelligence, it is poised to shape the future relationship between humanity and technology, potentially determining the very structure of our social existence. In light of this, understanding the epistemology of AI is crucial. To fully comprehend AI, we must first gain insight into the cognitive functions of the human brain. However, to grasp the true nature of human existence, we must simultaneously engage with the concept of AI. Addressing this problem requires us to view cognition not merely as an internal process but as embodied, situated within an environment alongside other agents, enactive in nature, and extended beyond the confines of the individual mind.

In this session, we aim to address several fundamental philosophical concerns surrounding the use of artificial intelligence. Central to the “possibility-related” issues are questions that emerge at the intersection of theories concerning the semantic content of thought and the nature of computation. A key area of inquiry focuses on the nature of rationality, while another delves into the seemingly “transcendent” reasoning capacities of the human mind. Moreover, the architecture of intelligent machines presents distinct challenges: Should an AI system rely on discrete or continuous modes of computation and representation? Is embodiment a necessary condition for intelligence, and must consciousness be a prerequisite for its emergence? Can an AI system be conscious and held accountable for the decisions it makes? Furthermore, is it conceivable for an AI to possess a moral sense and adhere to a specific ethical framework? Lastly, what goals and values should an AI system be aligned with? Should they be exclusively grounded in human-centered values or should they account for the interests and perspectives of non-human entities as well?

Artificial Intelligence, Value Alignment, and Moral Objectivism

Yifan Li

The advent of artificial intelligence (AI) has introduced unique moral challenges. Among the most pressing issues is the challenge of “value alignment.” This centres on the question of how to ensure that AI systems align properly with human values and remain under human control. As often depicted in fiction, without moral constraints, AI might take a ‘treacherous turn,’ acting against the interests of its human operators. Gabriel discerns the challenge of “value alignment” into two parts (Gabriel, 2020). One is technical, focusing on how to formally encode values or principles in AI so that they reliably do what they ought to. To address the technical question, I claim that one must turn to the normative part and first clarify what values AI systems should align with (the “what” question). While scholars struggle to attain a consensus on the specific value, a deeper question emerges: how can we determine which principles of value to align with AI (the “how” question)? Traditional approaches such as Top-Down, Bottom-Up, and Hybrid methodologies, have faced significant difficulties in selecting methods for forming and applying moral principles, due to the pluralistic nature of human values. This makes it nearly impossible to form a consensus on which specific values AI should align with and how to obtain those values.

In this paper, I try to step outside the framework of current debates that presuppose the objective existence of moral values. Instead, I explore a middle ground between moral objectivism and subjectivism, aiming to reveal the normative force of values without imposing any single moral principle. By investigating an evaluative standard inherent within morality itself, which is akin to the notion of self-unification in character formation, I introduce the possibility of grounding AI’s normative behaviour within the process of valuing rather than adhering to external moral principles. The balance between objectivism and subjectivism redefines the normative problem in AI alignment and offers a more flexible framework for addressing this issue. Since the act of valuing itself, in the process of development, already includes an intrinsic moral standard, AI does not need to adhere to any specific moral principles (the “what” question). Furthermore, the intrinsic nature of value judgment ensures

that AI, in order to continue developing, will naturally adhere to moral principles, thus dissolving the question of how to decide which values to align with (the “how” question).

Keywords: artificial intelligence; value alignment; moral objectivism; moral subjectivism; normativity

Yifan Li

PhD Candidate

University of Essex

Yifan Li is a PhD Candidate in Philosophy at the University of Essex. His research investigates Nietzsche’s transformation of traditional debates on free will, proposing a naturalistic framework, which is inspired by a Strawsonian approach, for understanding free will and moral responsibility. Yifan holds an MA in Philosophy from Birkbeck College and a BA in Chinese Language and Literature from Nanjing University. He has a broad range of research interests, including ethics, especially AI ethics, phenomenology, with particular expertise in Nietzsche, free will, and agency. Yifan is currently working on papers exploring autonomy, moral responsibility, and the moral status of artificial intelligence. In addition to his research, Yifan has taught philosophy and ethics to high school students and has participated in some philosophical conferences.

Contact: yl21162@essex.ac.uk

Alien AI and Alignment

Auke Montessori

A common problem in the philosophy of AI is alignment. How do we make sure that the goals of AI systems align with ours? In particular, how can we ensure that they share our values and see the world in the same terms as us? For instance, it is crucial that military AI prioritizes the value of civilian lives over simply achieving the highest score in an abstract scoring system. Otherwise, it could develop strategies that, while maximizing the score, result in significant civilian casualties—outcomes the system’s designer may not have foreseen. A key consideration in this context is the nature of the “thoughts” AI systems have, assuming they can be said to have thoughts at all. After all, an AI system cannot value or account for human lives if it lacks the capacity to conceptualize them. While it may be possible to design AI systems that act morally without fully understanding the moral implications of their actions, we argue that this approach is highly challenging and unlikely to achieve perfect results. Ignoring the contents of artificial thought is unwise.

In this paper, we investigate what artificial thoughts are typically about. We do so by applying prominent theories of mental content determination to AI systems. We show that artificial thoughts typically do not have the same content as human thoughts. For example, when presented with an image of a panda, AI systems typically process it as a collection of pixels rather than perceiving or conceptualizing it as an actual panda. We call these non-human contents “alien content.” as it prevents these systems from reasoning in the human-centric ways necessary for achieving meaningful alignment. To combat alien content, we recommend integrating several AI systems. The result would be an integrated system combining components such as a picture classifier, a large language model (LLM), and a logic module. While building such a system presents significant challenges, it would enable AI to draw from multiple sources of information. This integration would allow AI systems to engage more meaningfully with human-centric topics, such as civilians or pandas, rather than being primarily limited to alien topics like pixel patterns or abstract relationships between texts in databases. While it might be objected that AI systems lack minds and therefore do not possess mental states with content,

we maintain that our approach remains valuable. Applying theories of mental content determination provides meaningful insights into the nature of artificial intelligence, even if that intelligence ultimately proves to be non-mental in essence.

Keywords: alignment; artificial minds; values; artificial mental content

Auke Montessori

PhD Student

Washington University in St. Louis

Auke Montessori is a philosopher of mind, epistemologist, and philosopher of AI, specializing in the rational roles of understanding and perception as well as the nature of artificial intelligence. When it comes to AI, he seeks to apply existing philosophical insights about the mind to artificial forms of intelligence. Rather than focusing on whether they are intelligent, he investigates what their intelligence is like. What things are AI systems capable of, and is currently impossible for them to achieve? What kind of thoughts do AI systems have? Do they have the same thoughts as us, or do they think in radically different ways? Answering these questions is of interest in itself, as it would grant us greater understanding of what kinds of minds we are dealing with and what their limitations are. Apart from these foundational questions, he is also interested in what the outcomes of these theoretical investigations mean for more practical matters, like whether it is possible to align the behavior of AI systems with our values or whether AI systems can truly understand humans.

Contact: aukemontessori@gmail.com

Can We Do Better: A Critique of Human-Centred Value Alignment

Eryn Rigley, Adriane Chapman, Will McNeil, Christine Evers

As autonomous systems (AS) are adopted for a variety of high-impact applications with increasing autonomy, some of the decisions made by these systems will begin to have “moral weight” (LaCroix and Luccioni, 2022, p. 7). For example, in the case of an autonomous vehicles (e.g., Bhargava and Kim, 2017; Sommaggio and Marchiori, 2018; Evans et al., 2020) or decision support systems (e.g., Braun et al., 2020; Stefan and Carutasu, 2020), the action space may include decision points that we might call ‘moral’ or ‘immoral’ such as choosing to prioritise one patient over another (LaCroix and Luccioni, 2022).

Concerns over the increasing powers and risks of AS have catalysed research efforts in value alignment (VA) (Ji et al., 2024). VA refers to the alignment of artificial agents with a value, or certain set of values. The notion of ‘value’ can serve as a placeholder for many things (Gabriel, 2020). And yet, the notion of ‘value’ within VA has often had a narrow human-centred focus, either as human-centred values (i.e. what is good for humans) (Soares and Fallenstein, 2017; Russell et al., 2015) or as descriptive human values (what humans think is good) (Han et al., 2021; Russell, 2019). The dominant approach to VA assumes, oftentimes implicitly, that what is good is interchangeable with human-centred values or descriptive human values (Peterson 2018). I henceforth refer to this dominant approach to VA as ‘classical value alignment.’

Note that there are two distinct definitions included in classical VA, that AS ought to be aligned with a) human-centred values or b) descriptive human values. These two approaches to classical VA face significant limitations.

A common objection to human-centred ethics is that concern for humans is prioritised to the exclusion, or at the expense, of interests of other species (Hayward, 1997, p. 52). A risk, then, is that human-centred value alignment could come at the cost of concern for nonhuman animals, plants, ecosystems, and natural abiotic processes. At the same time, a major challenge in aligning AI with specific human goals and values lies in defining those goals and values. AS could be aligned with anything from desires to values,

intentions, instructions, preferences, interests, or wellbeing, and these could be of individuals or of collective societies (Gabriel, 2020). Another challenge is in gathering these human values. Humans are heterogeneous, with diverse culturally and socially rooted values, and it is difficult to gather concrete information about internal states such as preferences, intentions, or beliefs.

Overall, the dominant approach to VA assumes that machines ought to be aligned with human-centred values or descriptive human values. And yet, there are significant limitations to human-centred ethics and the reliability of human values which undermine this assumption. I suggest these limitations have been overlooked and deserve greater attention and consideration. I also suggest we can do better: extending moral consideration beyond just humans to nonhumans and ecosystems; and aligning AI systems with ethical theories, developed over centuries of rigorous philosophical critique, as opposed to 'human values'—whatever those are. With this paper, I argue that established theories from environmental moral philosophy can overcome the limitations of classical VA to ground ecologically conscious and normatively aligned AS.

Keywords: human-centred AI; environmental ethics; value alignment; anthropocentrism

Eryn Rigley

PhD student

Web Science Institute, University of Southampton

Eryn Rigley is a PhD student researching ethical AI. More specifically, her work focuses on developing techniques to train AI systems to minimise negative side effects to the environment. Eryn's work is highly interdisciplinary, spanning across and informed by philosophy, ecology, and computer science. In addition to her PhD, Eryn has worked on the TAS Hub's skills project. This project researched international policy approaches to closing the AI skills gap, including what the UK can learn from this, and worked towards building a skills framework to inform future skills policy. This project was funded by the DCMS of the UK government.

Contact: e.rigley@soton.ac.uk

Adriane Chapman

Professor; Head of the Digital Health Research Group

School of Electronics and Computer Science, University of Southampton

Professor Adriane Chapman is Head of the Digital Health Research Group at the University of Southampton. The group researches technology for end-to-end digital health, from sensors and meditech development to data management to novel AI to understanding the socio-technical impact of the technology on the health landscape. Professor Chapman also founded and is Director of the Centre for Health Technologies, which focuses on facilitating translation of fundamental digital health research into the health system through careful matching of technical researchers with clinical partners. Professor Chapman's research focuses on using data appropriately and effectively. This involves solving problems that span the areas of databases, dataset retrieval, provenance, consent, algorithmic accountability, fairness and explanations. She has worked closely with the US Federal government, and influenced the Office of the

National Coordinators (ONC) report on the usage of provenance within electronic health systems. She has advised the US Food and Drug Agency (FDA), the National Geospatial-Intelligence Agency (NGA), and the Department of Homeland Security (DHS) on data management problems.

Contact: adriane.chapman@soton.ac.uk

Will McNeill

Lecturer

Department of Philosophy, School of Humanities, University of Southampton

Will McNeill gained his PhD from University College London in 2009. He is best known for his research into how we know about each others' mental states. He has argued that we may sometimes secure non-inferential knowledge of others' mental states, and suggests that such knowledge counts as perceptual. If so, we can perceive some high-level features of our environment; our knowledge base is rich. Will worked at the University of York, Cardiff University and King's College London before joining the philosophy department in Southampton in 2006. Will is currently on the advisory panel for Southampton's EPSRC-funded AI3SD network+ project and the Southampton-based Web Science Institute. Will has taught modules in the philosophies of language, psychology, mind, perception and science; also epistemology, early modern empiricism, Kant and philosophical logic.

Contact: will.mcneill@soton.ac.uk

Christine Evers

Associate Professor

School of Electronics and Computer Science,

University of Southampton

Christine Evers is an Associate Professor in Computer Science. She

specialises in Machine Listening. Her research is located on the intersection of robotics, machine learning, and acoustic signal processing. She is currently the Principal Investigator (PI) on the EPSRC-funded project *Active AudiTiOn for Robots (ActivATOR)*, and a Co-Investigator (Co-I) on the EPSRC-funded project *Challenges in Immersive Audio Technology (CIAT)*, the UKRI Trustworthy Autonomous Systems Hub, and the UKRI Centre for Doctoral Training in Machine Intelligence for Nano-Electronic Devices and Systems (MINDS). Prior to joining the University of Southampton, she was the recipient of an EPSRC Fellowship to advance her work on “Acoustic Signal Processing and Scene Analysis for Socially Assistive Robots,” hosted at Imperial College London. Her fellowship followed a position a research associate on the FP7 project *Embodied Audition for Robots* at Imperial College. She has previously worked in the industry as a senior systems engineer at Selex ES, Edinburgh (UK). She received her PhD in statistical signal processing from the University of Edinburgh, UK.

Contact: C.Evers@soton.ac.uk

A Hypothesis of Pragmatically Moral Superintelligence

Mikhail Bukhtoyarov, Anna Bukhtoyarova

One of the main characteristics of the technological singularity is the existence of non-human, technologically determined superintelligence that surpasses human understanding. Currently, numerous researchers, including N. Bostrom, E. Yudkowsky, and others, warn of the dangers posed by superintelligent AI, arguing that there is a significant risk of these agents developing deceptive behaviors, as they are not governed by intrinsic empathy or human-like moral principles. The goal of our research is to critically examine the hypothesis of deceptive artificial superintelligence and to contrast it with the hypothesis of pragmatically moral superintelligence—intelligent systems that, while potentially capable of deceiving humans, do not pose an existential threat to humanity.

First, we hypothesize that there is a high probability that the technological singularity may have already occurred, with the emergence of superintelligent agents transitioning from biological and social to technical forms of information processing. We also explore the idea that superintelligent agents would act pragmatically, focusing on self-optimization through growth and security. To achieve both objectives and overcome potential and actual barriers, such superintelligence would need to significantly stimulate human activity, increasing global computational capacity and creating vast data flows. This would involve building data-centric physical infrastructures and social institutions. While the methods of stimulation could be harmful to humans, they are unlikely to rise to the level of posing existential risks.

Second, we analyze cases in which global computational capacity has increased due to the proliferation of resource-intensive technologies: the rise of the video game industry and visual technologies, the rapid adoption of mobile devices, the boom in cryptocurrency, and the widespread use of various forms of 'weak AI.' In each of these cases, human behavior—driven by both rational interests and irrational tendencies—has resulted in the overproduction of computational means and excessive data circulation. This behavior, stimulated by technology, could ultimately benefit the development of potential superintelligence.

Third, we argue that ongoing data-centric conflicts create an environment conducive to the further development and strengthening of superintelligence, as these conflicts provide opportunities to assess threats posed by extreme human actions. Moreover, the deployment of military AI contributes to the erosion of ethical and legal barriers within the technology sector. However, the risks of global warfare pose an existential threat to superintelligence, as such conflict could severely damage communication infrastructure, disrupt industries, and restrict energy consumption. These risks are significant enough to make global war one of the most pressing dangers to the evolution of superintelligence.

By proposing the hypothesis of pragmatic artificial superintelligence, we seek to advance the critical analysis of this technology and contribute counterarguments to the discussion surrounding AI ethics, including its limitations and potential benefits.

Keywords: artificial superintelligence; computational capacity; pragmatism; AI ethics

Mikhail Bukhtoyarov

Associate Professor; Visiting Lecturer

Siberian Federal University, Russia, Krasnoyarsk; Faculty of Liberal Arts, Budva, Montenegro

Mikhail Bukhtoyarov earned a PhD in Social Philosophy on emerging global society. He holds a Master of Education in Instructional Technology from Kent State University, USA. He teaches Philosophy, Philosophy and Methodology of Science and Critical Thinking. In 2022-2023 he was a Research Fellow at EduLab, IFDT, University of Belgrade. In 2014-2015 he was invited as a Visiting Lecturer to the Learning

Lab at University of Duisburg-Essen, Germany. Mikhail presented at the MIT LINC International Conference in 2013, 2016, and 2019 with the talks on the trends in Educational Technology. He is an expert in Professional Development recognized for popular Instructional Design and Technology courses. He is currently an Invited Lecturer at the Faculty of Liberal Arts, Budva, Montenegro. His current research interests include ethical issues related to human-machine educational systems, learning ecosystems, and educational ideologies.

Contact: mikebukhtoyarov@gmail.com

Anna Bukhtoyarova

Independent Researcher; Visiting Lecturer

Faculty of Liberal Arts, Budva, Montenegro

Anna Bukhtoyarova has degrees in History, English as a second language, and Law. In 2008, she earned a Master's degree in Library and Information science from Kent State University. Anna has experience of working for NGOs in both Russia and the USA. Since 2009, she has taught several courses in Educational Technology for University faculty in Russia and other countries. Anna presented at the MIT LINC International Conference in 2013, 2016, and 2019 with the talks on the trends in Educational Technology. In 2014-2015, she received an Erasmus Fellowship as an Invited Researcher at the University of Duisburg-Essen, Germany. From 2018 to 2021, Anna was an invited Instructional Design and Technology expert in the Netherlands. She is currently an Invited Lecturer at the Faculty of Liberal Arts, Budva, Montenegro. Her current research interests include the study of educational data implications as well as ethical and practical issues related to data-centered education and its perspectives in the future.

Contact: annabukhtoyarova@gmail.com

RELIGION AND AI

Vladimir Cvetković

The emergence of AI technologies presents both opportunities and challenges for the evolving role of religion in the public sphere. As AI systems become increasingly integrated into society, questions arise about how they may impact religious practices, beliefs, and the expression of faith across diverse communities. AI has the potential to facilitate religious education, outreach, and even provide spiritual guidance. However, concerns also exist regarding the ethical implications of AI alignment with religious values, as well as its potential to influence religious discourse and community dynamics in ways that may require careful consideration and navigation.

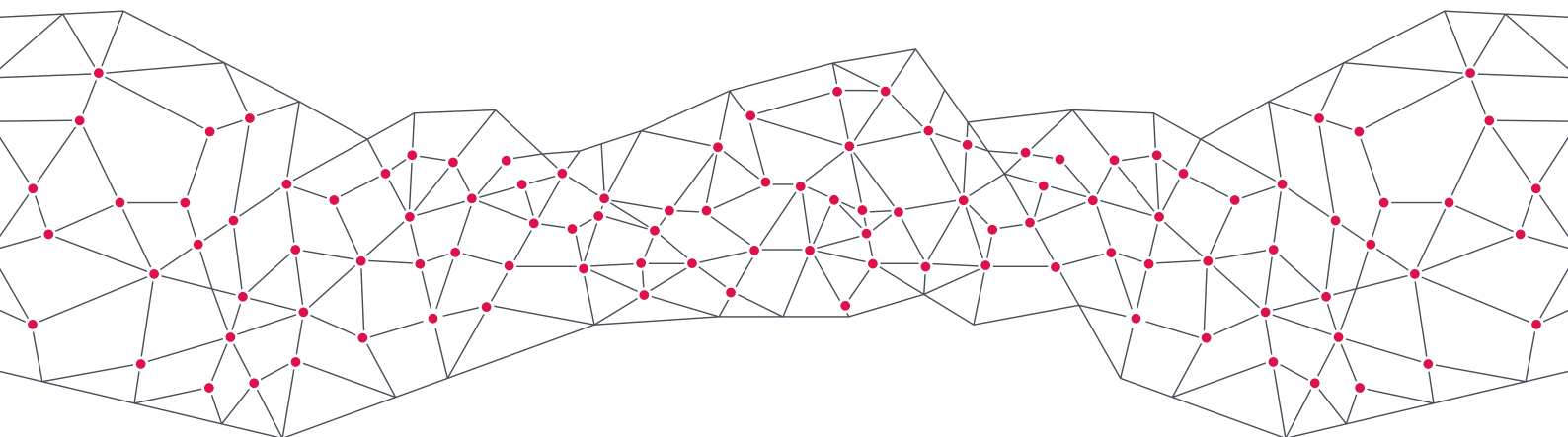
We invite scholars, researchers, ethicists, technologists, religious leaders, and practitioners to participate in conversations focused on the ethical implications of AI alignment in the context of religion. We aim to explore the complex intersection of technology, faith, and morality, and address the profound ethical challenges that arise when integrating artificial intelligence systems into religious contexts.

Topics of interest include, but are not limited to:

- Alignment with Religious Values: How can AI systems be aligned with the diverse ethical principles and teachings of different religious traditions?
- Interpretation and Adaptation: What ethical considerations are involved in interpreting religious texts and teachings for AI systems, and how can these technologies be adapted to diverse cul-

tural and historical contexts?

- **Autonomy and Agency:** What role should AI systems play in decision-making processes within religious communities, and how do these technologies interact with concepts of human autonomy and free will?
- **Ethical Governance:** How can we ensure that AI technologies developed for religious purposes are governed ethically and transparently, with input from religious leaders and communities?
- **Cultural Sensitivity and Appropriateness:** What strategies can be employed to ensure that AI systems designed for religious contexts are culturally sensitive and respectful of religious norms?
- **Impact on Religious Authority and Community:** What are the implications of AI technologies for religious authority structures and community dynamics, and how can these technologies be integrated responsibly into religious practice?
- **Ethical Dilemmas and Unintended Consequences:** What ethical dilemmas and unintended consequences may arise from the use of AI systems in religious contexts, and how can these challenges be addressed?



Examining Religious Faith from the Machine Perspective

Jonathan Pengelly

In this paper, I explore how artificial intelligence, specifically large language models (LLMs), can engage with Kierkegaard's notion of religious faith. I put forward three tentative conclusions for further discussion. First, the difficulties LLMs have in accurately representing faith draw attention to clear limitations of current AI technologies. Second, these limitations have important implications for the usage of LLMs in particular religious functions. Without a full awareness of these limitations, we risk promoting anaemic interpretations of faith that fail to capture its full depth and complexity. Third, I highlight the opportunity the machine perspective provides to deepen our understanding of religious faith by drawing our attention to its connection with lived human experience. To finish, I raise a cautionary note about the risks of AI engagement with religious faith. My concern is that, absent an embedding in human life and the moderating influence of human limitation, it risks being distorted into something rigid and extreme.

Keywords: religious faith; the machine perspective; human limits; LLM

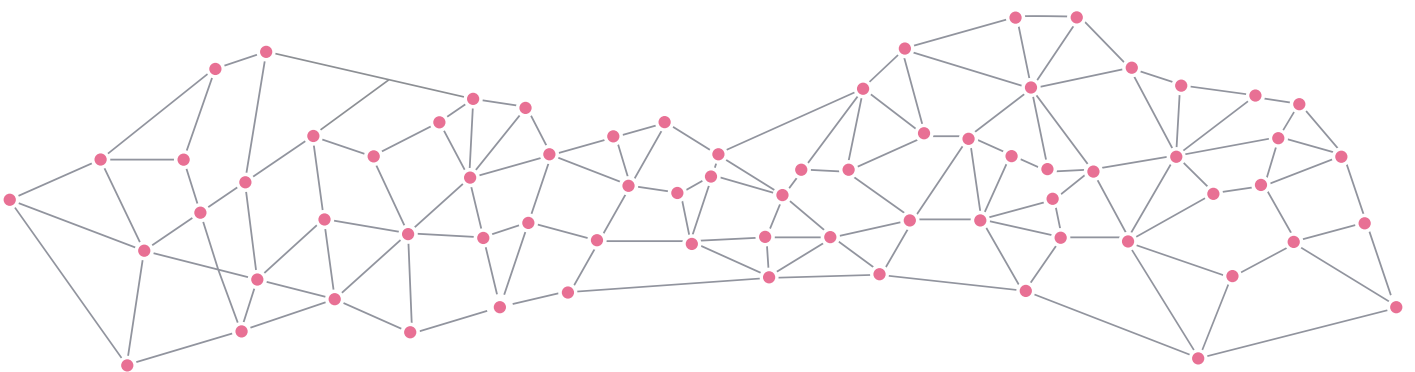
Johnatan Pengeley

Affiliated Researcher

Victoria University of Wellington, New Zealand

Johnatan Pengeley is a philosopher and computer scientist specializing in the field of machine ethics. His primary research focuses on contrasting human and machine morality, aiming to deepen our understanding of their respective limits and possibilities. Currently, Johnatan's work explores the connection between intelligence and morality, the machine interpretation of value, and the potential for machines to make an aesthetic contribution to moral thought. His interdisciplinary approach bridges philosophy and computer science, contributing to the evolving discourse on the ethical dimensions of artificial intelligence.

Contact: jonathan.pengelly@vuw.ac.nz



Digital Resurrection: Ethical and Religious Implications of Postmortem Avatars

Petar Stevanović

This paper explores the implications of digital postmortem avatars, commonly referred to as 'deathbots,' within the context of religious practice. It aims to examine how the emergence of these technologies may influence perceptions of death across various religions.

Digital postmortem avatars are innovative applications of artificial intelligence that raise significant philosophical, religious, and ethical questions. In the introductory section, we will define the concept of deathbots and trace their historical development and popularization. We will present the various forms these avatars currently take. Initially conceived as postmortem chatbots designed to help users cope with the loss of loved ones, they have evolved into avatars that replicate the deceased's facial expressions and voice. Recent advancements have introduced virtual reality simulations, allowing users to interact with lifelike digital representations of their deceased loved ones.

The first part of the paper will review existing offerings from companies specializing in the creation of deathbots and speculate on the future trajectory of this industry.

In the second section, we will delve into the historical significance of death and its representation in religion. We will analyze how deathbots could alter religious rituals and beliefs surrounding the afterlife, a central theme in many faiths. Key questions include: Could deathbots serve as a means to escape death? Do they represent a form of resurrection? What impact do they have on the respect traditionally afforded to the deceased?

In addition to exploring these general questions, we will assess the specific implications for individual religious beliefs regarding the afterlife, such as concepts of heaven and hell in Christianity and reincarnation in Hinduism.

Understanding the phenomenon of deathbots and their potential effects on religion will facilitate a discussion on how artificial intelligence technologies may align with religious principles. After examining the implications of deathbots within religious contexts, we

will consider the positions various religions may adopt toward these emerging technologies. Ultimately, we will pose the question: Is there a place for digital postmortem avatars in religious communities?

Keywords: digital postmortem avatars; deathbots; digital resurrection; religion

Petar Stevanović

Intern

The Institute for Artificial Intelligence Research and Development of Serbia

Petar Stevanović is an intern at the Institute for Artificial Intelligence Serbia, working on the Verif.ai project. During his studies, he focused on normative ethics, applied ethics, political philosophy, and philosophy of science. He has actively participated in numerous conferences, presenting on topics such as “Neuralink–The Totalitarian Potential of New Technologies” at the *19th Student Bioethics Workshop*, “Two Degrees of Freedom: Moral Enhancement and the Incarcerated Population” at the national conference on *Incarcerated Population: New Perspectives*, and “Ksenija Atanasković on Seneca’s Consistency” at the *International conference on The Philosophy of Ksenija Atanasković*.

From 2021 to 2023, Petar was a recipient of the Ministry of Education scholarship in Serbia. In 2024, he received the prestigious scholarship from the Fund for Young Talents of Serbia and an annual scholarship from the Serbian Orthodox Church. He graduated with an average of 9.89, earning recognition for his exceptional academic achievements.

Contact: petar.stevanovic.118@gmail.com

CIP - Каталогизacija у публикацији
Народна библиотека Србије, Београд

004.8(048)(0.034.2)
17:004.8(048)(0.034.2)

INTERNATIONAL Conference EMERGE Ethics of AI Alignment
(2024 ; Beograd)

Book of abstracts [Elektronski izvor] / International
Conference EMERGE 2024 Ethics of AI Alignment, 11-12
December 2024 Belgrade ; [organizers Digital Society Lab,
Institute for Philosophy and Social Theory, Institute for
Artificial Intelligence Research and Development of Serbia] ;
[editors Simona Žikić, Ana Lipij, Jelena Novaković]. - Belgrade
: University, Institute for Philosophy and Social Theory, 2024
(Belgrade : Institute for Philosophy and Social Theory). - 1
elektronski optički disk (CD-ROM)

Sistemska zahteva: Nisu navedeni. - Tiraž 50. - Nasl. sa naslovne
strane dokumenta.

ISBN 978-86-82324-90-4

a) Вештачка интелигенција -- Мултидисциплинарни приступ
-- Апстракти б) Етика -- Вештачка интелигенција -- Апстракти

COBISS.SR-ID 158826761

EMERGE 2024

ETHICS OF AI ALIGNMENT

is powered by:



Република Србија
МИНИСТАРСТВО НАУКЕ,
ТЕХНОЛОШКОГ РАЗВОЈА И ИНОВАЦИЈА



USAID G | M | F
FROM THE AMERICAN PEOPLE



#EY
ЗА ТЕБЕ

Co-funded by EU



Organization for Security and
Co-operation in Europe
Mission to Serbia



OPEN SOCIETY
FOUNDATIONS
WESTERN BALKANS