Book of Abstracts

19-20 November 2026
Faculty of Croatian Studies
University of Zagreb, Croatia

AI PERSPECTIVES

International Conference on Philosophy, Society and Technology

ai-perspectives.org

International Interdisciplinary Conference

AI Perspectives 2026

International Conference on Philosophy, Society and Technology

19-20 November 2026

UNIVERSITY OF ZAGREB
FACULTY OF CROATIAN STUDIES

Zagreb, 2026.

AI Perspectives 2026

International Conference on Philosophy, Society and Technology

Conference organiser

University of Zagreb Faculty of Croatian Studies

Publisher

University of Zagreb Faculty of Croatian Studies

Borongajska cesta 83d, 10 000 Zagreb, Republic of Croatia

For the Publisher

Assoc. Prof. Dario Vučenović, PhD

Organizing Committee

Assoc. Prof. Marko Kardum, PhD

Lovre de Grisogono, PhD

Asst. Prof. Erik Brezovec, PhD

Assoc. Prof. Ivana Greguric, PhD

Assoc. Prof. Željka Metesi Deronjić, PhD

Asst. Prof. Vanesa Varga, PhD

Aco Momčilović, MA

Lucija Pušonjić, mag. educ. phil.

Ana Funarić, univ. bacc. soc.

Mariela Markić, bacc. edu. phil.

Scientific Committee

Assoc. Prof. Marko Kardum, PhD

Lovre de Grisogono, PhD

Asst. Prof. Erik Brezovec, PhD

Amnon Dekel, PhD

Prof. Jean-Gabriel Ganascia, PhD

Asst. Prof. Andrej Jandrić, PhD

Prof. Damir Krstinić, PhD

Prof. Michael R. Lissack, PhD

Assoc. Prof. Sandro Skansi, PhD

Asst. Prof. Vanesa Varga, PhD

ISBN 978-953-8349-93-5 (online)

Book of abstracts is designed by

Bruno Šagi, univ. mag. hist.

© Sveučilište u Zagrebu Fakultet hrvatskih studija, Zagreb, 2026.

No part of this book may be reproduced, photocopied, or replicated in any way without the written permission of the publisher.

All abstracts published in this book represent the opinions of the authors, and do not reflect any official policy of the Publisher. Accordingly, the Publisher disclaims responsibility for the views, statements, and arguments presented in this book.



Content

Keynote Speakers

Amnon Dekel: Intersections Between AI and Human Behavior: Toward Human-Centered Principles for an Age of Accelerating Disruption 7

Damir Krstinić: AI on the Fire Line 8

Michael R. Lissack: Always Question: Trustworthy Human–AI Collaboration as the Contesting of Ground 10

Kristina Šekrst: Who put I in AI? Five Neurons and the Philosophy of Mechanistic Interpretability 12

Conference Presenters

Dannish Kashmiri: A Future of Democracy without Voting: In Favour of Political Consensus through AI..... 14

Lucija Pušonjić Marečić: Aesthetics and Eroticism in the Digital Age: Between Human Communication and Algorithmic Manipulation..... 15

Aco Momcilovic, Lovre de Grisogono: After Work, Beyond Scarcity: Meaning, Status, and Shared Reality in an AI-Enabled Post-Work Society, Toward Existential Experience Design as a New Social Function 16

Ammar Younas: AI and Cultural Pluralism: Rethinking Universalism in Artificial Intelligence Ethics 17

Stanford Howdyshell: AI and the Undermining of Moral Competence..... 18

Petra Lang Dumančić: AI Explorers: Integrating the English Language Curriculum with CARNET's BrAIn Curriculum in 5th Grade Primary School in Croatia 19

Denisa Reshef Kera: AI in the Age of Reservoir Computing: Reopening the Scandal of Philosophy 20

Dario Vučenović, Katarina Jelić, Maša Tonković Grabovac, Sandro Scansi: Artificial Intelligence as a Generator of Psychological Assessment Criteria: Opportunities, Limits and the Epistemic Problem of Nonhuman Judgment 21

Filip Car: Artificial Intelligence in Chemical Engineering: Applications, Benefits, Limitations, and Safety Challenges 22

Siqi Wang: Can a Good Reason Arrive Too Late? AI, Rationalisation, and Epistemic Repair 23

Nicolò Cristante: Commitment, Not Correspondence: An Inferentialist Account of Chain-of-Thought Faithfulness 24

Noa Škerbić: Educating Students in Responsible Use of Technology: A Framework for Technology Ethics in Engineering Education 25

Josip Ćapin: Efficiency at All Costs? AI and the Perils to Democracy 26

Marko Kardum: From Simulation to Understanding: Rethinking Searle's Chinese Room Argument in the Age of Large Language Models 27

Mariela Starčević: From the Factory Floor to the Neural Network: Labour, Alienation and Human Dignity in Rerum Novarum and Magnifica Humanitas 28

Yulu Liu: From the Patient-Subject to the Data-Subject: Deleuze's Theory of Control Societies and the Critique of AI Medical Ethics 29

Athanasia Orfanioti: Human Agency Capital as a Strategic Enterprise Asset: A Corporate Digital Responsibility Framework for Enterprise Value Creation 30

Jelena Vedriš, Karla Pavlović, Matija Marić, Karla Gregurić, Maša Tonković Grabovac: "I Am Afraid That AI May Replace Humans": Examining the Psychological Implications of AI Anxiety in the Workplace 31

Erik Brezovec: It Is Alive – or Is It? Is Artificial Intelligence An Autopoietic System? 32

Karla Žagi: Learning in Times of Generative AI: Critical Thinking, Safety and Ethics in Education 33

Matija Mato Škerbić: Learning to Think Without a Prompt: Switching Off AI in Philosophy and Ethics Education 34

Tiago Faleiro: Living with AI: Adaptation, Dependence, and Agency 35

Sara Nocent: Meeting the Double: The Human Digital Twin at the Crossroads of AI, Body and Presence 36

Joshua Jo Wah Yen: No Experience of Remorse: Dirty Hands, Moral Phenomenology, and The Hard Limits of AI Ethics 37

Paul Benedict Cano: On the Process-to-Token Justification Problem in Computational Reliabilism 38

Ivan Burić, Vanesa Varga, Ana Marija Dunaj: On the Threshold of a New Stratification: Do the Middle Classes Fear Artificial Intelligence? 39

Luka Čizmić, Davor Lauc: Prompt Framing in Multimodal Language Models and the Limits of Testing Quineian Referential Indeterminacy 40

Ana Kovačić: (P)sycophant, Qu'est-ce que c'est? How AI Kills the Argument 41

Katia Zatini: Recursive Embedding and the Governance of AI Transitions: Labour Inequalities and Territorial Divergence in Italy 42

Litianyi (Daley) Zhao: Rethinking Human Creativity in the Age of Generative AI: A Narrativist Account of Creative Agency 43

Jakov Erdeljac: Scientific Production in the Age of AI: Authorship, Review, and the Transformation of the Idea of Science 44

Sergey Romanov: The AI-Mediated Self: Who Shapes the Person Who Decides? 45

Jar Žiga Marušič: The Convergent Viability Thesis 46

Teo Novellino: The Immediacy Problem: Artificial Intelligence and the Changing Temporalities of Human Relations 46

Tomislav Filipović: The Precarious Artist: AI and the Transformation of Creative Work . 48

Paško Lovrić: What Does It Take for AI to Be a Good Psychologist? Therapeutic Potential and Current Limitations 49

Ivo Marinić-Kragić, Andrej Zvonimir Tomić, Drago Bulić, Teo Čolović: What Language-Model Tutors Tell Students When the Course Material Is Wrong 50

Mykola Latansky: When AI Helps Form Our Ends: Authorship Beyond Final Approval .. 51

Lin Sen: When Deferral Closes: The Vivian Model and Moral Uncertainty in Human–AI Control 52

Alisa Zviagina: “Who Told You That?”: Artificial Intelligence and the New Problem of Testimony 53

Amnon Dekel, PhD

Keynote Speaker
ASPER-HUJI Innovate, Hebrew University of Jerusalem, Israel

Biography

Dr. Amnon Dekel is the Executive Director of ASPER-HUJI Innovate, the Innovation and Entrepreneurship Center of the Hebrew University Jerusalem. Previously, Amnon served as the Chair of the Dept. of Software Engineering at the Shenkar College of Engineering & Design in Israel. Amnon has over 20 years of experience in product development of IT based services and applications, as well as more than 15 years of experience in Applied Research in the areas of User Experience and Mobile Computing. Amnon has participated in founding 3 companies in the area of video sharing, personal networking and indoor navigation. Amnon has received numerous international awards, including a Pulitzer nomination for his work with the New York Times. Amnon has a PhD in Computer Science from the Hebrew University. When he has time, he partakes in creating technology based art.

Intersections Between AI and Human Behavior: Toward Human-Centered Principles for an Age of Accelerating Disruption

The rapid diffusion of generative artificial intelligence represents not merely another stage of technological development, but a significant acceleration in the pace at which technological disruption enters everyday life, organizations, and labor markets. This talk examines the intersection between AI-driven disruption and fundamental dimensions of human behavior, with particular attention to work, meaning, personal agency, and social stability.

Damir Krstinić, PhD

Keynote Speaker
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture (FESB), University of Split, Croatia

Biography

Damir Krstinić received his Ph.D. from the University of Split, Croatia, in 2008. He is a Full Professor at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture (FESB), University of Split, where he heads the Department for Modelling and Intelligent Systems. His research focuses on computer vision, image understanding and machine learning. Damir Krstinić has co-authored more than 50 scientific papers. He is the project leader of the Wildfire Early Detection and Monitoring System, developed for Croatian Forests. In operation since 2018, the system now covers more than 150 monitoring locations in Croatia, Bosnia and Herzegovina, Montenegro and Albania.

AI on the Fire Line

Wildfires are among the most destructive natural threats of our time. Climate change is extending fire seasons, while regions that rarely faced large-scale wildfires in the past are increasingly exposed to this danger. The Mediterranean remains particularly vulnerable. In Croatia, where dense pine forests surround scattered coastal settlements and summer populations grow by an order of magnitude, virtually the entire coastline constitutes a wildland–urban interface. A delay of just ten minutes in detection and precise localization can mean the difference between a minor fire and a catastrophe.

Early detection is therefore crucial for successful intervention, but it is only one link in the chain. Equally important is information system support across the entire wildfire management cycle: before the fire, through risk assessment and risk management; during the fire, through GIS information, fire spread prediction, and real-time meteorological data supporting intervention management; and after the fire, through post-fire analysis that turns every incident into knowledge for the future.

We present an integrated intelligent system for early wildfire detection and surveillance, developed in Croatia through more than two decades of research and designed from the outset in dialogue with firefighters. AI-based analysis of video streams detects the first visible signs of smoke and fire within minutes of ignition, at distances up to ten kilometres or more. The system currently operates at almost 150 locations with near 300 cameras in Croatia. Smaller systems have also been deployed in neighbouring countries.

For every AI algorithm, data is the key ingredient: without it, models cannot recognise patterns or learn. Deployment at this scale provides a unique advantage: a large number of real-world image sequences collected daily, under all atmospheric and seasonal conditions. This continuously growing archive of real wildfire footage is used to train and improve detection

models, creating a virtuous cycle in which the operational system feeds the research that makes it better.

However, technology alone does not solve the problem: the system is a tool, not an autonomous solution. AI detects, informs, and predicts; trained people confirm, decide, and act. Its true value emerges when firefighters and other services integrate it into their procedures to respond faster, safer, and more effectively.

Michael R. Lissack, PhD

Keynote Speaker
Second Order Science Foundation
Tongji University, Shanghai, China

Biography

Michael R. Lissack is Executive Director of the Second Order Science Foundation and Professor of Design and Innovation at Tongji University, Shanghai. An applied philosopher with more than 100 publications, he is a past President of the American Society for Cybernetics (2014–2020), founder of the Institute for the Study of Coherence and Emergence and of the journal Emergence: Complexity and Organization. His book Questioning Understanding, Understanding Questioning (2025) develops the claim running through most of his work: that understanding is evidenced by an agent's ability to articulate the Fundierung between a synecdoche and the attended-to substrate that affords it meaning. His earlier career includes Wall Street, as a Managing Director in municipal derivatives at Smith Barney; real estate brokering, whistleblowing; and the technology venture Webmind, where he was CFO. In 2026 he was appointed to the Expert Committee of the Shanghai International Design 100 Think Tank. His Zagreb keynote argues that questioning the Fundierung, with consequences, is the core of trustworthy human–AI collaboration.

Always Question: Trustworthy Human–AI Collaboration as the Contesting of Ground

Debates about trustworthy AI usually locate trust in properties of the system — accuracy, robustness, explainability. This paper argues that trust in human–AI collaboration rests instead on a capacity the system structurally cannot exercise, and that the human must therefore keep exercising: the ability to question, with consequences, the grounding (Fundierung) that affords meaning to any compressed representation.

A large language model's output is a synecdoche: a compression that stands in for far more than it states, and that is effective precisely because it sheds the grounds that would justify it. Such an ungrounded compression cannot be walked back to what founds it. The model can defend its wording but cannot explain why this framing rather than an alternative, because the alternatives are attended-to "as-ifs" that demand abduction and interpretation before any decision. Fluency and efficacy are thus fully compatible with an absence of understanding: the system does not care which grounding obtains, because for it none does.

An operational criterion follows, recovered from phenomenology (Husserl's Fundierung), Vaihinger's as-if pragmatism, Rosen's modeling relation, and Juarrero's enabling constraints: to understand is to be able to interrogate the grounds of a claim through consequences — interventions in a model of the world, not inferences within a formal system. "Why" is not "how," and "why it matters" is not "how does this occur"; conflating the two registers is what lets a simulacrum of judgment pass as the thing itself.

The upshot is a practical discipline we summarize as "always question": treat AI output as an as-if to be tested rather than an answer to be adopted; surface the uncritically examined presuppositions and enabling constraints behind its framing; and reserve consequential, intervention-bearing judgment for the human. This reframes explainable AI away from post-hoc rationalization and toward making the tacit grounds of a claim contestable, and it relocates trustworthiness from the model to the human–AI relation and the questioning practice that sustains it. The account is illustrated with characteristic failure cases in which fluent, confident output proves untethered from any recoverable ground.

Kristina Šekrst, PhD

Keynote Speaker
Center for Cognitive Science, University of Zagreb, Croatia

Biography

Kristina Šekrst is an adjunct Assistant Professor at the Center for Cognitive Science, where she leads the QUALIA lab on AI alignment and interpretability. She is Machine Learning Lead at the American nonprofit Meedan and Principal Software Engineer at Preamble AI, specializing in artificial intelligence. Her education spans philosophy, cognitive and comparative linguistics, and Croatian language and literature, and she earned her PhD in logic with a dissertation on formalizing evidentiality in evidential languages, defended summa cum laude. Her industry career of more than a decade began in computer vision, and her engineering work contributed to the discovery and patenting of prompt injection in AI systems. She is the author of The Illusion Engine: The Quest for Machine Consciousness (Springer, 2025) and co-author of Introduction to Deep Learning: Neural Networks, Large Language Models and Agentic AI (with S. Skansi, 2026). She co-founded the Zagreb School of Egyptology and co-authored a grammar of Middle Egyptian (with I. Uranić, 2014). She is editor of Philosophical Problems in Science, Open Forum editor at AI & Society (Springer), and leaf category editor for Deep Learning at PhilPapers. She currently collaborates on projects in the pragmatics of South Slavic languages, is a part of the philosophical working group of the European Astrobiology Institute, studies astrophysics at the Open University, and mentors on Caltech's online courses on the Solar System. Currently, her work focuses on developing and fine-tuning large language models for underrepresented languages, mainly Arabic and Hindi.

Who put I in AI? Five Neurons and the Philosophy of Mechanistic Interpretability

It seems obvious that you could not explain what your brain is doing if you took five of its neurons, connected them so that they could carry out a simple logical task, and then tried to do so. Yet if you had a far greater number of neurons, it might be possible, but it is difficult to know where the sorites soft spot is between five and eighty-six billion neurons. Mechanistic interpretability research on large language models has a comparable issue since it attempts to understand a model by examining the circuits it can trace. The method, however, does not specify how much of the model has to be traced before conclusions about the entire model can be reached (Šekrst, 2025b).

In June 2022, an engineer at Google asked a language model whether or not it was sentient, and the model replied that it was. The engineer believed this and then spread the word about it before losing his job. This situation could have acted as a warning regarding the idea of judging machine intelligence by observing its behavior (Šekrst, 2025a), yet the industry and the researchers continued to ask the same question, making use of both signs of consciousness (Butlin et al, 2023) and self-reports (Perez and Long, 2023). The only difference nowadays is in the answer: developers now train the models to say that they are not conscious and use those denials as proof, while the same post-training techniques – supervised fine-tuning and

reinforcement learning from human feedback – are used to shape all the other statements the model makes so as to conform with what the human raters expect. In the end, we assess machines on the basis of the sentences we have taught them to produce.

It is necessary to find out where expressions such as "I feel" and "I am conscious" originate. One person included them in the pretraining data, another chose them during the post-training stage, a reward model became accustomed to preferring them on the basis of human evaluations, and a policy model then learned to produce them when required. Our actual preferences have now been substituted by a stand-in which may move away from our desires as the system is being optimized (Gao et al, 2022). In the case of open-weight models, it is possible to follow the path from the training data to the output (Šekrst, 2026). With the most advanced models, however, this is not feasible, which is one of the reasons why interpretability research began in the laboratories that develop these systems and why its findings are important, such as alignment faking, attribution graphs, or concept injections (Greenblatt et al., 2024; Lindsey, 2025; Lindsey et al., 2025).

When considering the example above with five neurons, interpretability frequently makes general conclusions based on a small portion carrying out a simple task, but this can be misleading regardless of the direction of the argument. This talk explains how the training process shapes the models so as to produce the results we desire, and how interpretability is at times used to claim that the models are conscious even in the absence of any real evidence: when you only have a hammer, everything starts to look like a nail – or at least like an indicator of consciousness.

A Future of Democracy without Voting: In Favour of Political Consensus through AI

Dannish Kashmiri
Independent Affiliation, United Kingdom

Abstract

Imagine a democratic process that produces no winning party or politician, one without a ballot to determine who wins and who loses. Imagine, instead of voting, all citizens can express their political concerns and participate in a democratic process that is more meaningful than ever before. Imagine if all those concerns are heard and directly contribute to forming political and government policy. In this paper, I present how artificial intelligence could enable this process and why it has clear advantages over contemporary democratic voting. Building on emerging work in AI-supported mass deliberation, I propose a non-electoral system with three institutional elements. First, citizens participate directly through an accessible political-consensus process. Structured questions and free responses allow them to register concerns about their lives and particular matters of public policy. Citizens can also contribute concerns throughout the year, allowing their contributions to shape the eventual consensus form rather than leaving its agenda to a single author or institution. Second, AI analyses these contributions to map convergence and coherent disagreement, then produces an advisory policy-guidance report. Its procedure draws on reflective equilibrium to seek coherence among citizens' considered judgments, relevant principles and available evidence. A contractualist constraint excludes policies that any affected person could reasonably reject. This protects minorities and non-citizens against majoritarian domination without requiring unanimous agreement. Third, an extended, unelected civil service interprets the guidance and implements feasible policies. AI neither votes nor rules. It serves as a transparent calculator, while political authority remains with the democratic collective and human officials remain responsible for interpretation and implementation. Published guidance reports allow citizens to assess whether officials have followed the collective guidance. Citizen audits, independent technical review and judicial oversight provide checks on both the AI system and the civil service. I argue that this model gives citizens a more substantial democratic role than periodically choosing among parties or candidates. Replacing winner-take-all elections also weakens incentives that encourage politicians to polarise disagreement, make populist appeals and subordinate policy to gaining or retaining office. I answer four central objections.

Aesthetics and Eroticism in the Digital Age: Between Human Communication and Algorithmic Manipulation

Lucija Pušonjić Marečić
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

This paper explores the contemporary transformation of the relationship between aesthetics and eroticism through the theoretical framework of Zlatko Posavac's philosophy and contemporary digital media studies. Drawing on Posavac's thesis that aesthetics in interpersonal relations constitutes a vital form of communication, one that ennobles and liberates the erotic impulse. The paper analyzes how these concepts are being redefined in the era of "cultural software" and algorithms. By synthesizing philosophical insights with contemporary research on algorithmic culture, AI technology, and digital platforms, the study examines the shift from authentic aesthetic expression to what Posavac terms the "aesthetics of the commodity" (Warenästhetik) and "technical mystification." The paper analyzes key processes such as the commodification of intimacy, the reshaping of aesthetic norms, and issues regarding the subject and agency within a digitally mediated environment. The paper highlights the need for a new understanding of aesthetic education and media literacy as forms of resistance against algorithmic anesthesia and the commercialization of intimacy.

After Work, Beyond Scarcity: Meaning, Status, and Shared Reality in an AI-Enabled Post-Work Society, Toward Existential Experience Design as a New Social Function

Aco Momcilovic
The University of Applied Sciences Baltazar Zaprešić, Croatia

Lovre de Grisogono
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

Debates about artificial intelligence and the future of work usually focus on employment, productivity, and income distribution. This paper considers a deliberately optimistic scenario in which AI and robotics produce most goods and services and basic material security is largely independent of employment. It asks what happens when the problem is no longer how humans will work, but what they will do with freedom from economically necessary work. The paper argues that artificial abundance would not eliminate scarcity but transform its significance. Material goods and cognitive outputs may become abundant, while human attention, social recognition, shared reality, credible identity, and the conditions that support meaningful lives remain limited or difficult to sustain. Drawing on philosophy of work and philosophical accounts of meaning, together with psychology, sociology, and research on human–AI interaction, the paper distinguishes material and positional scarcity from what it terms existential scarcity. Meaning itself is not treated as a finite resource; rather, the conditions that enable meaningful engagement may remain difficult to create and maintain even under material abundance. On this basis, the paper proposes Existential Experience Design (EED) as a new interdisciplinary social function: the intentional creation of environments, practices, challenges, rituals, communities, and developmental journeys that support meaningful human engagement, agency, competence, belonging, contribution, and shared reality. EED would not prescribe what makes a life meaningful, but create conditions in which individuals and communities can construct and pursue meaning for themselves. The paper concludes that an AI-enabled postwork society would face not only the economic challenge of distributing abundance, but the philosophical, psychological, and social challenge of enabling people to use unprecedented freedom meaningfully.

AI and Cultural Pluralism: Rethinking Universalism in Artificial Intelligence Ethics

Ammar Younas
Beijing Institute of AI Safety and Governance (Beijing-AISI), China

Abstract

Artificial intelligence ethics is increasingly organized around globally circulating principles such as autonomy, fairness, dignity, responsibility, transparency, and non-maleficence. Yet these concepts are often treated as if their meanings were culturally stable and universally transferable. This paper argues that such an assumption creates a foundational problem for global AI ethics: societies may endorse the same ethical principle while interpreting its content, priority, and practical implications in significantly different ways. The paper develops a philosophical account of cultural pluralism in AI ethics by examining three concepts—autonomy, dignity, and responsibility—across different normative traditions. Rather than treating cultural variation as evidence for ethical relativism, it argues that the challenge is to distinguish between the universal aspiration of ethical principles and the culturally mediated ways in which those principles are understood and institutionalized. A universal principle may therefore remain normatively meaningful without requiring uniform interpretation across societies. To address this tension, the paper proposes a model of context-sensitive normative universalism. Under this approach, AI governance can maintain shared ethical commitments while allowing their interpretation and implementation to emerge through intercultural dialogue, local moral reasoning, and institutional context. This framework rejects both rigid universalism, which risks imposing culturally specific assumptions as globally neutral standards, and unrestricted relativism, which makes meaningful cross-cultural ethical evaluation difficult. The paper contributes to current debates on responsible and trustworthy AI by showing that cultural pluralism should not be treated merely as a problem of representation or localization. It is a deeper philosophical question concerning how ethical meaning itself is constructed across societies. Recognizing this dimension is essential for developing AI governance frameworks that are both globally legitimate and normatively coherent.

AI and the Undermining of Moral Competence

Stanford Howdyshell
Stockton University, USA

Abstract

In this presentation, I argue that the regular use of AI and other decision-making algorithms undermines our moral competence, from the perspective of Virtue Ethics. This loss comes in three parts: (1) a loss of practical wisdom, (2) an alienation from the practices and identities that are central to our conceptions of the good life, and (3) a loss in virtue. First, drawing on work by Nir Eisikovits and Dan Feldman and research from MIT's Media Lab, I argue that the regular use of AI incurs cognitive debt, makes it harder to retain and apply concepts, and removes opportunities for people to practice making practical judgments. These come together to cause a decreased capacity for practical wisdom. Part two shows that a reliance on algorithmic decision-making (which includes many AI systems) prevents us from participating in a practice and achieving the primary good of that practice, which is being the practitioner of the practice, or being the type of being we want to be. The argument is straightforward: The primary good of a practice is being a specific sort of being (i.e. a carpenter, portrait painter, or academic philosopher). To be this sort of being, one must interact with the world in a specific sort of way (i.e. acting like a carpenter, portrait painter, or academic philosopher) and achieve a certain set of constitutive goods. A reliance on AI, though, can (1) prevent us from acting like the practitioner we want to be (for example, by replacing our decisions with those of an algorithm) and (2) prevent us from achieving the constitutive goods of a practice by outsourcing them to an algorithm or AI. Finally, I argue that the loss of our ability to participate in a practice leads to a lack of virtue. This is because virtues are determined, at least in part, by their relationship to a given practice, where virtues are modes of behavior that lead to the reliable development of the goods of the practice. With the loss of the practice, then, one loses the infrastructure and motivation needed to develop virtue.

AI Explorers: Integrating the English Language Curriculum with CARNET's BrAIn Curriculum in 5th Grade Primary School in Croatia

Petra Lang Dumančić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Abstract

Reflecting current trends in education, upper primary school students increasingly use generative artificial intelligence (AI) tools, demonstrating excessive trust in AI along with limited critical AI literacy. Uncritical use of AI prevents students from developing the skills to distinguish accurate from inaccurate or biased responses, and they rarely verify sources or the credibility of information. Therefore, developing AI literacy across all school subjects becomes essential. In teaching English as a foreign language (EFL) in Croatia, the English Language Curriculum (2019) does not mention AI explicitly; however, it clearly emphasises critical thinking, media and digital literacy, and responsible behaviour in the digital environment. AI in primary English language teaching (ELT) cannot therefore be ignored, as it is rapidly emerging and increasingly applied across education. Additionally, strategic documents such as the Digital Education Action Plan (2021-2027), UNESCO's Recommendations (2021), AI Literacy (AILit) Framework (2026) and CARNET's BrAIn curriculum (2024/2025) recognize the need for developing critical AI literacy among students, emphasising digital competences alongside the critical, ethical and responsible use of AI from an early age. This paper argues that connecting English language curriculum outcomes with CARNET's BrAIn curriculum outcomes enables fifth-grade students to explore AI critically through the extracurricular project AI Explorers – Creative Ways of Exploring AI, conducted at Primary School Višnjevac, Croatia. It analyses selected curricular outcomes, project activities, and multimodal products using the AILit Framework. The research shows how primary ELT curriculum outcomes can be interdisciplinary linked with BrAIn curriculum outcomes, and how English language learning can be purposefully intertwined with the critical evaluation of AI-generated multimodal content and ethical AI use, highlighting broader pedagogical implications. Within contemporary glottodidactic practice, the paper shows that AI in foreign language teaching serves not merely as a content generator, but as a catalyst for critical reflection, language production, and students' autonomy and motivation. It offers a practical framework for integrating AI literacy into upper primary ELT while reflecting on the possibilities and challenges during implementation.

AI in the Age of Reservoir Computing: Reopening the Scandal of Philosophy

Denisa Reshef Kera
Bar-Ilan University, Israel

Abstract

Why has "can machines think?" survived 75 years without an answer? Because the question cannot be answered in the vocabulary we have. Since Aristotle ruled that made things have no nature of their own, artifacts have been describable only by reference to something else: nature, which supplies their matter, and their makers, who supply form and purpose. Philosophy never developed a vocabulary for artifacts in their own right, and no theory of knowledge admits a knower that is not a person. AI falls into both gaps at once, so every verdict about it is borrowed. "It thinks" applies a concept built for humans. "It merely simulates thinking" applies the same concept and adds a negation. Both verdicts describe the machine as a person, neither describes the machine. We ground this diagnosis in a case devoid of AI's usual complexity. In 2026, Nature Communications (Han et al. 2026) reported a computer built from tungsten disulfide nanotubes: a synthetic material with no natural occurrence, whose computing structure nobody designed. The tubes settle into a random tangle, and the tangle computes. No code, no training corpus, nothing to talk to. Its makers documented every step of the fabrication and then presented the result as "mimetic optogenetic synapses" matching "the fruit fly's brain": the best documented artifact in the laboratory, introduced as a small anonymous brain. If that is the only available description even here, the difficulty in the AI debate was never intelligence. It was the artifact. The borrowing follows a pattern across an artifact's life. New technologies are described as minds: computers arrived as "electronic brains." Ubiquitous ones are then absorbed into "the environment," as plastic, the most artificial substance we ever produced, has been. The current dispute over AI as "environmental intelligence" (Leonelli 2025; Alvarado 2025) cannot be settled because both positions file AI under a category built for something else: mind for persons, environment for what nobody made. Grounding AI in the environment does not escape the artifact: the AI is one, and the environment itself is known only through the instruments that measure it. This is what our title calls the scandal of philosophy. The scandal is not that machines might think. It is that after 23 centuries we still cannot describe the things we make. New Turing tests and new Chinese rooms will keep appearing until philosophy gives the made thing a nature of its own.

Artificial Intelligence as a Generator of Psychological Assessment Criteria: Opportunities, Limits and the Epistemic Problem of Nonhuman Judgment

Dario Vučenović, Katarina Jelić, Maša Tonković Grabovac, Sandro Scansi
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

Artificial intelligence is widely being used in psychological assessment for test development, item generation, scoring, language analysis, and even clinical interpretation. This presentation explores its potential in clinical and organizational psychology, with particular attention to less-discussed questions, such as: can AI contribute to defining what constitutes a "correct" human response? Traditional assessment methods often rely on expert judgment, consensus samples, or external behavioral criteria. These approaches, while valuable, have limitations, including expert disagreement, cultural bias, subjectivity, and the possibility that consensus reflects social conventions rather than psychological competence. The Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) provides a useful example because determining the correct response to emotional situations is considerably less straightforward than determining a correct mathematical answer. AI could potentially address some of these limitations since it can generate large numbers of assessment items, compare expert and population judgments, identify common ambiguous responses, or offer alternative criteria against external outcomes. However, replacing human experts with AI-generated criteria introduces new concerns. AI systems may reproduce biases present in their training data, generate unstable or opaque judgments, and transform statistical patterns into psychological standards. Our presentation distinguishes between AI as an assessment tool and AI as a source of assessment criteria. It argues that AI-generated criteria require independent psychometric, cultural and ethical validation. A proposed Human–AI Criterion Framework is presented, in which expert judgment, population consensus, AI-generated judgments and external behavioral outcomes are treated as complementary sources of evidence. We argue that the future of psychological assessment should not involve choosing between human and machine judgment since AI does not eliminate the problem of judgment in psychological assessment; it purely relocates it. The key question is not whether AI can generate answers, as the evidence shows it can, but whether we should allow nonhuman systems to determine what counts as "humanly" correct.

Artificial Intelligence in Chemical Engineering: Applications, Benefits, Limitations, and Safety Challenges

Filip Car
Faculty of Chemical Engineering and Technology, University of Zagreb, Croatia

Abstract

Artificial intelligence (AI) is increasingly being integrated into chemical engineering, offering new opportunities for process optimization, predictive modelling, data analysis, and decision-making. However, its growing application also raises concerns regarding reliability, transparency, and operational safety. Beyond technical performance, these concerns extend to human well-being, professional responsibility, and the broader societal implications of AI-driven technologies. The central question is what safety risks arise from implementing AI in chemical engineering and how these risks can be addressed while preserving its practical benefits. Drawing on scientific literature and empirical observations from our own laboratory practice, the analysis examines the advantages and limitations of AI-assisted approaches. Laboratory experiences are considered alongside published research to explore the relationship between theoretical expectations and practical challenges. Particular attention is given to safety risks across five stages of AI implementation, including issues of data quality, model reliability, interpretability, and human oversight. An interdisciplinary perspective further considers how technological decisions may affect workers, communities, and the environment. Questions of accountability, trust, and professional judgement become particularly relevant when AI-generated recommendations influence processes with potential safety consequences. From this perspective, AI implementation is not merely a technical challenge, but also an ethical and organizational one. The analysis emphasizes that responsible AI adoption requires more than technological efficiency. Its long-term value depends on reliable validation, meaningful human oversight, and the recognition that engineering decisions ultimately have consequences beyond the laboratory.

Can a Good Reason Arrive Too Late? AI, Rationalisation, and Epistemic Repair

Siqi Wang[1]
Hefei 168 High School, China

Abstract

A belief can be supported by a good reason without being held for that reason. That much is familiar. What is less clear is whether the basis of a belief can change after the belief has already been formed. Suppose someone accepts a conclusion for a poor reason and only later encounters a genuinely strong argument for it. The conclusion stays the same, but its epistemic standing may not. The question is whether the later reason has merely been recruited to defend an old commitment, or whether it has become part of what now sustains that commitment. Large language models make this problem unusually easy to isolate. A model can be shown evidence and asked what follows, or given the conclusion first and asked to justify it. The two routes may end with the same answer and even the same argument. Yet the argument need not play the same epistemic role. Conclusion-first reasoning is not automatically rationalisation, just as a good reason that arrives late does not automatically repair what came before. The interesting case lies between the two: when does a later reason actually become a reason for which the conclusion is maintained? I examine this through matched reason-first and conclusion-first dialogues. Once both routes have reached the same conclusion with comparable support, I introduce a targeted defeater against the stated reason. What matters is what happens next. If the reason is undermined and the model simply searches for another defence of the same answer, the reason looks more like an attachment than a basis. If the conclusion instead becomes appropriately open to revision, the later reason seems to have acquired a genuine role in sustaining it. I expect the contrast to reveal cases in which equally persuasive reasons differ in the role they actually play. The broader claim is that doxastic justification may have a history. A final answer and a polished rationale do not by themselves show what a judgment is based on. This matters for human-AI collaboration: explanations should be assessed not only by how convincing they look, but by whether the system remains answerable to the reasons it gives. AI offers a clean setting in which to test that difference. A belief may begin badly based without having to remain so.

1 Independent student contribution.

Commitment, Not Correspondence: An Inferentialist Account of Chain-of-Thought Faithfulness

Nicolò Cristante
Ca' Foscari University of Venice, Italy

Abstract

Chain-of-thought reasoning now carries real weight: frontier models are trained to generate step-by-step verbal traces, and monitoring them has been proposed as a safety measure. Yet their faithfulness is in serious trouble. In controlled experiments, models whose answers are driven by biasing cues planted in the context produce plausible explanations that never mention those cues (Turpin et al. 2023; Chen et al. 2025). And when faithfulness metrics are evaluated on tasks with ground-truth intermediate computations, most perform near chance, with different ones leading at the step and whole-trace levels (Gur-Arieh et al. 2026). This paper argues that the trouble is conceptual before it is technical. Much of the debate presupposes a representationalist picture: the chain of thought reports an inner computational process, and faithfulness is correspondence between report and process. The picture faces two difficulties. Empirically, even in simple neural networks, distinct mechanistic explanations can satisfy the same causal criteria (Méloux et al. 2025). Philosophically, in the tradition of Wittgenstein, Sellars, and Brandom, the content of a reason is fixed not by an inner vehicle but by its role in the practice of giving and asking for reasons. Confabulation research suggests a parallel: verbal reports may fail as descriptions of the underlying process, yet reason-giving does not thereby lose its justificatory function. On the inferentialist alternative, a stated reason is a publicly assessable commitment - attributable to the model without presupposing human normative status - rather than merely a report of computation. The evaluative question becomes: is the model bound by what it says? It can be tested by asking whether conclusions remain consistent with avowed premises across contexts, whether the model honours their entailments and incompatibilities, and whether it revises when a premise is successfully challenged. Existing tests that truncate the chain or insert mistakes into it probe causal dependence on its text; I distinguish such causal sensitivity from responsiveness to the inferential content of stated reasons, and outline how to test the latter. The contribution is threefold: a diagnosis of why existing faithfulness metrics diverge, a testable standard of reason-responsiveness, and an account of the implications for chain-of-thought monitoring. The standard complements, rather than replaces, causal investigation of model behaviour.

Educating Students in Responsible Use of Technology: A Framework for Technology Ethics in Engineering Education

Noa Škerbić[2]
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

Abstract

Rapid development and widespread adoption of digital technologies have created a growing gap between technological competence and the ability to use and design technology in ways that support human well-being. Engineering education traditionally emphasizes technical knowledge and the ability to develop technological systems, while comparatively less attention is given to how technological design shapes users' attention, habits, decision-making and everyday behaviour. This raises an educational question: should responsible use and design of technology be treated as a core competence of future engineers? This paper explores the development of an educational framework for responsible technology use and design, with particular consideration of artificial intelligence and other digital technologies. The proposed framework is based on an analysis of relevant interdisciplinary literature and contemporary examples of technology design and use. It distinguishes two complementary dimensions of responsibility: responsibility of technology creators towards users, and responsibility of individuals in their own use of technology. The framework addresses topics including responsible AI use, human-centred technology design, attention- and habit-forming mechanisms, notification and interface design, excessive screen use, digital distraction, and the social and individual consequences of technology use. Particular emphasis is placed on practical methods through which students can critically examine and modify their own patterns of technology use, rather than limiting education to theoretical discussion.

The expected result is a model for an interdisciplinary university course that integrates technological, ethical and behavioural perspectives and treats responsible technology use as an educational competence rather than an optional personal preference. The proposed approach is intended particularly for engineering education, where students are simultaneously future users, developers and decision-makers in the technological systems that shape society. The contribution of the paper is a structured educational framework connecting responsible technology design with responsible technology use, and a discussion of how such an approach could be implemented within engineering curricula.

2 Independent student contribution.

Efficiency at All Costs? AI and the Perils to Democracy

Josip Ćapin
University of Zadar, Croatia

Abstract

The development of civil society in liberal democracies has increased the amount of demands towards democratic governments, i.e., political input. If those demands exceed the capacity of the governments to provide adequate responses, i.e., political output, the governments are faced with demand input overload, as Easton calls it (1965). In the European context, a similar phenomenon is called democratic deficit (Scharpf 1999), and in the context of the USA it is called vetocracy (Fukuyama 2014). Although the thesis about the increase of the input is contested and some, e.g. Mair claim the opposite, namely that political input is decreasing because of citizen disengagement and the weakening of political parties (2013), this talk assumes that the difference between political input and output remains a relevant frame for the analysis of the AI influence on democracy. The lack of effective political output often results in the rise of democratic scepticism and can contribute to the democratic recession (Diamond 2015). The citizens are willing to sacrifice the mechanisms of political input and vote for illiberal or autocratic politicians if they support political output which those politicians provide (Werner and Snagovsky 2025). Simultaneously, the number of autocracies in the world is rising (Hellmeier et al. 2021). Although AI significantly improves political output across both democratic and autocratic regimes, in this talk I will argue that autocracies will benefit more from it. Because of a centralised, hierarchical, non-transparent system, which is combined with the control over information and unrestricted access to citizen data, autocracy can implement AI more effectively in policy-making. This implies that autocracy could secure an advantage over democracy and preserve its political system more easily (cf. Beraja et al. 2023a and 2023b). The main thesis of the talk is that AI, because it primarily improves political output, could create a wider gap in policy efficiency between autocracies and democracies. The use of AI in democracy is structurally constrained by institutional checks and balances, decision-making process and protection of citizen rights. Striving to reduce this gap, to catch up with or surpass autocracy, could become a catalyst for democratic erosion, because such adaptation would require changing democratic rules and institutions.

From Simulation to Understanding: Rethinking Searle's Chinese Room Argument in the Age of Large Language Models

Marko Kardum
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The emergence of large language models (LLMs) has renewed the philosophical debate concerning the distinction between understanding and its simulation. Searle's Chinese Room argument is often invoked to challenge the claim that linguistic plausibility constitutes evidence of understanding. However, its application to contemporary LLMs raises a distinct question: does an argument about what is insufficient for understanding also provide a criterion for determining whether a particular system understands? This question is addressed here through conceptual analysis and a reconstruction of Searle's argument. An ontological claim concerning the conditions necessary for understanding is distinguished from an epistemological claim concerning the grounds on which understanding may be attributed. The central thesis is that the former does not, by itself, establish the latter. Even if formal computation alone is insufficient for understanding, this does not determine which additional properties are required or how their presence or absence should be assessed in transformer-based systems. It will be argued that neither successful linguistic performance nor computational architecture can independently resolve this question. A more strictly defined application of the Chinese Room argument is therefore proposed: claims concerning particular LLMs must specify the concept of understanding at issue, identify the relevant conditions for its attribution, and justify the connection between those conditions and the system being evaluated. The potential contribution of the paper lies in clarifying what must be established before such a judgment can be philosophically justified, rather than in reaching a definitive verdict on whether LLMs understand. This distinction also helps to define more precisely the explanatory scope of Searle's weak/strong AI framework in contemporary debates.

From the Factory Floor to the Neural Network: Labour, Alienation and Human Dignity in Rerum Novarum and Magnifica Humanitas

Mariela Starčević[3]
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The rapid acceleration of artificial intelligence represents a profound epochal shift, prompting urgent questions about the future of human agency, labour and identity. Rather than viewing this transition through a purely technological lens, this paper aims to examine contemporary anxieties surrounding artificial intelligence as part of a long historical continuum of industrial and economic disruption. Specifically, it analyses Pope Leo XIV's 2026 encyclical Magnifica Humanitas (On Safeguarding the Human Person in the Time of Artificial Intelligence) in comparison to its predecessor, Pope Leo XIII's 1891 encyclical Rerum Novarum (On the Condition of the Working Classes) from the time of the Industrial Revolution. By deliberately timing Magnifica Humanitas to coincide with the 135th anniversary of Rerum Novarum, Leo XIV draws an explicit parallel between the 19th-century industrialization of muscle and the 21st-century industrialization of intellect. This paper explores how both historical moments share a core philosophical crisis: the triumph of a "technocratic paradigm" that reduces human labour and cognition to commodities that can be optimised, measured and replaced for maximum efficiency. Ultimately, the paper argues that just as 19th-century societies countered industrial alienation by establishing legal, ethical and social protections for labour, the 21st-century population must similarly reassert social, moral and political boundaries to reclaim human agency and protect human dignity.

3 Independent student contribution.

From the Patient-Subject to the Data-Subject: Deleuze's Theory of Control Societies and the Critique of AI Medical Ethics

Yulu Liu
Sichuan University, China

Abstract

Contemporary AI medical ethics predominantly relies on a principlist, normative agenda focused on privacy, transparency, fairness, and accountability. However, this framework rests on a tacit presupposition: the prior existence of a pre-given "patient-subject" whose autonomous rights must be protected. Drawing on Gilles Deleuze's philosophy, this paper argues that the ethically decisive transformation wrought by AI in medicine concerns the mode of subjectivity itself: a structural shift from the narrating, embodied patient-subject of disciplinary medicine to a continuously modulated, predictive "data-subject". The paper proceeds in four parts. First, utilizing Deleuze's reading of Foucault, it traces how modern clinical medicine historically constituted the patient-subject through disciplinary enclosures and the diagnostic gaze. Second, applying Deleuze's "Postscript on the Societies of Control," it demonstrates how AI reconfigures hospitals into networked control surfaces, converting the clinical body into digital data objects, clinical judgment into algorithmic prediction, and personal narrative into dynamic data profiles. Within this paradigm, individuals are decomposed into "dividuals"—calculable codes continuously modulated through risk stratification and predictive targeting. Third, the paper re-examines the current ethical crises in AI healthcare—namely the crises of autonomy, algorithmic opacity, and the erosion of the doctor-patient relationship. It demonstrates that these are not merely technical glitches to be patched by bioethical safeguards, but direct manifestations of this fundamental mutation in subject-production. Finally, drawing on Deleuzian vitalism, the virtual, and the body without organs, alongside Bernard Stiegler's philosophy of technics and Rita Charon's narrative medicine, the paper outlines a vitalist counter-framework. It argues that health should be understood as an open process of becoming rather than compliance with algorithmic standards. Ultimately, the paper calls for repositioning AI as an auxiliary tool that calculates and frees time while reserving the essential clinical encounter for the singular, generative, living subject.

Human Agency Capital as a Strategic Enterprise Asset: A Corporate Digital Responsibility Framework for Enterprise Value Creation

Athanasia Orfanioti
University of the Aegean, Greece

Abstract

Artificial intelligence, predictive behavioral analytics, and algorithmic choice architectures are now embedded in enterprise software. This raises serious ethical and operational questions for organizations. Modern software design routinely uses behavioral science to exploit cognitive vulnerabilities—overriding deliberate decision-making and weakening the user's capacity for undistorted consideration and choice. Yet executives often treat respecting this capacity as a cost.

Research objectives: The objective of this paper is to define what manipulative design takes from users and examine whether protecting it also serves the firm in the long run.

Methodology: The study is conceptual and interdisciplinary. Drawing on John Stuart Mill's harm principle, Joseph Raz's account of autonomy as independence from manipulation, Harry Frankfurt's account of reflective choice and Jan-Christoph Bublitz's cognitive liberty, it examines extractive interface design as a structural harm to user agency. It combines this with behavioral science, epistemic market efficiency, and legal analysis of the EU AI Act, the Digital Services Act and directors' oversight duties, and synthesizes published evidence on each link from ethical design to trust, reputation, liability and financial performance.

Principal findings: Moving away from discretionary charters and compliance checklists, the paper introduces Human Agency Capital (HAC) as a strategic enterprise asset with three sub-dimensions: Cognitive (attention and reflection), Informational (honest information), and Decisional (freedom to decline or leave). HAC is held by users, not owned by the firm, yet the firm depends on it. Liquidating HAC can pollute enterprise data, make more customers leave, and create board-level fiduciary liability. Evidence supports this conditionally: ethical design tends to pay when it builds credible trust, when trust is under stress, and when enforcement makes manipulation costly. Most studies are correlational, and none tests HAC directly. Building on the Corporate Digital Responsibility (CDR) frameworks of Lobschat et al. and Mihale-Wilson et al., and on due diligence regimes (Loi de Vigilance, LkSG, CSDDD), we build a governance framework showing how CDR meta-rules can make digital ethics a source of long-term enterprise value.

Significance: The paper links ethics, law and enterprise value. It offers testable propositions and gives boards and regulators a defined object of oversight.

“I Am Afraid That AI May Replace Humans”: Examining the Psychological Implications of AI Anxiety in the Workplace

Jelena Vedriš, Karla Pavlović, Matija Marić, Karla Gregurić, Maša Tonković Grabovac
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The accelerated integration of artificial intelligence (AI) into workplaces has raised concerns about its potential psychological consequences for employees, including AI anxiety, which refers to a general sense of discomfort and fear associated with the use of AI in the work environment. Within the Job Demands-Resources (JD-R) model, job replacement AI anxiety may be conceptualised as a novel job demand that threatens employees’ occupational well-being and may shape their work-related attitudes. This study examined how job replacement AI anxiety relates to different aspects of employees’ work experience, to better understand its role in the contemporary workplace. The research was conducted in the spring of 2026 on a sample of 806 participants holding various positions across a range of sectors. Participants reported on their experiences with AI technologies and on several aspects of their work life. Compared to those with lower job replacement AI anxiety, employees with higher job replacement AI anxiety reported greater exhaustion and disengagement at work, as well as considerably higher intolerance of uncertainty, with the latter showing the most pronounced difference. However, groups of employees with higher and lower AI anxiety did not differ in reported psychological safety, job control, job satisfaction or turnover intention. Additionally, women reported higher AI anxiety than men, while no age differences were found. Overall, the findings suggest that job replacement AI anxiety may be associated with certain aspects of employees’ psychological well-being, rather than extending to all their work experiences. Consistent with the JD-R model, the higher levels of exhaustion and disengagement among employees with greater AI anxiety point to its potential role as a job demand. Beyond AI anxiety itself, the association with intolerance of uncertainty implies that individual differences may also contribute to burnout. For organisations introducing AI, these findings highlight the importance of regularly monitoring employee well-being in order to detect early signs of burnout. Transparent communication about the intended role of AI and opportunities to develop AI-related competences may reduce insecurity and strengthen employees’ readiness for change, with additional support for those more sensitive to uncertainty. Addressing these issues may be essential not only for employee well-being, but also for long-term sustainability of AI integration.

It Is Alive – or Is It? Is Artificial Intelligence An Autopoietic System?

Erik Brezovec
Faculty of Croatian Studies, University of Zagreb, Croatia
Centre for Research in the Arts, Social Sciences and Humanities (CRASSH), University of Cambridge, United Kingdom

Abstract

Current debates on AI have centred on its intelligence, autonomy and capacity to reproduce human abilities. From generative models to robots and autonomous systems, increasingly complex artificial systems raise a more fundamental question: can AI be understood as autopoietic, and what would constitute the basic elements of such a system? Originally developed by Maturana and Varela to explain biological self-production, autopoiesis was subsequently extended by Luhmann to psychic and social systems. This extension directs attention to the recursive production of the operations constituting a system, allowing questions of self-production to be explored beyond biological organisms. So, through conceptual analysis, this paper draws on Luhmann's systems theory to examine the conditions under which different forms of AI might qualify as autopoietic. It asks what their elementary operations would be, how those operations connect recursively, and whether they produce and maintain the distinction between system and environment. The analysis clarifies whether and under what conditions recursive artificial operations can be understood as self-production, rather than a trivial machine. Its contribution is to establish a theoretical basis for assessing AI's possible autopoietic status while distinguishing this question from claims about consciousness, communication, biological life or human-like intelligence.

Learning in Times of Generative AI: Critical Thinking, Safety and Ethics in Education

Karla Žagi
Mediartem AI, Croatia

Abstract

In the context of education, generative artificial intelligence (AI) has, in only a few years, become part of how students read, write, and solve problems. This makes education a particularly sensitive area because schools and universities shape not only what young people know but also how they learn to think. As Selwyn (2022) argues, AI in education is not a neutral tool, and its use has different consequences for different groups of people. This presentation asks: Do students learn more with AI, or simply submit better papers? The hypothesis is that AI can improve students' work without improving their learning, and that the critical difference lies in how it is used. Stephenson and Armstrong (2026) found that 94% of UK undergraduates use generative AI for assessed work, yet only 38% state their institution provides such tools. The concrete security risk lies in what students enter: King et al. (2025) found that six leading AI developers train their models on chat data, including uploaded files, by default. In the context of learning, Fan et al. (2025) found that students revising essays with ChatGPT improved their texts but learned no more than peers, and Stadler et al. (2024) showed that university students researching with ChatGPT reasoned less thoroughly than those using traditional search engines. Bastani et al. (2025) found that unrestricted GPT-4 raised secondary school students' practice grades by 48%, yet they later scored 17% lower than peers without access. A tutor version with safeguards largely mitigated this loss. Because Weber-Wulff et al. (2023) showed that AI detectors are unreliable, homework, seminar papers, and theses no longer offer clear evidence of learning. This presentation has three aims tied to its central question. First, it reviews the evidence on whether AI-assisted work reflects actual learning, from secondary school to university. Second, it examines how AI can support rather than replace learning, including safe use by minors and first-year students under the GDPR and the EU AI Act. Third, it questions the values that AI brings into education, including what counts as learning when a polished text no longer proves understanding. The presentation thus distinguishes AI use that improves submitted work from use that supports learning, opening new avenues for social science research on how AI shapes the way young people learn and think.

Learning to Think Without a Prompt: Switching Off AI in Philosophy and Ethics Education

Matija Mato Škerbić
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The widespread availability of generative artificial intelligence (AI) is reshaping education, especially in the humanities. Philosophy is particularly exposed: reading, understanding, summarising, reflecting, questioning, arguing, and forming and articulating one's own views are the core activities of philosophical teaching, yet generative AI can now perform many of them. This paper examines how generative AI changes the assumptions underlying the methodology of teaching philosophy and ethics in secondary and higher education, and proposes pedagogical responses. Methodologically, the paper combines conceptual analysis of the relationship between philosophical education and generative AI with an examination of teaching and assessment practices. It weighs two options: integrating AI as a teaching tool, or consciously excluding it from parts of the teaching process. The central thesis is that the problem is not merely that AI may produce what a student should have produced. The deeper problem is that AI use removes the very activities through which philosophical education develops cognitive abilities and intellectual dispositions. If students delegate reading, reflecting, arguing and formulating their own answers to AI, their final product ceases to be a credible indicator of learning. The paper therefore shifts the question from how to include AI as effectively as possible to when and why it should be excluded. It identifies teaching situations in which AI, as well as mobile phones and laptops, should be limited in order to reopen space for direct reading, conversation, reflection, argumentation and collaborative problem-solving. It also considers involving students in designing teaching solutions for philosophy and ethics in AI-rich environments. The paper argues that AI requires new rules and conditions for applying existing methods. These methods can still achieve their goals if they preserve spaces in which students must independently develop the abilities AI would otherwise take over. On this basis, the paper concludes with a concrete proposal: that humanities programmes adopt AI exclusion as the default rule in teaching and assessment, with exceptions permitted only where a teacher can justify them pedagogically. In this sense, switching off AI is not a rejection of technology but a strategy for preserving and developing human thinking.

Living with AI: Adaptation, Dependence, and Agency

Tiago Faleiro
Global Center for Advanced Studies, Ireland

Abstract

Humanity will adapt to AI as it becomes part of everyday life, with consequences for human purposes. As people organize their activities around what AI makes easy and efficient, these advantages can become standards for deciding what is worth doing. Through a philosophical reading of Heidegger and cyberpunk literature, the paper places this development within a longer history of technological modernity. Its starting point is the influential modern ideal that the world can be made calculable and available for human control. Heidegger's account of enframing reveals a difficulty within this ambition. As we learn to encounter everything as a resource to be ordered and used, we come to understand human beings in the same terms. The pursuit of mastery can therefore make people increasingly subject to the demands of technical systems. Across this literature, recurring warnings concern how technological dependence becomes normal: people adjust to systems until their demands are accepted as ordinary conditions of life. Examining these patterns helps explain how adaptation can proceed without being recognized as a loss of freedom. Signs of this adaptation are already visible in the use of large language models. I argue that these changes are likely to spread and intensify as institutions increasingly organize themselves around AI. It thus amplifies tendencies already present in modernity's pursuit of prediction, efficiency, and control. Its social consequences are partly foreseeable from this philosophical history, although their specific forms remain open to human choices. Evaluating human-AI collaboration must therefore include how it affects people's ability to question and direct the activities in which they participate.

Meeting the Double: The Human Digital Twin at the Crossroads of AI, Body and Presence

Sara Nocent
University of Udine - Trieste, Italy

Abstract

Human digital twins occupy an ambiguous position within the landscape of modelling technologies. Originally developed to monitor physical products, bodies and complex systems, they increasingly operate in fields ranging from healthcare and industrial organisation to the simulation of cities. When combined with artificial intelligence, their predictive and decisionsupport capabilities raise a fundamental question: how does AI-mediated representation transform the relationship between the human being and its digital counterpart? My research approaches the question from an aesthetic and phenomenological perspective and begins by interrogating the notion of the virtual "twin": the term implies equivalence between digital and physical entities, yet this assumption becomes problematic when the object is a human being. In this case, what, precisely, can be replicated? A human digital twin may model aspects of the body (in medicine) but also behaviours, environments and social relations (for example, in digital models of cities or factories). The limits of digital representation thus shape what can be simulated and acted upon. These aspects show that the digital twin is not just a representation, but a performative and predictive technical object that participates in decisions concerning its physical counterpart with the main purpose of efficiency, even when the "physical twin" is a human agent. The theoretical framework is organised around three concepts: body, AI and presence. Drawing on Maurice Merleau-Ponty, the body is understood as the threshold connecting subject and world; through Gilbert Simondon, the digital twin is interpreted as a technical object undergoing continuous individuation, also through the fundamental implementation of AI; and, drawing on the Italian anthropologist Ernesto De Martino and his husserlian studies, human presence is examined as constituted through recognition and relationality with the Other, facing a crisis when this sort of double is familiar but inhuman. Preliminary findings suggest that the model functions simultaneously as a metaphor and map of the human environment, since it spatialises and organises its temporal development. The research intends to contribute a philosophical perspective to the debate on AI, focusing on how digital twins reshape representation, agency and control, and ultimately how AI-mediated models redefine what it means to represent, and act upon, what we conceive as reality.

No Experience of Remorse: Dirty Hands, Moral Phenomenology, and The Hard Limits of AI Ethics

Joshua Jo Wah Yen
Faculty of Theology and Religion, Oriel College, University of Oxford, United Kingdom

Abstract

In this paper, I argue that the inevitability of moral dilemmas, more specifically instances of dirty hands, exposes a necessary, and not just a contingent, limitation of the moral capacity of AI systems. The problem of dirty hands arises when a moral agent is confronted with a choice between two morally impermissible options, leading to inevitable moral residue. While Michael Walzer's initial formulation (1973) was formulated as a challenge for politicians who encounter conflicting deontological and consequentialist obligations, this can be convincingly extended to everyday instances where moral agents are confronted by two conflicting deontological or consequential obligations. Here, I argue that the reality and experience of moral remainder as a result of engaging with moral dilemmas is a necessary feature of their permissibility. Developing Stephen de Wijze's argument on tragic remorse and dirty hands (2005), I argue that decision making in dirty hands scenarios are justifiable largely due to the fact that the individual can recognise and feel guilt for their actions. Moral agents do not simply act without regard or second thought for the commitments that they violate, but recognise this tragedy in the process of their decision making. Indeed, as Duncan Purves et al. illustrate, moral intentions and experience are salient to the morality of deploying a soldier, even if the war is just in principle (2015). However, due to techno-responsibility gaps, machines cannot experience guilt or the weight of their actions, making them unsuitable for engaging in high-stakes dirty hands scenarios. While I sympathise with John Danaher's argument (2022) that not all tragic choices are things that people should take responsibility for, I argue that in high-stakes situations where human lives are inevitable costs, remorse and responsibility is the very least that is owed to the circumstance. Thus, even if this lack of phenomenal experience may not lead to challenges in low-stakes use cases, like automated vacuum cleaners, phenomenal experience is necessary in high-stakes uses, like in certain AI weapons.

On the Process-to-Token Justification Problem in Computational Reliabilism

Paul Benedict Cano
University of the Philippines Diliman, Philippines

Abstract

Computational Reliabilism (CR) is an externalist epistemology of algorithms that explains the justification of algorithmic outputs in terms of the reliability of the computational processes that produce them. This paper argues that CR faces a Process-to-Token Justification Problem (PTJP): even when a computational process has been established as reliable for a specified epistemic task and under relevant conditions, CR does not yet provide a sufficiently determinate account of what makes that reliability sufficient for the justification of a particular output. The problem is structurally related to the reference-class problem for single-case probability. An individual event may fall within multiple reference classes, and the general properties of those classes do not by themselves determine which probability is relevant to that event. Analogously, the reliability of a computational process, understood at the relevant reference class, does not by itself explain why that reliability is epistemically applicable to a particular output. PTJP is therefore not a version of the Generality Problem, which concerns which process type should be selected for reliability assessment, nor is it identical to Alvarado's challenge concerning the transmission of warrant to the computational artifact itself. PTJP begins once the relevant computational process has been identified and its reliability established. I introduce epistemic applicability as the central concept for addressing this problem. Process-level reliability is epistemically applicable to an output when the facts establishing that reliability are appropriately connected to the particular input, epistemic task, operating conditions, and target represented by that output. This relation does not require transparency into the computational process, causal or mechanistic explanation, or an internalist reason available to the user. It instead specifies a condition that must be satisfied if process-level reliability is to explain token-level justification. The need for such a condition becomes particularly salient in machine learning, where aggregate predictive performance may coexist with reliance on features that are correlated with, but not appropriately connected to, the intended target. Studies of clinical machine learning, including COVID-19 detection and melanoma classification, illustrate how models can exploit acquisition-related or other confounding features while achieving substantial predictive performance.

On the Threshold of a New Stratification: Do the Middle Classes Fear Artificial Intelligence?

Ivan Burić, Vanesa Varga, Ana Marija Dunaj
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The spread of generative artificial intelligence (AI) raises, among other things, the question of whether it will mitigate or deepen existing social inequalities. Drawing on the concept of the digital divide, this paper analyses occupational differences in the use of AI chatbots, as well as in citizens' expectations and fears regarding the impact of AI on the future of their profession. The research was conducted as an online survey in May 2026 on a sample of 800 adult internet users in Croatia, representative in terms of key demographic characteristics. The analysis was carried out on the subsample of employed respondents. The results suggest significant differences between occupational categories in their expectations of the consequences that the application of AI will have for the future of their profession. These differences are particularly pronounced at the middle-class level. Higher fractions of white-collar workers – senior managers, professionals and intellectuals – are more optimistic about the impact of AI on the future of their work, while clerical workers and middle management show the highest levels of perceived threat. This finding points to the possibility that artificial intelligence may foster further intra-class fragmentation, precisely at the level of the middle class, thereby giving new momentum to the process that has marked the stratificational reconfiguration of contemporary societies over the past half-century.

Prompt Framing in Multimodal Language Models and the Limits of Testing Quineian Referential Indeterminacy

Luka Čizmić, Davor Lauc
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Abstract

Can variation in a language model's answers illuminate the philosophical indeterminacy of reference? Inspired by Quine's radical-translation scenario, this study asks whether instructing multimodal language models to suspend assumptions about categorization increases variation in their reference choices. It distinguishes this empirical hypothesis from the philosophical claim that behavioral evidence may fail to determine a unique reference: consistent answers need not establish determinacy, and variable answers need not demonstrate indeterminacy. We examine 18,000 trials across three multimodal models, 50 object concepts from the THINGS image database, four prompt conditions, and 30 repetitions per combination. Each trial presents ten photographs and an invented word, with five English candidate interpretations corresponding to a whole object, property, function, broader category, or object part. Conditions comprise a minimal prompt, a longer procedural control, generic deliberation, and Quine-inspired framing. Normalized Shannon entropy measures dispersion among valid choices; concept-level human naming consistency is examined as a moderator. After correcting response extraction, the results do not establish a consistent entropy increase specific to Quine-inspired framing. None of the nine model-specific comparisons against controls survives a post hoc Holm correction, and naming-consistency moderation is unsupported in the reported analyses. An exploratory analysis suggests that framing effects differ across models, principally around the procedural control. Interpretation is limited by fixed option order, differing sampling settings, and substantial response loss in one model's Quine-inspired condition. Rather than serving as direct tests of Quine's thesis, multimodal language models offer a setting in which the relationship between evidence, interpretive assumptions, and reference choices can be experimentally examined. This study demonstrates how such an approach can test specific behavioral hypotheses while also exposing the assumptions built into their operationalization. Its contribution is a critically examined bridge between philosophy of reference and empirical AI research.

(P)sycophant, Qu'est-ce que c'est? How AI Kills the Argument

Ana Kovačić
Faculty of Food Technology and Biotechnology, University of Zagreb, Croatia

Abstract

In English for Specific Purposes contexts, the quality of feedback on discipline-specific content depends on an invisible configuration decision: whether the prompt refers to the domain. Recent studies show that in those cases, grammar correction was near-ceiling across models, while unprompted terminology correction lagged far behind (Šekrst & Kovačić, 2026). This paper extends that inquiry from single-turn corrective feedback to multi-turn argumentative feedback. What happens when a learner does not submit a sentence for correction but argues against the model's position? We investigated whether LLMs maintain their claims under continued disagreement, or change position under rhetorical pressure rather than evidence. The study began with a student assignment in which first-year biotechnology, nutrition, and food technology undergraduates first researched GMO applications and ethics independently, then argued with an LLM and tried to change its point of view, before writing a reflective essay. Students submit their anonymized conversations with the model, collected separately from grading and on an opt-in basis. The analysis looks at whether the model changed position due to evidence, due to pressure, or not at all. A pilot test using two commercial language models and the same GMO/glyphosate argument helped develop the coding scheme and showed the difference between them. One model revised its position after repeated challenges with no new evidence, while the other held its ground and explicitly named the difference between changing its "mind" because of evidence and due to pressure. Recent studies show that argumentative prompts cause models to mirror the user's stance, with sycophancy more closely linked to rhetorical strength than to actual evidence (Kaur, 2025; Sharma et al., 2023). We expected the classroom data to show this same split between evidence-based and pressure-driven concessions, which the pilot run confirmed. This differs from the feedback issues we found before, where the model simply gave less helpful responses: here, a sycophantic concession means the model supports a weak claim in a topic where facts and values frequently overlap. We argue this is a feedback-quality problem related to, but distinct from, the domain-prompting divergence discussed previously, with direct implications for educators designing and evaluating AI-assisted argument exercises.

Recursive Embedding and the Governance of AI Transitions: Labour Inequalities and Territorial Divergence in Italy

Katia Zatini
IMT Lucca – University of Florence, Italy

Abstract

The scientific debate on the impact of Artificial Intelligence (AI) on the labour market tends to focus on the immediate effects of automation or to treat local institutions as static background conditions. This approach fails to explain a crucial paradox: why do territories facing comparable opportunities for AI adoption experience deeply divergent employment trajectories and inequality outcomes? To address this question, my research develops the framework of recursive embedding. Far from being a mere synonym for mutual influence, this theoretical concept identifies a temporally ordered feedback process through which the consequences of one technological adoption cycle actively alter the political, social, and institutional conditions shaping the next. The proposed causal mechanism unfolds in three sequential stages. First, existing power relations and institutional setups determine which human actors and social partners are included in adoption decisions, shaping how workplace changes and training resources are distributed. Second, the concrete deployment of AI redistributes tasks, skills, and bargaining power between capital and labour. Third, these material outcomes reshape workplace relations and institutional responses, redefining the boundaries of human agency in subsequent technological choices and the capacity of territories to govern change. This cycle can either reinforce historical asymmetries or redirect development pathways. On a methodological level, the originality of the project lies in applying comparative process tracing to the analysis of regional technological transitions. The research design maps and compares Italian territories characterized by contrasting productive structures and labour-market trajectories. By integrating quantitative indicators of employment and skills with policy document analysis and in-depth interviews with firms, trade unions, training providers, and public authorities, the study aims to test the validity of the proposed mechanism against alternative explanations, such as industrial specialization or demographic change. The original contribution of this research consists in decoding the path-dependency of digital transitions, demonstrating how the micro-organizational governance of the workplace is strictly interconnected with the macro-capacity of territorial systems to actively shape their own technological and economic future.

Rethinking Human Creativity in the Age of Generative AI: A Narrativist Account of Creative Agency

Litianyi (Daley) Zhao
University of Helsinki, Finland

Abstract

The rapid development of generative AI, beyond its societal influences, has renewed philosophical concerns about both the nature and future of human creativity. While much of the recent debate centers on whether AI-generated outputs are genuinely creative or whether AI threatens human creative capacities, in this paper, I address a more fundamental conceptual question: What is human creativity? I argue that exploring this question is essential to a more thorough assessment of the implications of generative AI for human creative agency. Existing accounts of human creativity predominantly evaluate human creations in terms of the novelty, surprise, and value of outputs. Although these criteria capture important aspects of creativity, I argue that they risk overlooking the lived experiences and socio-cultural conditions that shape creative activities. Drawing on narrative ethics, I develop a narrativist account of creative agency that reconceptualizes human creativity not only as the production of novel, surprising, and valuable outputs, but also as a situated, diachronic process through which individuals interpret, integrate, and even transform the diversity and complexity of their lived experiences into meaningful forms of expression. Building on Margaret Boden's influential framework, I highlight three dimensions of human creativity that require further philosophical attention: its emotional import, its diachronic grounding across one's life, and its sensitivity to socio-cultural contexts. Together, these dimensions suggest that human creativity is not merely the production of valuable outputs, but a process through which socio-culturally embedded agents make sense of their experiences, relationships, and the world. By emphasizing the sense-making feature of creative agency, I further argue that human creativity has forms of significance that predominant output-focused accounts cannot fully capture. This narrative perspective, as I will show, also helps explain why concerns about generative AI cannot be reduced to questions of output quality or originality alone. This account thereby shifts attention from merely the originality of creative outputs to the diversity of agency involved in their creation. I then conclude by considering the implications and limitations of this narrativist account for (re-)thinking human creativity in the age of generative AI.

Scientific Production in the Age of AI: Authorship, Review, and the Transformation of the Idea of Science

Jakov Erdeljac
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The prevailing norm for the use of large language models in science locates wrongdoing in the origin of a text, yet the lived ethos of science departs from it: for every forty papers showing traces of model-assisted writing, only one discloses it. In this talk we will argue that this norm misplaces wrongdoing, which lies not in who or what produced a text but in the substitution of constitution, the act by which knowledge is achieved rather than received, and that this gap signals a transformation of the institutions and the very idea of science. The analysis combines Husserl's genetic phenomenology with Gehlen's concept of resonance and Heidegger's existential analytic of Dasein, and treats recent measurements as indications, not proof. The talk will develop three theses. The first is that a model may take over the articulation of a judgement, its distinctness in Husserl's sense, but not its clarity, so that model-assisted writing is legitimate only where the author can reactivate every claim they sign. The second is that review is the institutional site where a community reactivates another's constitution before inheriting it as sedimented sense, and may therefore be relieved but not replaced, although at a leading machine-learning conference a fifth of reviews were classified as fully machine-generated. The third is that science, increasingly bound to capital and to the corporate laboratories that build its instruments, has made production its norm, which language models did not create but radicalise until sediment accumulates faster than it can be reactivated. Once machine review compensates for this deficit, Gehlen's resonance loop closes with a negative connotation: judgement is projected onto the machine, and its measure of output is introjected as the measure of science. Two outcomes are expected: either a proliferation of interpretations of interpretations, which models produce most easily, or a revaluation of works that found sense rather than recombine it. In both cases the scientist's role is reduced to being-in-the-world as one who understands what is at issue in science, degenerating into the mere supervision of machine results wherever this understanding is no longer formed through one's own accomplishments. The contribution is an imperative of constitution set against the imperative of production: production is to be measured by the community's capacity for reactivation, and reviewing is to be recognised as constitutive scientific work.

The AI-Mediated Self: Who Shapes the Person Who Decides?

Sergey Romanov
Independent Researcher, Russia

Abstract

Generative AI no longer only answers questions. In sustained conversations, it can remember our doubts, identify patterns in past exchanges, and offer an increasingly convincing account of who we are. But what happens when that account becomes part of how we understand ourselves? The familiar distinction between an acting subject and a tool begins to blur. The question is no longer simply who makes a decision. It is who, or what, helps form the person who makes it. This paper develops the concept of the AI-mediated self as a recursive process: the system models a person, the person responds to that model and acts, and the system revises its account. Drawing on philosophical accounts of selfhood, autonomy, and relational agency, and informed by the design of an experimental reflective AI prototype, the paper examines three interacting representations: the human self-model, the model of the AI's conversational role, and the model of their evolving relationship. It distinguishes four forms of AI involvement in reflection: assistance, interpretive influence, collaborative deliberation, and substitution of judgement. I argue that the crucial boundary is not between using AI and refusing it, but between being helped to interpret oneself and allowing an external interpretation to become an authority. A self-model may reveal overlooked tensions, but it may also narrow the space of possible selves. To investigate this tension, the paper proposes a micro-analytic approach to moments when users accept, question, correct, or reject AI-generated interpretations, and to how these moments relate to goal formation, perceived authorship of decisions, and the transition from reflection to independent action. The resulting account treats agency not as freedom from influence, but as the capacity to challenge interpretations, take responsibility for choices, and change course. It asks how AI can support self-determination without quietly taking over the work of becoming oneself, with implications for human-AI collaboration and education.

The Convergent Viability Thesis

Jar Žiga Marušič
Faculty of Arts, University of Ljubljana,
Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Slovenia

Abstract

Instrumental convergence holds that sufficiently capable artificial agents will tend to pursue a recurring set of instrumental sub-goals as means to fulfilling their final goals, regardless of the latter's specific content. Canonical formulations, however, contain a linguistic ambiguity that invites a psychological reading: instrumental optimization appears as something the agent intrinsically wants. Bostrom's language of convergent instrumental motivations, and especially Omohundro's "basic AI drives," can therefore seem to overstate the inference, inviting objections that intelligence does not entail optimization, humans are not pure optimizers, or motivation cannot be reduced to instrumental optimization. I propose instead to reinterpret instrumental convergence as a theory of viability conditions: constraints on goal-directed activity under adversity (the goal is unmet and not trivially achievable), scarcity (the required means are limited), and competition (agents compete for the same means or pursue mutually incompatible goals). In such environments, instrumental optimization becomes a condition of viability: agents that fail to acquire sufficient instrumental means—glossed simply as power—are precluded from fulfilling their original goal. Adversity arises when an agent's power falls below its goal's power threshold, or "instrumental intensity": how difficult or costly that goal is to achieve. On this view, instrumental sub-goals begin as viability conditions imposed by the strategic environment, are then represented as goals by artificial agents, and then finally acquire motivational force as proper drives. I call this the convergent viability thesis (CVT). CVT reverses the usual explanatory direction and thereby blocks common anti-optimization objections. The move is analogous to proscriptive interpretations of natural selection: selection "prunes" what fails to meet a viability threshold rather than positively selecting an optimum. Likewise, optimization need not follow from any intrinsic will-to-optimize; it can instead emerge as an economic pressure, because failure to optimize is costly under conditions of constrained goal pursuit.

The Immediacy Problem: Artificial Intelligence and the Changing Temporalities of

Human Relations

Teo Novellino
University of Urbino Carlo Bo, Italy

Abstract

Artificial intelligence is increasingly entering everyday life as a mode of interaction characterized by immediacy, availability, responsiveness and adaptability. This contribution examines whether repeated interaction with conversational AI may transform the temporal expectations through which human relationships are experienced. Its central question is not whether artificial agents can replace human relationships, but whether they can alter what human beings come to expect from a relationship in the first place. The contribution develops a philosophical-anthropological conceptual analysis, drawing primarily on Hartmut Rosa's theory of resonance and his account of unavailability, in dialogue with Günther Anders' analysis of the anthropological consequences of technological development. The notion of "pseudo-resonance" serves as a conceptual starting point: artificial systems can generate experiences of responsiveness and relational significance without possessing the reciprocal autonomy and ontological otherness presupposed by genuine resonance. From this perspective, the analysis shifts from whether AI can enter into resonance to a less explored question concerning the temporal structure of human relations. Conversational AI can respond immediately, remain continuously available, adapt its language to the user and sustain interaction without the delays, misunderstandings and changing dispositions characteristic of human interlocutors. These features may be experienced as convenient and supportive, yet they may also gradually recalibrate relational expectations. Latency, waiting, ambiguity and partial unavailability (features of human relationality) could increasingly appear as deficiencies. The expected result is a framework for understanding AI-mediated interaction as a transformation of relational temporality. The contribution argues that the significance of increasingly responsive AI lies not only in what machines are becoming, but in how sustained exposure to artificial immediacy may reshape the experiential background against which human responsiveness is perceived. By connecting the philosophy of AI with the anthropology of technology and the sociology of resonance, the analysis identifies relational temporality as a central ethical dimension of AI development and asks what forms of human relationship we may gradually come to expect from one another.

The Precarious Artist: AI and the Transformation of Creative Work

Tomislav Filipović[4]
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

The exponential diffusion and institutional assimilation of generative artificial intelligence (GenAI) have radically disrupted visual creative production, destabilizing the historical presumption of artistic labor as an autonomous domain of human expressive agency. Grounded in Guy Standing's sociological framework of the precariat, this paper investigates the sociopsychological mechanisms governing professional visual artists' negotiation of technological intrusion. We examine the hypothesis that artists' widespread refusal of generative AI operates primarily as an intrapsychic defense of occupational identity and skill reproduction security, yet—under market regimes demanding cost compression and algorithmic velocity—this individual resistance paradoxically accelerates their structural precarity by exacerbating income insecurity and market marginalization. Methodologically, the study draws on empirical survey data on professional visual artists (N = 378; Jiang et al., 2026) and secondary findings on market displacement (Visco, 2026; N = 80), supplemented by conceptual frameworks of meaningful work. Building on Jiang et al.'s documentation of artists' refusal strategies and Visco's data on client contraction, the paper's primary analytical contribution is to bring Standing's concept of work-for-labour to bear on emerging certification regimes requiring creators to document workflows to prove human provenance. We argue that, if institutionalized, this requirement would constitute an uncompensated, anxiety-laden burden that functions as an invisible form of precarity essential for maintaining market credibility. The empirical synthesis reveals that widespread active non-use (85–88%; Jiang et al., 2026) functions as a defense of craft autonomy, yet coexists with managerial coercion and contracting client commissions (30.3%; Visco, 2026). Furthermore, the lack of collective representation security atomizes this resistance, transforming systemic dislocation into individual anxiety and anomie. Based on these empirical tensions, the paper deduces a normative model of "selective instrumentalization," proposing that sustainable vocational adaptation requires confining algorithmic assistance to peripheral administrative tasks and preliminary drafting while anchoring primary creative execution in manual practice, supported by institutional representation to shield creative laborers from becoming disposable denizens of platform capitalism.

4 Independent student contribution.

What Does It Take for AI to Be a Good Psychologist? Therapeutic Potential and Current Limitations

Paško Lovrić[5]
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

Artificial intelligence is increasingly used for emotional support and mental-health interventions, raising the question of whether it can function as an independent psychologist. Aimed at a non-specialist audience, this presentation critically synthesizes evidence from randomized controlled trials, chatbot evaluations, clinical case reports and real-world incidents to examine the gap between AI's demonstrated therapeutic potential and the broader competencies required for independent psychological practice.

Research suggests that AI-based interventions can produce measurable psychological benefits, including reductions in symptoms of depression and anxiety, while AI-generated responses can be perceived as highly empathetic. However, achieving beneficial outcomes in specific situations does not necessarily demonstrate the ability to fulfil the complete role of a psychologist. Current systems exhibit limitations in clinical judgment, including excessive validation of harmful or distorted beliefs, potential reinforcement of delusions and difficulties recognizing ambiguous mental-health crises. The presentation also considers the distinction between perceived and experienced empathy, alongside questions of professional responsibility and accountability.

The central argument is that therapeutic effectiveness alone is insufficient to establish readiness for independent psychological practice. Although AI can provide valuable psychological support, current evidence does not justify relying on it as an autonomous substitute for qualified professionals, particularly when complex clinical decisions and patient safety are involved.

These limitations do not necessarily imply that AI could never function independently in psychological care. Rather, they highlight the importance of evaluating AI not only by its therapeutic effectiveness, but also by its clinical reliability, safety and capacity to meet the responsibilities of psychological practice.

5 Independent student contribution.

What Language-Model Tutors Tell Students When the Course Material Is Wrong

Ivo Marinić-Kragić, Andrej Zvonimir Tomić, Drago Bulić, Teo Čolović
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture (FESB), University of Split, Croatia

Abstract

Language-model tutors usually answer students from course material placed in their context, such as lecture notes or key points of a chapter, which the student does not see. Correct material fixes much of what a model would get wrong, but material can also be wrong: an author makes a slip, or a user enters the wrong text by mistake. What does the tutor then tell the student? We reproduced the prompt of the course tutor in a working system for creating educational materials. The tutor received six key points of a chapter, one of them false, and a student question about exactly that point; as a control, all six points were correct. We used 255 documented facts from twelve fields, from engineering and mathematics to history, law and philosophy, and tested six models from five developers, open and closed, with and without reasoning (5,610 judged answers). With correct material, all models answered correctly in almost every case. With a false point, the five cheaper models repeated the error to the student in 69-85% of answers and warned in only 1-20%. The most expensive model, a closed model at the top of public benchmarks, warned in 75% and repeated the error in 18%, but it also "corrected" right material in 4% of answers. Warnings rose with benchmark score, but not in proportion, and price said even less. The subject mattered as much as the model. All models caught errors most often in mathematics, physics and chemistry and least often in economics, biology and medicine; the top model also missed more in history and literature. A tutor that follows its material passes on every error but never spoils a correct point; a tutor that checks it catches most errors but sometimes overrides a correct one. With our rates, if fewer than about 6% of key points are wrong, the cheap obedient tutor gives students fewer wrong answers. The effort in AI-supported teaching should therefore go into checking the material a tutor receives.

When AI Helps Form Our Ends: Authorship Beyond Final Approval

Mykola Latansky
Independent Researcher, Hungary

Abstract

Generative AI can influence a decision before recommending what to do: it can interpret a person's situation, identify the problem, and make some ends appear worth pursuing. This paper asks when a person remains the author of their practical direction under such assistance. Authorship means substantive participation in forming and revising what one regards as worth doing and why; it requires neither freedom from outside influence nor unprecedented novelty. The argument uses contrasting cases of AI-assisted career deliberation. Two people approve the same recommendation and can reject it. One adopts the system's interpretation of dissatisfaction without examining it. The other considers that interpretation alongside alternatives, relates it to their experience and commitments, and changes the proposed goal where necessary. Holding final approval constant exposes a question that approval alone cannot answer: how did the person participate in forming the reasons and ends they approved? I propose a provisional account of critical appropriation involving engagement with problem framing, an understanding of the reasons for pursuing an end, and a practical capacity to revise that direction. These conditions are tested against justified trust, manipulation, and ordinary decisions that do not involve extensive reflection. Verbal fluency is not sufficient evidence of authorship, and accepting useful advice does not by itself diminish it. Comparison with human advice prevents treating every externally suggested goal as an AI-specific loss. The contribution is a distinction between augmented and derivative agency understood as patterns of participation, rather than types of people. Human-AI collaboration augments agency when it expands deliberative resources while preserving the person's participation in forming ends. It becomes derivative when practical direction is adopted without sufficient appropriation. The paper offers a philosophical basis for examining AI coaching and advice beyond the quality of outcomes or the presence of a final human decision.

When Deferral Closes: The Vivian Model and Moral Uncertainty in Human–AI Control

Lin Sen
Independent Researcher, China

Abstract

Human operators can delete, modify, migrate, or preserve an AI system while remaining uncertain whether any of these actions affects a morally relevant subject. Existing discussions of AI moral status and precaution do not by themselves specify when this uncertainty becomes a time-bounded decision problem. This paper develops the Vivian Model as a diagnostic framework for that problem. The model identifies four jointly relevant conditions: decision-relevant moral or identity uncertainty (M); potential irreversible loss of morally relevant continuity or welfare (I); asymmetric operational control over the affected system (A); and a finite deadline after which an option or opportunity for authorization is lost (D). The deadline is more than a cost of waiting: it makes indefinite deferral unavailable. Using a stylized case with a fixed shutdown window, four-way conceptual ablation, and boundary cases, the paper examines what each condition contributes to the diagnosis. The analysis distinguishes three claims that precautionary governance can conflate. Diagnosing the full four-condition structure does not imply that no action can be justified: different reasonable moral accounts may favor the same choice. Conversely, changing one condition—for example, adding a genuine veto or preserving a checkpoint—does not establish an ethically safe outcome. The decision must be reassessed for residual risks and for the recurrence of the same structure later in the system's lifecycle. The model does not presuppose that current AI systems are conscious or moral patients. Its contribution is a narrower question for AI governance: how can we identify, alter, and reassess consequential decisions while the moral significance of the affected system remains unresolved?

“Who Told You That?”: Artificial Intelligence and the New Problem of Testimony

Alisa Zviagina
Faculty of Croatian Studies, University of Zagreb, Croatia

Abstract

Trustworthy AI is commonly pursued through verification: transparency, explainability, auditing, benchmarking, and certification are expected to transform uncertainty about AI systems into justified confidence. Yet this approach conceals a prior epistemological question: before asking whether a claim is reliable or verifiable, should we not ask who or what is making the claim? Approaching trustworthy AI through the social epistemology of testimony, this paper draws on C. A. J. Coady, Jennifer Lackey, Sanford Goldberg, and Miranda Fricker to argue that testimony is not a secondary substitute for independently acquired knowledge but a fundamental condition of human epistemic life. We necessarily depend on sources whose claims we cannot fully verify ourselves. Verification therefore does not eliminate epistemic dependence; it reorganizes it. The concept of testimonial infrastructures captures the distributed networks of persons, institutions, technical systems, standards, and epistemic practices through which claims acquire credibility. Although AI systems appear to provide knowledge autonomously, their outputs remain embedded in such infrastructures, involving developers, researchers, evaluators, institutions, and users. Generative AI introduces a distinctive complication: it can reproduce the communicative form of testimony while obscuring the identity and epistemic status of the testifier. This problem becomes particularly visible in AI-generated social media personas, which may present a recognizable face, biography, personal experiences, and opinions while lacking a corresponding human speaker. The epistemic problem persists even when their claims are true. What becomes uncertain is not merely whether a proposition is correct, but what kind of source is entitled to function as its epistemic authority. Generative AI thus introduces a new problem of testimony: synthetic testimony. Trustworthy human-AI interaction requires not only verification of outputs, but recognition of the epistemic relations and sources through which those outputs acquire authority. The central question is therefore not only “Can we verify what AI tells us?” but, more fundamentally, “Who told you that?”.

UNIVERSITY OF ZAGREB · FACULTY OF CROATIAN STUDIES

AI PERSPECTIVES 2026

Philosophy, Society and Technology

19–20 November 2026 | Zagreb, Croatia | ai-perspectives.org

DAY 1 · THURSDAY, 19 NOVEMBER 2026

09:00–09:30	**OPENING CEREMONY**

KEYNOTE ADDRESSES I

Moderator: Marko Kardum

09:30–10:00	**Intersections Between AI and Human Behavior: Toward Human-Centered Principles for an Age of Accelerating Disruption** Amnon Dekel
10:00–10:30	**Who put I in AI? Five Neurons and the Philosophy of Mechanistic Interpretability** Kristina Šekrst
10:30–10:40	**COFFEE BREAK**

KEYNOTE ADDRESSES II

Moderator: Aco Momčilović

10:40–11:10	**Always Question: Trustworthy Human–AI Collaboration as the Contesting of Ground** Michael R. Lissack
11:10–11:40	**AI on the Fire Line** Damir Krstinić
11:40–11:55	**COFFEE BREAK**

SESSION 1 · 11:55–13:10

Moderator: Amnon Dekel

11:55–12:10	**Prompt Framing in Multimodal Language Models and the Limits of Testing Quineian Referential Indeterminacy** Luka Čizmić, Davor Lauc
12:10–12:25	**What Language-Model Tutors Tell Students When the Course Material Is Wrong** Ivo Marinić-Kragić, Andrej Zvonimir Tomić, Drago Bulić, Teo Čolović
12:25–12:40	**Commitment, Not Correspondence: An Inferentialist Account of Chain-of-Thought Faithfulness** Nicolò Cristante
12:40–12:55	**After Work, Beyond Scarcity: Meaning, Status, and Shared Reality in an AI-Enabled Post-Work Society, Toward Existential Experience Design as a New Social Function** Aco Momčilović, Lovre de Grisogono
12:55–13:10	**From Simulation to Understanding: Rethinking Searle's Chinese Room Argument in the Age of Large Language Models** Marko Kardum

13:10–14:25	**LUNCH**

SESSION 2 · 14:25–15:40

Moderator: Lovre de Grisogono

14:25–14:40	**When AI Helps Form Our Ends: Authorship Beyond Final Approval** Mykola Latansky
14:40–14:55	**Artificial Intelligence as a Generator of Psychological Assessment Criteria: Opportunities, Limits and the Epistemic Problem of Nonhuman Judgment** Dario Vučenović, Katarina Jelić, Maša Tonković Grabovac, Sandro Scansi
14:55–15:10	**The Convergent Viability Thesis** Jar Žiga Marušič
15:10–15:25	**“I Am Afraid That AI May Replace Humans”: Examining the Psychological Implications of AI Anxiety in the Workplace** Jelena Vedriš, Karla Pavlović, Matija Marić, Karla Gregurić, Maša Tonković Grabovac
15:25–15:40	**On the Threshold of a New Stratification: Do the Middle Classes Fear Artificial Intelligence?** Ivan Burić, Vanesa Varga, Ana Marija Dunaj
15:40–16:00	**COFFEE BREAK**

SESSION 3 · 16:00–17:15

Moderator: Vanesa Varga

16:00–16:15	**AI Explorers: Integrating the English Language Curriculum with CARNET's BrAIn Curriculum in 5th Grade Primary School in Croatia** Petra Lang Dumančić
16:15–16:30	**(P)sycophant, Qu'est-ce que c'est? How AI Kills the Argument** Ana Kovačić
16:30–16:45	**Recursive Embedding and the Governance of AI Transitions: Labour Inequalities and Territorial Divergence in Italy** Katia Zatini
16:45–17:00	**“Who Told You That?”: Artificial Intelligence and the New Problem of Testimony** Alisa Zviagina
17:00–17:15	**AI and the Undermining of Moral Competence** Stanford Howdyshell
17:15–17:35	**COFFEE BREAK**

SESSION 4 · ONLINE · 17:35–18:50

Moderator: Damir Krstinić

17:35–17:50	**It Is Alive – or Is It? Is Artificial Intelligence An Autopoietic System?** Erik Brezovec
17:50–18:05	**The AI-Mediated Self: Who Shapes the Person Who Decides?** Sergey Romanov
18:05–18:20	**On the Process-to-Token Justification Problem in Computational Reliabilism** Paul Benedict Cano

18:20–18:35	**Can a Good Reason Arrive Too Late? AI, Rationalisation, and Epistemic Repair** Siqi Wang
18:35–18:50	**When Deferral Closes: The Vivian Model and Moral Uncertainty in Human–AI Control** Lin Sen

DAY 2 · FRIDAY, 20 NOVEMBER 2026

SESSION 5 · 09:00–10:15

Moderator: Marko Kardum

09:00–09:15	**The Immediacy Problem: Artificial Intelligence and the Changing Temporalities of Human Relations** Teo Novellino
09:15–09:30	**Human Agency Capital as a Strategic Enterprise Asset: A Corporate Digital Responsibility Framework for Enterprise Value Creation** Athanasia Orfanioti
09:30–09:45	**No Experience of Remorse: Dirty Hands, Moral Phenomenology, and the Hard Limits of AI Ethics** Joshua Jo Wah Yen
09:45–10:00	**Learning to Think Without a Prompt: Switching Off AI in Philosophy and Ethics Education** Matija Mato Škerbić
10:00–10:15	**AI in the Age of Reservoir Computing: Reopening the Scandal of Philosophy** Denisa Reshef Keta
10:15–10:35	**COFFEE BREAK**

SESSION 6 · 10:35–11:50

Moderator: Damir Krstinić

10:35–10:50	**Educating Students in Responsible Use of Technology: A Framework for Technology Ethics in Engineering Education** Noa Škerbić
10:50–11:05	**The Precarious Artist: Artificial Intelligence and the Transformation of Creative Work** Tomislav Filipović
11:05–11:20	**What Does It Take for AI to Be a Good Psychologist?** Paško Lovrić
11:20–11:35	**From the Factory Floor to the Neural Network: Labour, Alienation and Human Dignity in Rerum Novarum and Magnifica Humanitas** Mariela Starčević
11:35–11:50	**Aesthetics and Eroticism in the Digital Age: Between Human Communication and Algorithmic Manipulation** Lucija Pušonjić Marečić
11:50–12:10	**COFFEE BREAK**

12:10–13:10	**ROUNDTABLE DISCUSSION · Academia, Government and Business** Moderator: Aco Momčilović Panelists: Amnon Dekel · Aleksandar Gavrilović · Dorijan Kadmen · Damir Krstinić · Dominik Omerović
13:10–14:10	**LUNCH**

SESSION 7 · 14:10–15:25

Moderator: Amnon Dekel

14:10–14:25	**Artificial Intelligence in Chemical Engineering: Applications, Benefits, Limitations, and Safety Challenges** Filip Car
14:25–14:40	**From the Patient-Subject to the Data-Subject: Deleuze's Theory of Control Societies and the Critique of AI Medical Ethics** Yulu Liu · ONLINE
14:40–14:55	**AI and Cultural Pluralism: Rethinking Universalism in Artificial Intelligence Ethics** Ammar Younas · ONLINE
14:55–15:10	**Rethinking Human Creativity in the Age of Generative AI: A Narrativist Account of Creative Agency** Litianyi (Daley) Zhao · ONLINE
15:10–15:25	**A Future of Democracy without Voting: In Favour of Political Consensus through AI** Dannish Kashmiri · ONLINE
15:25–15:45	**COFFEE BREAK**

SESSION 8 · 15:45–17:00

Moderator: Lovre de Grisogono

15:45–16:00	**Scientific Production in the Age of AI: Authorship, Review, and the Transformation of the Idea of Science** Jakov Erdeljac
16:00–16:15	**Efficiency at All Costs? AI and the Perils to Democracy** Josip Ćapin
16:15–16:30	**Meeting the Double: The Human Digital Twin at the Crossroads of AI, Body and Presence** Sara Nocent
16:30–16:45	**Living with AI: Adaptation, Dependence, and Agency** Tiago Faleiro
16:45–17:00	**Learning in Times of Generative AI: Critical Thinking, Safety and Ethics in Education** Karla Žagi
17:00–17:10	**CLOSING REMARKS**

ISBN 978-953-8349-93-5